Lectures on
Electromagnetic Field Theory

Weng Cho CHEW[1]

Fall 2020, Purdue University

[1]Updated: December 3, 2020

Contents

Preface **xi**

Acknowledgements **xii**

1 Introduction, Maxwell's Equations **1**
1.1 Importance of Electromagnetics 1
1.1.1 A Brief History of Electromagnetics 4
1.2 Maxwell's Equations in Integral Form 7
1.3 Static Electromagnetics 8
1.3.1 Coulomb's Law (Statics) 8
1.3.2 Electric Field **E** (Statics) 8
1.3.3 Gauss's Law (Statics) 11
1.3.4 Derivation of Gauss's Law from Coulomb's Law (Statics) 12
1.4 Homework Examples 13

2 Maxwell's Equations, Differential Operator Form **17**
2.1 Gauss's Divergence Theorem 17
2.1.1 Some Details 19
2.1.2 Gauss's Law in Differential Operator Form 21
2.1.3 Physical Meaning of Divergence Operator 21
2.2 Stokes's Theorem 22
2.2.1 Faraday's Law in Differential Operator Form 25
2.2.2 Physical Meaning of Curl Operator 26
2.3 Maxwell's Equations in Differential Operator Form 26
2.4 Homework Examples 27
2.5 Historical Notes 28

3 Constitutive Relations, Wave Equation, Electrostatics, and Static Green's Function **29**
3.1 Simple Constitutive Relations 29
3.2 Emergence of Wave Phenomenon, Triumph of Maxwell's Equations 30
3.3 Static Electromagnetics–Revisted 34
3.3.1 Electrostatics 34

3.3.2 Poisson's Equation . . . 35
3.3.3 Static Green's Function . . . 36
3.3.4 Laplace's Equation . . . 36
3.4 Homework Examples . . . 37

4 Magnetostatics, Boundary Conditions, and Jump Conditions 39
4.1 Magnetostatics . . . 39
4.1.1 More on Coulomb Gauge . . . 41
4.2 Boundary Conditions–1D Poisson's Equation . . . 41
4.3 Boundary Conditions–Maxwell's Equations . . . 43
4.3.1 Faraday's Law . . . 44
4.3.2 Gauss's Law for Electric Flux . . . 45
4.3.3 Ampere's Law . . . 46
4.3.4 Gauss's Law for Magnetic Flux . . . 48

5 Biot-Savart law, Conductive Media Interface, Instantaneous Poynting's Theorem 49
5.1 Derivation of Biot-Savart Law . . . 50
5.2 Shielding by Conductive Media . . . 52
5.2.1 Boundary Conditions–Conductive Media Case . . . 52
5.2.2 Electric Field Inside a Conductor . . . 53
5.2.3 Magnetic Field Inside a Conductor . . . 54
5.3 Instantaneous Poynting's Theorem . . . 56

6 Time-Harmonic Fields, Complex Power 61
6.1 Time-Harmonic Fields—Linear Systems . . . 62
6.2 Fourier Transform Technique . . . 64
6.3 Complex Power . . . 65

7 More on Constitute Relations, Uniform Plane Wave 69
7.1 More on Constitutive Relations . . . 69
7.1.1 Isotropic Frequency Dispersive Media . . . 69
7.1.2 Anisotropic Media . . . 71
7.1.3 Bi-anisotropic Media . . . 72
7.1.4 Inhomogeneous Media . . . 72
7.1.5 Uniaxial and Biaxial Media . . . 72
7.1.6 Nonlinear Media . . . 73
7.2 Wave Phenomenon in the Frequency Domain . . . 73
7.3 Uniform Plane Waves in 3D . . . 75

8 Lossy Media, Lorentz Force Law, Drude-Lorentz-Sommerfeld Model 79
8.1 Plane Waves in Lossy Conductive Media . . . 79
8.1.1 Highly Conductive Case . . . 80
8.1.2 Lowly Conductive Case . . . 81
8.2 Lorentz Force Law . . . 82
8.3 Drude-Lorentz-Sommerfeld Model . . . 82

8.3.1 Cold Collisionless Plasma Medium 83
8.3.2 Bound Electron Case 85
8.3.3 Damping or Dissipation Case 85
8.3.4 Broad Applicability of Drude-Lorentz-Sommerfeld Model 86
8.3.5 Frequency Dispersive Media 88
8.3.6 Plasmonic Nanoparticles 89

9 Waves in Gyrotropic Media, Polarization 91
9.1 Gyrotropic Media and Faraday Rotation 91
9.2 Wave Polarization 94
9.2.1 Arbitrary Polarization Case and Axial Ratio[1] 96
9.3 Polarization and Power Flow 98

10 Spin Angular Momentum, Complex Poynting's Theorem, Lossless Condition, Energy Density 101
10.1 Spin Angular Momentum and Cylindrical Vector Beam 102
10.2 Momentum Density of Electromagnetic Field 103
10.3 Complex Poynting's Theorem and Lossless Conditions 104
10.3.1 Complex Poynting's Theorem 104
10.3.2 Lossless Conditions 105
10.4 Energy Density in Dispersive Media 107

11 Transmission Lines 111
11.1 Transmission Line Theory 112
11.1.1 Time-Domain Analysis 113
11.1.2 Frequency-Domain Analysis–the Power of Phasor Technique Again! . . 116
11.2 Lossy Transmission Line 117

12 More on Transmission Lines 121
12.1 Terminated Transmission Lines 121
12.1.1 Shorted Terminations 124
12.1.2 Open Terminations 125
12.2 Smith Chart 126
12.3 VSWR (Voltage Standing Wave Ratio) 128

13 Multi-Junction Transmission Lines, Duality Principle 133
13.1 Multi-Junction Transmission Lines 133
13.1.1 Single-Junction Transmission Lines 135
13.1.2 Two-Junction Transmission Lines 136
13.1.3 Stray Capacitance and Inductance 139
13.2 Duality Principle 140
13.2.1 Unusual Swaps[2] 141
13.3 Fictitious Magnetic Currents 142

[1] This section is mathematically complicated. It can be skipped on first reading.
[2] This section can be skipped on first reading.

14 Reflection, Transmission, and Interesting Physical Phenomena 145
14.1 Reflection and Transmission—Single Interface Case . . . 145
14.1.1 TE Polarization (Perpendicular or E Polarization)[3] . . . 146
14.1.2 TM Polarization (Parallel or H Polarization)[4] . . . 148
14.2 Interesting Physical Phenomena . . . 149
14.2.1 Total Internal Reflection . . . 150

15 More on Interesting Physical Phenomena, Homomorphism, Plane Waves, and Transmission Lines 155
15.1 Brewster's Angle . . . 155
15.1.1 Surface Plasmon Polariton . . . 158
15.2 Homomorphism of Uniform Plane Waves and Transmission Lines Equations . 160
15.2.1 TE or TE_z Waves . . . 161
15.2.2 TM or TM_z Waves . . . 162

16 Waves in Layered Media 165
16.1 Waves in Layered Media . . . 165
16.1.1 Generalized Reflection Coefficient for Layered Media . . . 166
16.1.2 Ray Series Interpretation of Generalized Reflection Coefficient . . . 167
16.1.3 Guided Modes from Generalized Reflection Coefficients . . . 168
16.2 Phase Velocity and Group Velocity . . . 168
16.2.1 Phase Velocity . . . 168
16.2.2 Group Velocity . . . 170
16.3 Wave Guidance in a Layered Media . . . 173
16.3.1 Transverse Resonance Condition . . . 173

17 Dielectric Waveguides 175
17.1 Generalized Transverse Resonance Condition . . . 175
17.2 Dielectric Waveguide . . . 176
17.2.1 TE Case . . . 177
17.2.2 TM Case . . . 183
17.2.3 A Note on Cut-Off of Dielectric Waveguides . . . 184

18 Hollow Waveguides 185
18.1 General Information on Hollow Waveguides . . . 185
18.1.1 Absence of TEM Mode in a Hollow Waveguide . . . 186
18.1.2 TE Case ($E_z = 0$, $H_z \neq 0$, TE_z case) . . . 187
18.1.3 TM Case ($E_z \neq 0$, $H_z = 0$, TM_z Case) . . . 189
18.2 Rectangular Waveguides . . . 190
18.2.1 TE Modes ($H_z \neq 0$, H Modes or TE_z Modes) . . . 190

[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.

[4]Also known as TM_z polarization.

19 More on Hollow Waveguides **193**
19.1 Rectangular Waveguides, Contd. 194
19.1.1 TM Modes ($E_z \neq 0$, E Modes or TM_z Modes) 194
19.1.2 Bouncing Wave Picture 195
19.1.3 Field Plots 196
19.2 Circular Waveguides 198
19.2.1 TE Case 198
19.2.2 TM Case 200

20 More on Waveguides and Transmission Lines **205**
20.1 Circular Waveguides, Contd. 205
20.1.1 An Application of Circular Waveguide 206
20.2 Remarks on Quasi-TEM Modes, Hybrid Modes, and Surface Plasmonic Modes 209
20.2.1 Quasi-TEM Modes 209
20.2.2 Hybrid Modes–Inhomogeneously-Filled Waveguides 210
20.2.3 Guidance of Modes 211
20.3 Homomorphism of Waveguides and Transmission Lines 212
20.3.1 TE Case 212
20.3.2 TM Case 214
20.3.3 Mode Conversion 216

21 Cavity Resonators **219**
21.1 Transmission Line Model of a Resonator 219
21.2 Cylindrical Waveguide Resonators 221
21.2.1 $\beta_z = 0$ Case 223
21.2.2 Lowest Mode of a Rectangular Cavity 224
21.3 Some Applications of Resonators 225
21.3.1 Filters 226
21.3.2 Electromagnetic Sources 227
21.3.3 Frequency Sensor 230

22 Quality Factor of Cavities, Mode Orthogonality **233**
22.1 The Quality Factor of a Cavity–General Concept 233
22.1.1 Analogue with an LC Tank Circuit 234
22.1.2 Relation to Bandwidth and Pole Location 236
22.1.3 Wall Loss and **Q** for a Metallic Cavity 237
22.1.4 Example: The Q of TM_{110} Mode 239
22.2 Mode Orthogonality and Matrix Eigenvalue Problem 240
22.2.1 Matrix Eigenvalue Problem (EVP) 240
22.2.2 Homomorphism with the Waveguide Mode Problem 241
22.2.3 Proof of Orthogonality of Waveguide Modes[5] 242

[5]This may be skipped on first reading.

23 Scalar and Vector Potentials **245**
23.1 Scalar and Vector Potentials for Time-Harmonic Fields 245
23.2 Scalar and Vector Potentials for Statics, A Review 246
23.2.1 Scalar and Vector Potentials for Electrodynamics 247
23.2.2 More on Scalar and Vector Potentials 249
23.3 When is Static Electromagnetic Theory Valid? 250
23.3.1 Quasi-Static Electromagnetic Theory 255

24 Circuit Theory Revisited **257**
24.1 Kirchhoff Current Law 257
24.2 Kirchhoff Voltage Law 258
24.3 Inductor 262
24.4 Capacitance 263
24.5 Resistor 263
24.6 Some Remarks 264
24.7 Energy Storage Method for Inductor and Capacitor 264
24.8 Finding Closed-Form Formulas for Inductance and Capacitance 265
24.9 Importance of Circuit Theory in IC Design 267
24.10Decoupling Capacitors and Spiral Inductors 269

25 Radiation by a Hertzian Dipole **271**
25.1 History 271
25.2 Approximation by a Point Source 273
25.2.1 Case I. Near Field, $\boldsymbol{\beta r} \ll \mathbf{1}$ 275
25.2.2 Case II. Far Field (Radiation Field), $\boldsymbol{\beta r} \gg \mathbf{1}$ 276
25.3 Radiation, Power, and Directive Gain Patterns 276
25.3.1 Radiation Resistance 279

26 Radiation Fields **283**
26.1 Radiation Fields or Far-Field Approximation 284
26.1.1 Far-Field Approximation 285
26.1.2 Locally Plane Wave Approximation 286
26.1.3 Directive Gain Pattern Revisited 289

27 Array Antennas, Fresnel Zone, Rayleigh Distance **293**
27.1 Linear Array of Dipole Antennas 293
27.1.1 Far-Field Approximation 294
27.1.2 Radiation Pattern of an Array 295
27.2 When is Far-Field Approximation Valid? 298
27.2.1 Rayleigh Distance 299
27.2.2 Near Zone, Fresnel Zone, and Far Zone 300

28 Different Types of Antennas—Heuristics **303**
28.1 Resonance Tunneling in Antenna . . . 304
28.2 Horn Antennas . . . 307
28.3 Quasi-Optical Antennas . . . 309
28.4 Small Antennas . . . 311

29 Uniqueness Theorem **317**
29.1 The Difference Solutions to Source-Free Maxwell's Equations . . . 317
29.2 Conditions for Uniqueness . . . 320
29.2.1 Isotropic Case . . . 320
29.2.2 General Anisotropic Case . . . 321
29.3 Hind Sight Using Linear Algebra . . . 322
29.4 Connection to Poles of a Linear System . . . 323
29.5 Radiation from Antenna Sources and Radiation Condition . . . 324

30 Reciprocity Theorem **327**
30.1 Mathematical Derivation . . . 328
30.2 Conditions for Reciprocity . . . 331
30.3 Application to a Two-Port Network and Circuit Theory . . . 331
30.4 Voltage Sources in Electromagnetics . . . 333
30.5 Hind Sight . . . 334
30.6 Transmit and Receive Patterns of an Antennna . . . 335

31 Equivalence Theorems, Huygens' Principle **339**
31.1 Equivalence Theorems or Equivalence Principles . . . 339
31.1.1 Inside-Out Case . . . 340
31.1.2 Outside-in Case . . . 341
31.1.3 General Case . . . 341
31.2 Electric Current on a PEC . . . 342
31.3 Magnetic Current on a PMC . . . 343
31.4 Huygens' Principle and Green's Theorem . . . 343
31.4.1 Scalar Waves Case . . . 344
31.4.2 Electromagnetic Waves Case . . . 346

32 Shielding, Image Theory **349**
32.1 Shielding . . . 349
32.1.1 A Note on Electrostatic Shielding . . . 349
32.1.2 Relaxation Time . . . 350
32.2 Image Theory . . . 351
32.2.1 Electric Charges and Electric Dipoles . . . 352
32.2.2 Magnetic Charges and Magnetic Dipoles . . . 353
32.2.3 Perfect Magnetic Conductor (PMC) Surfaces . . . 354
32.2.4 Multiple Images . . . 355
32.2.5 Some Special Cases—Spheres, Cylinders, and Dielectric Interfaces . . . 356

33 High Frequency Solutions, Gaussian Beams **361**
33.1 Tangent Plane Approximations . . . 362
33.2 Fermat's Principle . . . 363
33.2.1 Generalized Snell's Law . . . 365
33.3 Gaussian Beam . . . 366
33.3.1 Derivation of the Paraxial/Parabolic Wave Equation . . . 366
33.3.2 Finding a Closed Form Solution . . . 367
33.3.3 Other solutions . . . 369

34 Scattering of Electromagnetic Field **371**
34.1 Rayleigh Scattering . . . 371
34.1.1 Scattering by a Small Spherical Particle . . . 373
34.1.2 Scattering Cross Section . . . 375
34.1.3 Small Conductive Particle . . . 378
34.2 Mie Scattering . . . 379
34.2.1 Optical Theorem . . . 380
34.2.2 Mie Scattering by Spherical Harmonic Expansions . . . 381
34.2.3 Separation of Variables in Spherical Coordinates[6] . . . 381

35 Spectral Expansions of Source Fields—Sommerfeld Integrals **383**
35.1 Spectral Representations of Sources . . . 383
35.1.1 A Point Source . . . 384
35.2 A Source on Top of a Layered Medium . . . 389
35.2.1 Electric Dipole Fields–Spectral Expansion . . . 389
35.3 Stationary Phase Method—Fermat's Principle . . . 392
35.4 Riemann Sheets and Branch Cuts[7] . . . 396
35.5 Some Remarks[8] . . . 396

36 Computational Electromagnetics, Numerical Methods **399**
36.1 Computational Electromagnetics, Numerical Methods . . . 401
36.2 Examples of Differential Equations . . . 401
36.3 Examples of Integral Equations . . . 402
36.3.1 Volume Integral Equation . . . 402
36.3.2 Surface Integral Equation . . . 404
36.4 Function as a Vector . . . 405
36.5 Operator as a Map . . . 406
36.5.1 Domain and Range Spaces . . . 406
36.6 Approximating Operator Equations with Matrix Equations . . . 407
36.6.1 Subspace Projection Methods . . . 407
36.6.2 Dual Spaces . . . 408
36.6.3 Matrix and Vector Representations . . . 408
36.6.4 Mesh Generation . . . 409

[6] May be skipped on first reading.
[7] This may be skipped on first reading.
[8] This may be skipped on first reading.

36.6.5 Differential Equation Solvers versus Integral Equation Solvers 410
36.7 Solving Matrix Equation by Optimization 410
36.7.1 Gradient of a Functional 411

37 Finite Difference Method, Yee Algorithm 415
37.1 Finite-Difference Time-Domain Method 415
37.1.1 The Finite-Difference Approximation 416
37.1.2 Time Stepping or Time Marching 418
37.1.3 Stability Analysis 420
37.1.4 Grid-Dispersion Error 422
37.2 The Yee Algorithm 424
37.2.1 Finite-Difference Frequency Domain Method 427
37.3 Absorbing Boundary Conditions 428

38 Quantum Theory of Light 431
38.1 Historical Background on Quantum Theory 431
38.2 Connecting Electromagnetic Oscillation to Simple Pendulum 434
38.3 Hamiltonian Mechanics 438
38.4 Schrödinger Equation (1925) 440
38.5 Some Quantum Interpretations–A Preview 443
38.5.1 Matrix or Operator Representations 444
38.6 Bizarre Nature of the Photon Number States 445

39 Quantum Coherent State of Light 447
39.1 The Quantum Coherent State 447
39.1.1 Quantum Harmonic Oscillator Revisited 448
39.2 Some Words on Quantum Randomness and Quantum Observables 450
39.3 Derivation of the Coherent States 451
39.3.1 Time Evolution of a Quantum State 453
39.4 More on the Creation and Annihilation Operator 454
39.4.1 Connecting Quantum Pendulum to Electromagnetic Oscillator[9] 457
39.5 Epilogue 460

[9]May be skipped on first reading.

Preface

This set of lecture notes is from my teaching of ECE 604, 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 endured and continued to be important, from statics to optics, from classical to quantum, and from nanometer 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, computer system, and quantum computer designs, quantum communication and many more. Electromagnetic field theory is not everything, but is remains an important 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. Some times, 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 some 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 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, as 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 them 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.

Weng Cho CHEW
December 3, 2020 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. Mark Lundstrom gave me useful feedback on Lectures 38 and 39 on the quantum theory of light. Also, I am thankful to Dong-Yeop Na for helping teach part of this course. Robert Hsueh-Yung Chao also took time to read the lecture notes and gave me very useful feedback.

Lecture 1

Introduction, Maxwell's Equations

In the beginning, this field is either known as electricity and magnetism or optics. But later, as we shall discuss, these two fields are found to be based on the same set equations known as 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.

It 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. 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, and has impacted a large sector of modern technologies.

To understand why electromagnetics is so useful, we have to understand 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, going from static to ultra-violet wavelengths.[1]

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

- 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 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].

- 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.[2] Now, it has been validated to one part in a trillion (Aoyama et al, Styer, 2012).[3]

- As a consequence, electromagnetics has had a tremendous impact in science and technology. This is manifested in electrical engineering, optics, wireless and optical communications, computers, remote sensing, bio-medical engineering etc.

[2]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.

[3]This means an error of a hairline, if one were to fly from the earth to the moon.

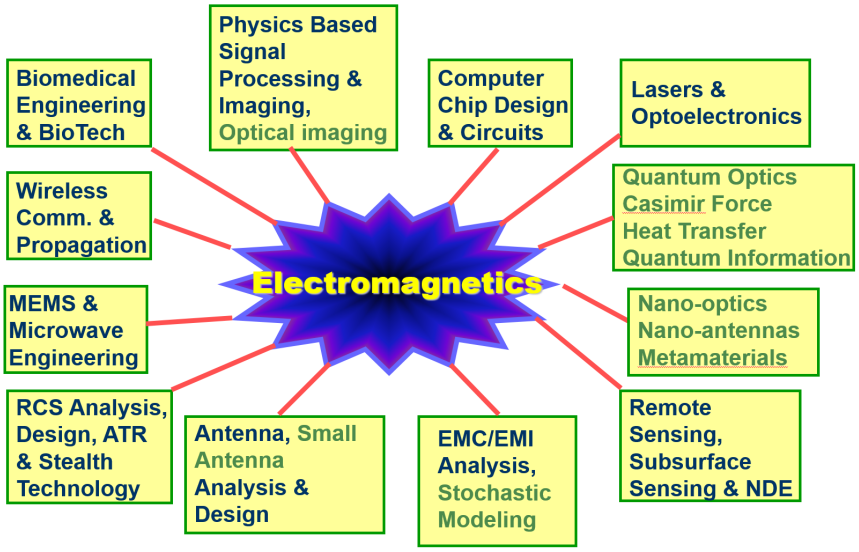


Figure 1.1: The impact of electromagnetics in many technologies. The areas in blue are prevalent areas impacted by electromagnetics some 20 years ago [9], and the areas in brown are modern emerging areas impacted by electromagnetics.

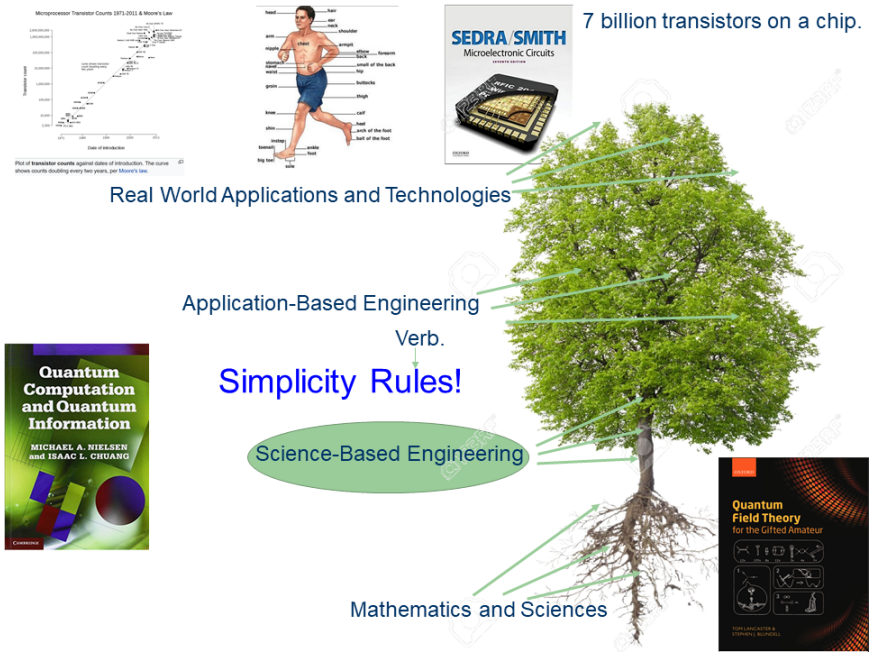


Figure 1.2: Knowledge grows like a tree. Engineering knowledge and real-world applications are driven by fundamental knowledge from math and sciences. At a university, we do science-based engineering research that can impact wide-ranging real-world applications. But everyone is equally important in transforming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

Figure 1.2 shows how knowledge are driven by basic math and science knowledge. Its growth is like a tree. It is important that we collaborate to develop technologies that can transform this world.

1.1.1 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 electricity and magnetism, now termed electromagnetics in the modern world, has been thought to be governed by different physical laws as opposed to 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.

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 are 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.[4]

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.

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 [18] did experiment to prove that electromagnetic field can propagate through space across a room. This illustrates the difficulty of knowledge dissemination when new knowledge is 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) [19]. The observation of interference phenomena in light has been known as well. When these pieces of information were pieced 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!

[4] 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.

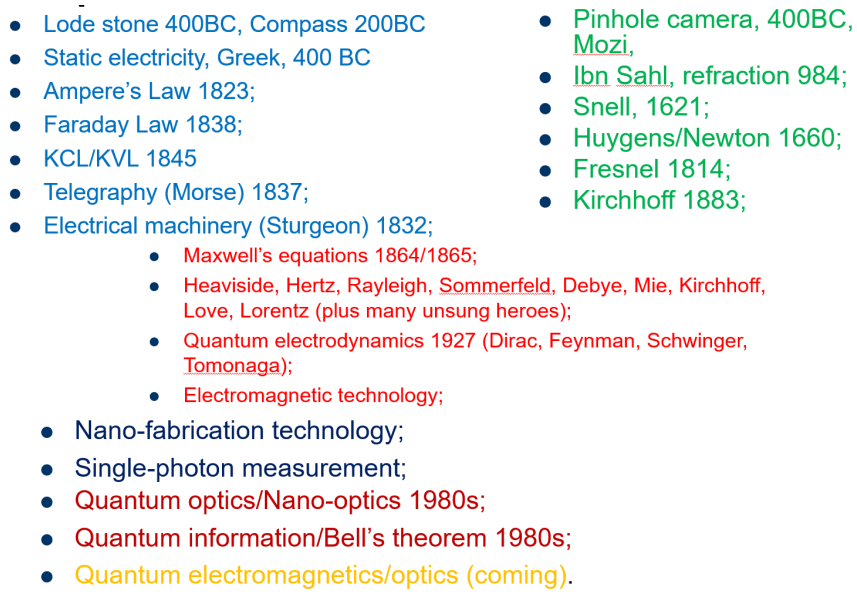


Figure 1.3: A brief history of electromagnetics and optics as depicted in this figure.

In Figure 1.3, 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 mathematical endeavor.

The photo-electric effect [20,21], and Planck radiation law [22] 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 $\hbar = 6.626 \times 10^{-34}$ J s, the Planck constant, 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 [23,24]. 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 optical photon.

The progress in nano-fabrication [25] allows one to make optical components that are subwavelength as the wavelength of blue light is about 450 nm. As a result, interaction of light with nano-scale optical components requires the solution of Maxwell's equations in its full glory.

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 controlled the 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 [26].

In 1960s, Bell's theorem (by John Steward Bell) [27] said that an inequality should be satisfied if Einstein's hidden variable theory was correct. Otherwise, the Copenhagen school of interpretation should prevail. However, experimental measurement showed that the inequality was violated, favoring the Copenhagen school of quantum interpretation [26]. 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, a quantum particle, such as a photon, its state is unknown until after its measurement. In other words, quantum theory is "spooky". 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. 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!"

1.2 Maxwell's Equations in Integral Form

Maxwell's equations can be presented as fundamental postulates.[5] We will present them in their integral forms, but will not belabor them until later.

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

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

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

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

The units of the basic quantities above are given as:

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

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

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

1.3 Static Electromagnetics

1.3.1 Coulomb's Law (Statics)

This law, developed in 1785 [28], 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

$$\mathbf{f}_{1\to 2} = \frac{q_1 q_2}{4\pi\varepsilon r^2}\hat{\mathbf{r}}_{12} \tag{1.3.1}$$

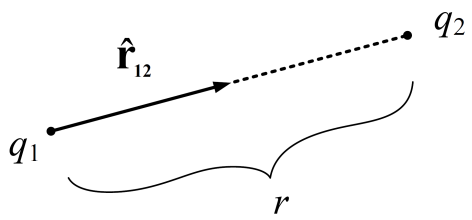


Figure 1.4: The force between two charges q_1 and q_2. The force is repulsive if the two charges have the same sign.

where the units are: $\mathbf{f}$ (force): newton
q (charge): coulombs
ε (permittivity): farads/meter
r (distance between q_1 and q_2): m
$\hat{\mathbf{r}}_{12}$= unit vector pointing from charge 1 to charge 2

$$\hat{\mathbf{r}}_{12} = \frac{\mathbf{r}_2 - \mathbf{r}_1}{|\mathbf{r}_2 - \mathbf{r}_1|}, \qquad r = |\mathbf{r}_2 - \mathbf{r}_1| \tag{1.3.2}$$

Since the unit vector can be defined in the above, the force between two charges can also be rewritten as

$$\mathbf{f}_{1\to 2} = \frac{q_1 q_2(\mathbf{r}_2 - \mathbf{r}_1)}{4\pi\varepsilon|\mathbf{r}_2 - \mathbf{r}_1|^3}, \qquad (\mathbf{r}_1, \mathbf{r}_2 \text{ are position vectors}) \tag{1.3.3}$$

1.3.2 Electric Field E (Statics)

The electric field $\mathbf{E}$ is defined as the force per unit charge [29]. 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 (1.3.1), is

$$\mathbf{f} = \frac{q\Delta q}{4\pi\varepsilon r^2}\hat{\mathbf{r}} \tag{1.3.4}$$

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

$$\mathbf{E} = \frac{\mathbf{f}}{\triangle q}, \qquad \text{(V/m)} \tag{1.3.5}$$

This electric field $\mathbf{E}$ from a point charge q at the orgin is hence

$$\mathbf{E} = \frac{q}{4\pi\varepsilon r^2}\hat{\mathbf{r}} \tag{1.3.6}$$

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

$$\mathbf{E}(\mathbf{r}) = \frac{q(\mathbf{r}-\mathbf{r}')}{4\pi\varepsilon|\mathbf{r}-\mathbf{r}'|^3} \tag{1.3.7}$$

where the unit vector

$$\hat{\mathbf{r}} = \frac{\mathbf{r}-\mathbf{r}'}{|\mathbf{r}-\mathbf{r}'|} \tag{1.3.8}$$

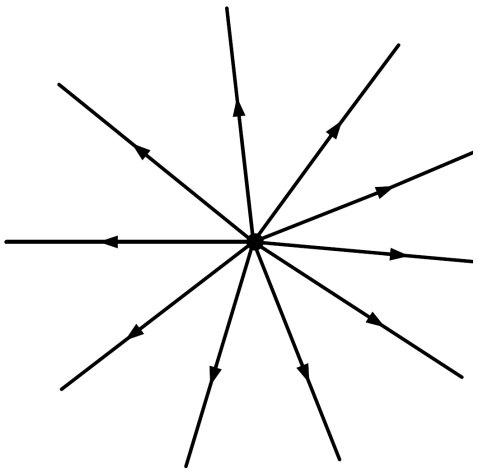


Figure 1.5: Emanating $\mathbf{E}$ field from an electric point charge as depicted by depicted by (1.3.7) and (1.3.6).

If one knows $\mathbf{E}$ due to a point charge, one will know $\mathbf{E}$ 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, the total field produced by these N charges is

$$\mathbf{E}(\mathbf{r}) = \sum_{i=1}^{N} \frac{q_i(\mathbf{r} - \mathbf{r}_i)}{4\pi\varepsilon|\mathbf{r} - \mathbf{r}_i|^3} \tag{1.3.9}$$

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, one gets

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

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 (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 [30] and later, reexpressed by Gauss in 1813 (Wikipedia).

This law can be expressed as

$$\oiint_S \mathbf{D} \cdot d\mathbf{S} = Q \tag{1.3.11}$$

where $\mathbf{D}$: electric flux density with unit $\mathrm{C/m^2}$ and $\mathbf{D} = \varepsilon \mathbf{E}$.

$d\mathbf{S}$: 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.

Q: total charge enclosed by the surface S.

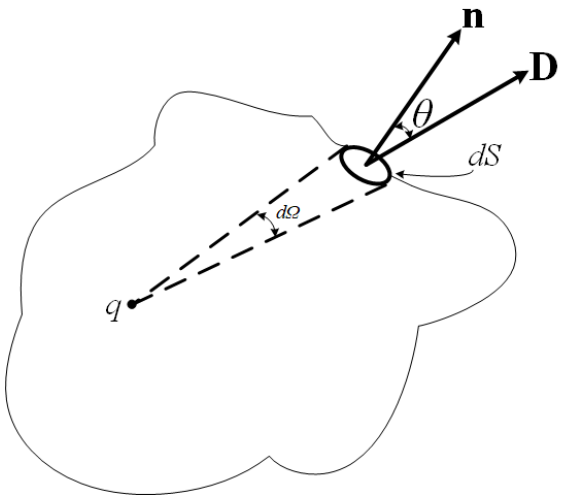


Figure 1.6: Electric flux (courtesy of Ramo, Whinnery, and Van Duzer [31]).

The left-hand side of (1.3.11) 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

$$\oiint_S \mathbf{D} \cdot d\mathbf{S} \approx \sum_i \mathbf{D}_i \cdot \hat{\mathbf{n}}_i \Delta S_i = \sum_i \mathbf{D}_i \cdot \Delta \mathbf{S}_i \tag{1.3.12}$$

where one has defined $\Delta \mathbf{S}_i = \hat{\mathbf{n}}_i \Delta S_i$. 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 and the ensuing electric field due to a point charge, the electric flux is

$$\mathbf{D} = \varepsilon\mathbf{E} = \frac{q}{4\pi r^2}\hat{\mathbf{r}} \tag{1.3.13}$$

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,

$$\oiint_S \mathbf{D}\cdot d\mathbf{S} = 4\pi r^2 \mathbf{D}\cdot\hat{\mathbf{n}} = 4\pi r^2 D_r = q \tag{1.3.14}$$

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

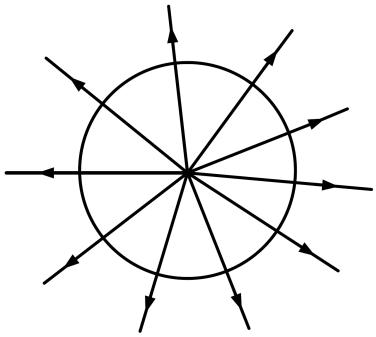


Figure 1.7: Electric flux from a point charge satisfies Gauss's law.

Even when the shape of the spherical surface S changes 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 1.8 are the same.

This can be appreciated by taking a sliver of the angular sector as shown in Figure 1.9. 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 weaker 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.

$$Q = \sum_i q_i \tag{1.3.15}$$

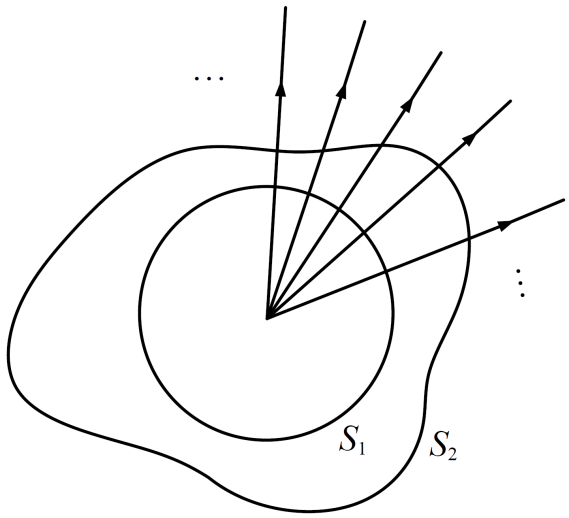


Figure 1.8: Same amount of electric flux from a point charge passes through two surfaces S_1 and S_2.

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.

1.4 Homework Examples

Example 1

Field of a ring of charge of density ϱ_l C/m.
Question: What is **E** along z axis? Hint: Use symmetry.

Example 2

Field between coaxial cylinders of unit length.
Question: What is **E**?
Hint: Use symmetry and cylindrical coordinates to express $\mathbf{E} = \hat{\rho} E_\rho$ and appply Gauss's law.

Example 3

Fields of a sphere of uniform charge density.
Question: What is **E**?
Hint: Again, use symmetry and spherical coordinates to express $\mathbf{E} = \hat{r} E_r$ and appply Gauss's law.

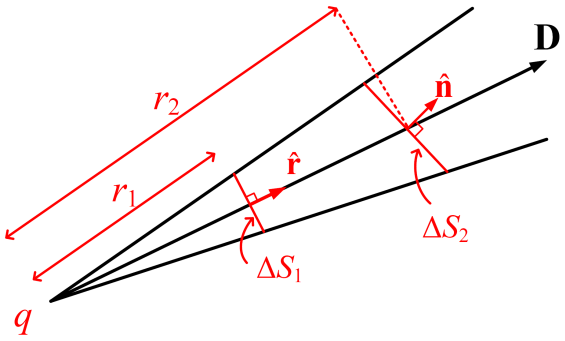


Figure 1.9: When a sliver of angular sector is taken, same amount of electric flux from a point charge passes through two incremental surfaces ΔS_1 and ΔS_2.

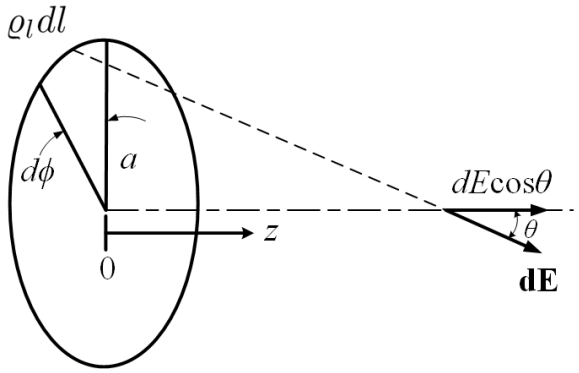


Figure 1.10: Electric field of a ring of charge (courtesy of Ramo, Whinnery, and Van Duzer [31]).

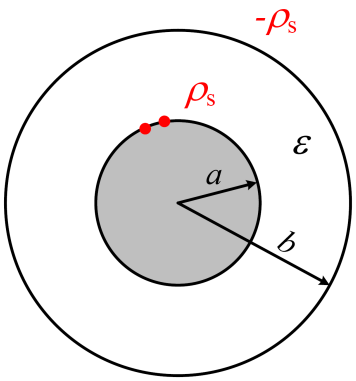


Figure 1.11: Figure for Example 2 for a coaxial cylinder.

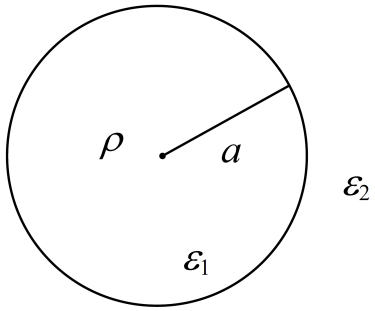


Figure 1.12: Figure for Example 3 for a sphere with uniform charge density.

Lecture 2

Maxwell's Equations, Differential Operator Form

Maxwell's equations were originally written in integral forms 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 forms[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

The divergence theorem is one of the most important theorems in vector calculus [31–34]. First, we will need to prove Gauss's divergence theorem, namely, that:

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

In the above, $\nabla \cdot \mathbf{D}$ 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 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 as shown on the right-hand side. As shall be shown, the ratio has a limit and eventually, we will find an expression for it. We know that if $\Delta V \approx 0$ or small, then the

[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.

above,

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

First, we assume that a volume V has been discretized[2] into a sum of small cuboids, where the i-th cuboid has a volume of ΔV_i as shown in Figure 2.1. Then

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

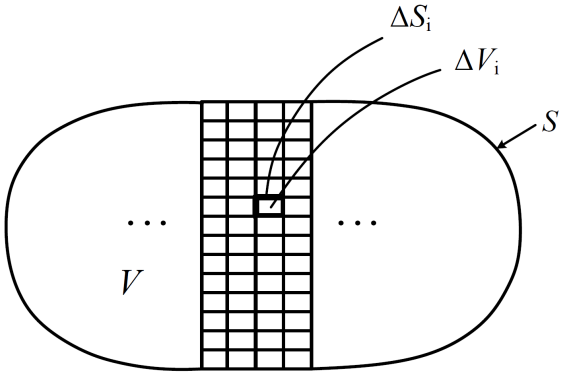


Figure 2.1: The discretization of a volume V into a sum of small volumes ΔV_i each of which is a small cuboid. Stair-casing error occurs near the boundary of the volume V but the error diminishes as $\Delta V_i \to 0$.

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

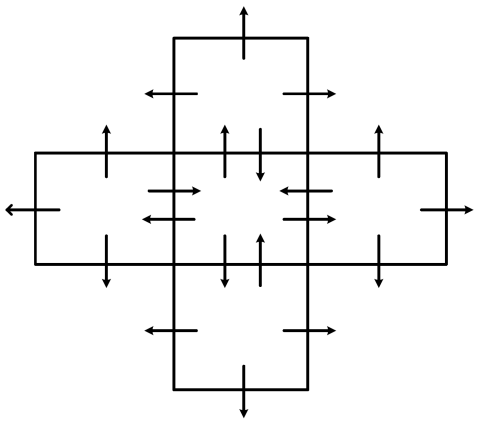


Figure 2.2: Fluxes from adjacent cuboids cancel each other leaving only the fluxes at the boundary that remain uncancelled. Please imagine that there is a third dimension of the cuboids in this picture where it comes out of the paper.

Then from (2.1.2),

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

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

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

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 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 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.

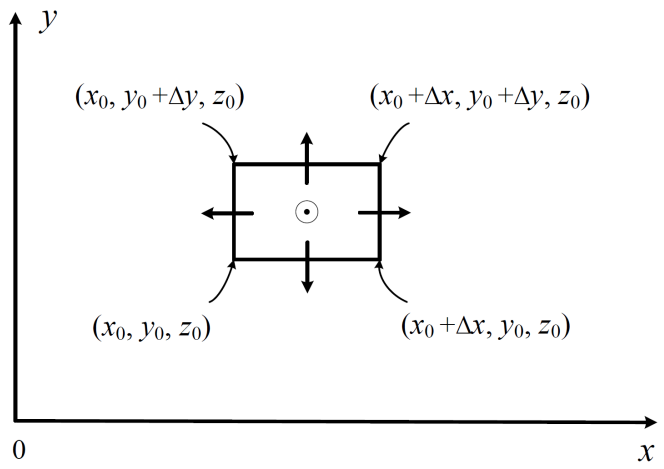


Figure 2.3: Figure to illustrate the calculation of fluxes from a small cuboid where a corner of the cuboid is located at (x_0, y_0, z_0). There is a third z dimension of the cuboid not shown, and coming out of the paper. Hence, this cuboid, unlike that shown in the figure, has six faces.

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

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

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

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

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

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.

Consequently, one gets 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 Gauss's Law in Differential Operator Form

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

$$\oiint_S \mathbf{D} \cdot d\mathbf{S} = Q = \iiint dV \varrho \tag{2.1.15}$$

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

$$\iiint_V dV \nabla \cdot \mathbf{D} = \iiint_V dV \varrho \tag{2.1.16}$$

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

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

2.1.3 Physical Meaning of Divergence Operator

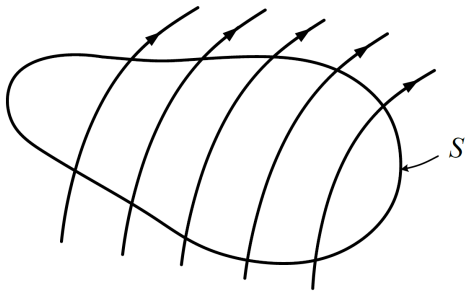
The physical meaning of divergence is that if $\nabla\cdot\mathbf{D} \neq 0$ at a point in space, it implies that there are fluxes oozing or exuding from that point in space [35]. On the other hand, if $\nabla \cdot \mathbf{D} = 0$, it implies no flux oozing out from that point in space. In other words, whatever flux that goes into the point must come out of it. The flux is 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

$$\oiint_S \mathbf{D} \cdot d\mathbf{S} = 0 \tag{2.1.18}$$

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, electric flux flow in a source free region, as well as magnetic flux flow, where the flux is conserved.



$$\nabla \cdot \mathbf{D} = 0 \quad \Rightarrow \oint_S \hat{\mathrm{n}} \cdot \mathbf{D} \mathrm{d}S = 0$$

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

2.2 Stokes's Theorem

The mathematical description of fluid flow was well established before the establishment of electromagnetic theory [36]. Hence, much mathematical description of electromagnetic theory uses the language of fluid. In mathematical notations, Stokes's theorem is

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

In the above, the contour C is a closed contour, whereas the surface S is not closed.[3]

First, applying Stokes's theorem to a small surface ΔS, we define a curl operator[4] $\nabla\times$ 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, $\nabla \times \mathbf{E}$ is a vector. Taking $\oint_{\Delta C} \mathbf{E} \cdot d\mathbf{l}$ as a measure of the 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 $\nabla \times \mathbf{E}$.

[3] 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.

[4] Sometimes called a rotation operator.

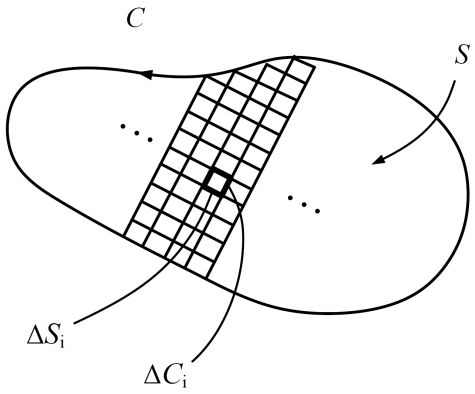


Figure 2.5: In proving Stokes's theorem, a closed contour C is assumed to enclose an open surface S. Then the surface S is tessellated into sum of small rects as shown. Stair-casing error vanishes in the limit when the rects are made vanishingly small.

First, the 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. Stokes's theorem is then 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 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

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

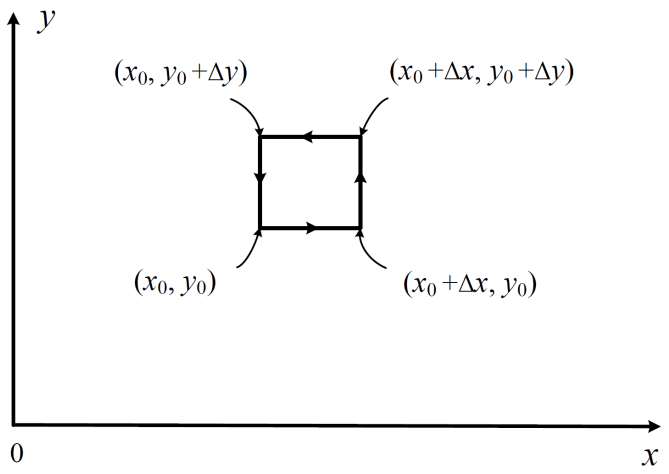


Figure 2.6: We approximate the integration over a small rect using this figure. There are four edges to this small rect.

Next, we need to prove the details of definition (2.2.2) using Figure 2.6. Performing the integral over the small rect, one gets

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

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 $|hatn = \hat{x}$ 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

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

where

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

2.2.1 Faraday's Law in Differential Operator Form

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.

Faraday's law in integral form is given by[5]

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

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

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

One can replace the left-hand side with the use of Stokes' theorem to arrive at

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

The normal of the surface element $d\mathbf{S}$ can be pointing in an arbitrary direction, and the surface S can be made very small. Then the integral can be removed, and one has

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

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

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

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

[5]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.

2.2.2 Physical Meaning of Curl Operator

The curl operator $\nabla\times$ is a measure of the rotation or the circulation of a field at a point in space. 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. 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.

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 theorem, 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 2.7.

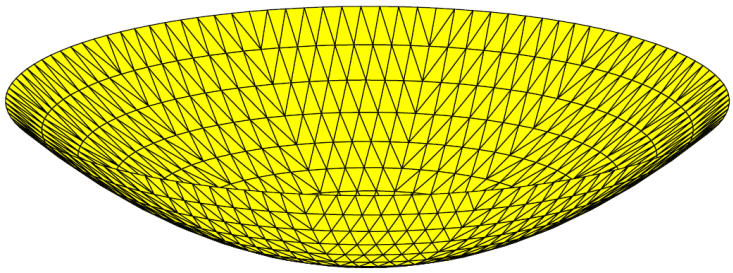


Figure 2.7: An arbitrary curved surface can be triangulated with flat triangular patches. The triangulation can be made arbitrarily accurate by making the patches arbitrarily small.

2.3 Maxwell's Equations in Differential Operator Form

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

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

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

$$\nabla\cdot\mathbf{D} = \varrho \tag{2.3.3}$$

$$\nabla\cdot\mathbf{B} = 0 \tag{2.3.4}$$

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, 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 Homework Examples

Example 1

If $\mathbf{D} = (2y^2 + z)\hat{x} + 4xy\hat{y} + x\hat{z}$, find:

1. Volume charge density ρ_v at $(-1, 0, 3)$.
2. Electric flux through the cube defined by

$$0 \leq x \leq 1, 0 \leq y \leq 1, 0 \leq z \leq 1.$$

3. Total charge enclosed by the cube.

Example 2

Suppose $\mathbf{E} = \hat{\mathbf{x}}3y + \hat{\mathbf{y}}x$, calculate $\int \mathbf{E} \cdot d\boldsymbol{l}$ along a straight line in the x-y plane joining (0,0) to (3,1).

2.5 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, 37, 38]. It was Oliver Heaviside who distilled those equations and presented them in the present form found in textbooks. Putatively, most cannot read his treatise [37] beyond the first 50 pages [39].
- Maxwell wrote many poems in his short lifespan (1831-1879) and they can be found at [40].
- 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 would be named Clerk Maxwell. That was why Maxwell is addressed as Professor Clerk Maxwell.

Lecture 3

Constitutive Relations, Wave Equation, Electrostatics, 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 that derivation of this phenomenon here. To make matter simple, we will reduce the problem to electrostatics to simplify the math to introduce the concept of the Green's function.

As mentioned previously, for time-varying problems, only the first two of the four Maxwell's equations suffice. The latter two are derivable from the first two. But the first two equations have four unknowns **E**, **H**, **D**, and **B**. Hence, two more equations are needed to solve for these unknowns. These equations come from the constitutive relations.

3.1 Simple Constitutive Relations

The constitution relation between electric flux **D** and the electric field **E** in free space (or vacuum) is

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

When material medium is present, one has to add the contribution to **D** by the polarization density **P** which is a dipole density.[1] Then [31, 32, 41]

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

[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 **D**.

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 due to 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 [42]. 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.

For many media, they are approximately linear media. Then $\mathbf{P}$ is linearly proportional to $\mathbf{E}$, or $\mathbf{P} = \varepsilon_0\chi_0\mathbf{E}$, or

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

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,[2]

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

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

$$\mathbf{B} = \mu\mathbf{H}, \qquad \mu = \text{permeability H/m} \tag{3.1.5}$$

In free space or vacuum,

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

As shall be explained later, this 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 for 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.

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 [18], some 23 years after the

[2]It is to be noted that we will use MKS unit in this course. Another possible unit is the CGS unit used in many physics texts [43]

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.[3] 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

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

In the above, the left-hand side can be simplified by using the identity that $\mathbf{a} \times (\mathbf{b} \times \mathbf{c}) = \mathbf{b}(\mathbf{a} \cdot \mathbf{c}) - \mathbf{c}(\mathbf{a} \cdot \mathbf{b})$,[4] 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

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

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 [32, 44].

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

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

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

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

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

[4] For mnemonics, this formula is also known as the "back-of-the-cab" formula.

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 [45]. From the above factorization, it is seen that the solutions of these one-way wave equations are also the solutions of the original wave equation given by (3.2.8). Their general solutions are then

$$f_+(z,t) = F_+(z - c_0 t) \tag{3.2.14}$$
$$f_-(z,t) = F_-(z + c_0 t) \tag{3.2.15}$$

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 of any shape (see Figure 3.1); they can be used to encode information for communication!

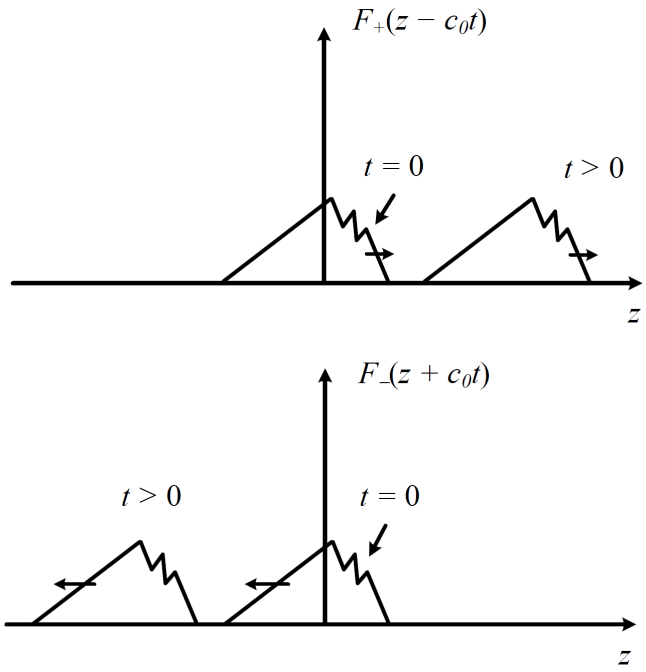


Figure 3.1: Solutions of the wave equation can be a single-valued function of any shape. In the above, F_+ travels in the positive z direction, while F_- travels in the negative z direction as t increases.

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 $c_0 = \sqrt{1/\mu_0\varepsilon_0}$.

Maxwell's equations (3.2.1) implies that **E** and **H** are linearly proportional to each other. Thus, there is only one independent constant in the wave equation, and 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 defined to be the distance traveled by light in $1/(299792458)$ seconds. Hence, the more accurate that unit of time or second can be calibrated, the more accurate can we calibrate the unit of length or meter. Therefore, the design of an accurate clock like an atomic clock is an important research problem.

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.[5] This

[5]The velocity of light was known in astronomy by (Roemer, 1676) [19].

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 [31, 32, 46]

$$\nabla \times \mathbf{E} = 0 \tag{3.3.1}$$
$$\nabla \times \mathbf{H} = \mathbf{J} \tag{3.3.2}$$
$$\nabla \cdot \mathbf{D} = \varrho \tag{3.3.3}$$
$$\nabla \cdot \mathbf{B} = 0 \tag{3.3.4}$$

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 effect is negligible. In other words, there is no time delay 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 the solution in the static regime or the long-wavelength regime.

3.3.1 Electrostatics

We see that Faraday's law in the static limit is

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

One way to satisfy the above is to let $\mathbf{E} = -\nabla\Phi$ because of the identity $\nabla \times \nabla = 0$.[6] Alternatively, one can assume that $\mathbf{E}$ is a constant. But we usually are interested in solutions that vanish at infinity, and hence, the latter is not a viable solution. Therefore, we let

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

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

3.3.2 Poisson's Equation

As a consequence of the above,

$$\nabla \cdot \mathbf{D} = \varrho \Rightarrow \nabla \cdot \varepsilon \mathbf{E} = \varrho \Rightarrow -\nabla \cdot \varepsilon \nabla \Phi = \varrho \tag{3.3.8}$$

In the last equation above, if ε is a constant of space, or independent of $\mathbf{r}$, then one arrives at the simple Poisson's equation, which is a partial differential equation

$$\nabla^2 \Phi = -\frac{\varrho}{\varepsilon} \tag{3.3.9}$$

Here, the Laplacian operator

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

For a point source, we know from Coulomb's law that

$$\mathbf{E} = \frac{q}{4\pi\varepsilon r^2}\hat{r} = -\nabla\Phi \tag{3.3.10}$$

From the above, we deduce that[7]

$$\Phi = \frac{q}{4\pi\varepsilon r} \tag{3.3.11}$$

Therefore, we know the solution to Poisson's equation (3.3.9) when the source ϱ represents a point source. 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.

A point source located at $\mathbf{r}'$ is described by a charge density as

$$\varrho(\mathbf{r}) = q\delta(\mathbf{r} - \mathbf{r}') \tag{3.3.12}$$

where $\delta(\mathbf{r}-\mathbf{r}')$ is a short-hand notation for $\delta(x-x')\delta(y-y')\delta(z-z')$. Therefore, from (3.3.9), the corresponding partial differential equation for a point source is

$$\nabla^2 \Phi(\mathbf{r}) = -\frac{q\delta(\mathbf{r} - \mathbf{r}')}{\varepsilon} \tag{3.3.13}$$

The solution to the above equation, from Coulomb's law in accordance to (3.3.11), has to be

$$\Phi(\mathbf{r}) = \frac{q}{4\pi\varepsilon|\mathbf{r} - \mathbf{r}'|} \tag{3.3.14}$$

whereas (3.3.11) is for a point source at the origin, (3.3.14) is for a point source located and translated to $\mathbf{r}'$. The above is a coordinate independent form of the solution. Here, $\mathbf{r} = \hat{x}x + \hat{y}y + \hat{z}z$ and $\mathbf{r}' = \hat{x}x' + \hat{y}y' + \hat{z}z'$, and $|\mathbf{r} - \mathbf{r}'| = \sqrt{(x-x')^2 + (y-y')^2 + (z-z')^2}$.

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

3.3.3 Static Green's Function

Let us define a partial differential equation given by

$$\nabla^2 g(\mathbf{r}-\mathbf{r}') = -\delta(\mathbf{r}-\mathbf{r}') \tag{3.3.15}$$

The above is similar to Poisson's equation with a point source on the right-hand side as in (3.3.13). Such a solution, a response to a point source, is called the Green's function.[8] By comparing equations (3.3.13) and (3.3.15), then making use of (3.3.14), we deduced that the static Green's function is

$$g(\mathbf{r}-\mathbf{r}') = \frac{1}{4\pi|\mathbf{r}-\mathbf{r}'|} \tag{3.3.16}$$

An arbitrary source can be expressed as

$$\varrho(\mathbf{r}) = \iiint_V dV' \varrho(\mathbf{r}')\delta(\mathbf{r}-\mathbf{r}') \tag{3.3.17}$$

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 (3.3.9), we have

$$\nabla^2\Phi(\mathbf{r}) = -\frac{1}{\varepsilon}\iiint_V dV' \varrho(r')\delta(\mathbf{r}-\mathbf{r}') \tag{3.3.18}$$

We can let

$$\Phi(\mathbf{r}) = \frac{1}{\varepsilon}\iiint_V dV' g(\mathbf{r}-\mathbf{r}')\varrho(\mathbf{r}') \tag{3.3.19}$$

By substituting the above into the left-hand side of (3.3.18), exchanging order of integration and differentiation, and then making use of equation (3.3.15), it can be shown that (3.3.19) indeed satisfies (3.3.9). The above is just a convolutional integral. Hence, the potential $\Phi(\mathbf{r})$ due to an arbitrary source distribution $\varrho(\mathbf{r})$ can be found by using convolution, namely,

$$\Phi(\mathbf{r}) = \frac{1}{4\pi\varepsilon}\iiint_V \frac{\varrho(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|} dV' \tag{3.3.20}$$

In a nutshell, the solution of Poisson's equation when it is driven by an arbitrary source ϱ, is the convolution of the source with the static Green's function, a point source response.

3.3.4 Laplace's Equation

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

$$\nabla^2\Phi = 0 \tag{3.3.21}$$

which is the Laplace's equation. Laplace's equation is usually solved as a boundary value problem. In such a problem, the potential Φ is stipulated on the boundary of a region with a certain boundary condition, and then the solution is sought in the region so as to match the boundary condition.

Examples of such boundary value problems are given at the end of the lecture.

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

3.4 Homework Examples

Example 1

Fields of a sphere of radius a with uniform charge density ρ:

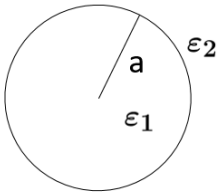


Figure 3.2: Figure of a sphere with uniform charge density for the example above.

Assuming that $\Phi|_{r=\infty} = 0$, what is Φ at $r \leq a$? And Φ at $r > a$.

Example 2

A capacitor has two parallel plates attached to a battery, what is **E** field inside the capacitor?

First, one guess the electric field between the two parallel plates. Then one arrive at a potential Φ in between the plates so as to produce the field. Then the potential is found so as to match the boundary conditions of $\Phi = V$ in the upper plate, and $\Phi = 0$ in the lower plate. What is the Φ that will satisfy the requisite boundary condition?

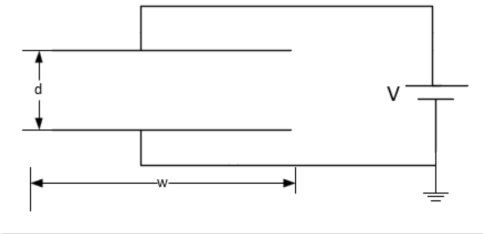


Figure 3.3: Figure of a parallel plate capacitor. The field in between can be found by solving Laplace's equation as a boundary value problem [31].

Example 3

A coaxial cable has two conductors. The outer conductor is grounded and hence is at zero potential. The inner conductor is at voltage V. What is the solution?

For this, one will have to write the Laplace's equation in cylindrical coordinates, namely,

$$\nabla^2\Phi = \frac{1}{\rho}\frac{\partial}{\partial\rho}\left(\rho\frac{\partial\Phi}{\partial\rho}\right) + \frac{1}{\rho^2}\frac{\partial^2\Phi}{\partial\phi^2} = 0 \tag{3.4.1}$$

In the above, we assume that the potential is constant in the z direction, and hence, $\partial/\partial z = 0$, and ρ, ϕ, z are the cylindrical coordinates. By assuming axi-symmetry, we can let $\partial/\partial\phi = 0$ and the above becomes

$$\nabla^2\Phi = \frac{1}{\rho}\frac{\partial}{\partial\rho}\left(\rho\frac{\partial\Phi}{\partial\rho}\right) = 0 \tag{3.4.2}$$

Show that $\Phi = A\ln\rho + B$ is a general solution to Laplace's equation in cylindrical coordinates inside a coax. What is the Φ that will satisfy the requisite boundary condition?

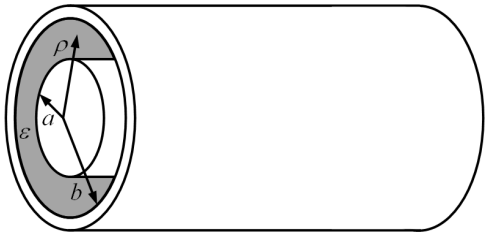


Figure 3.4: The field in between a coaxial line can also be obtained by solving Laplace's equation as a boundary value problem (courtesy of Ramo, Whinnery, and Van Duzer [31]).

Lecture 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 are derived from Maxwell's equations. Maxwell's equations can be first solved in different domains. Then the solutions are pieced (or sewn) together by imposing boundary conditions at the boundaries or interfaces of the domains. 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 $\partial/\partial t = 0$, which are [31, 32, 46]

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

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

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

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

because of the vector identity

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

The above is zero for the same reason that $\mathbf{a} \cdot (\mathbf{a} \times \mathbf{b}) = 0$. 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,

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

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 back-of-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 $\nabla \cdot \nabla$. As a result, we arrive at [47]

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

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. 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 $\mathbf{r}$. Here, $dV' = dx'dy'dz'$. It is also variously written as $d\mathbf{r}'$ or $d^3\mathbf{r}'$.

[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 [48], 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 $\mathbf{B}$; hence, $\mathbf{A}$ is non-unique.

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

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

Then

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

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

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 what $\mathbf{b} \cdot \mathbf{a}$ is as well. Here, $\mathbf{a}$, $\mathbf{b}$, and $\mathbf{c}$ are independent vectors.

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. 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 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 [49].

[2] 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 [31] is useful for proving this assertion.

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 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. 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.

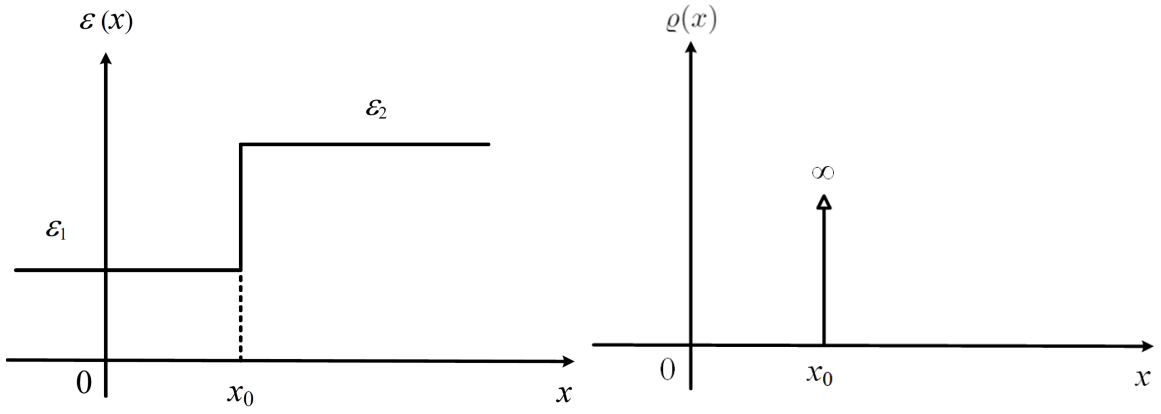


Figure 4.1: A figure showing a charge sheet at the interface between two dielectric media. Because it is a surface charge sheet, the volume charge density $\varrho(x)$ is infinite at the sheet location x_0.

To see this in greater detail, we illustrate it with the following example. In the above, $\varrho(x)$ represents a 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

$$\int_{x_0-\Delta}^{x_0+\Delta} dx \left[\frac{d}{dx}\varepsilon(x)\frac{d}{dx}\Phi(x)\right] = -\int_{x_0-\Delta}^{x_0+\Delta} dx\varrho(x) = -\int_{x_0-\Delta}^{x_0+\Delta} dx\varrho_0\delta(x - x_0) \tag{4.2.2}$$

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

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

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

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.

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

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

The above implies 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

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

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 whose derivatives will yield the delta function on the right-hand side.

4.3 Boundary Conditions–Maxwell's Equations

As seen previously, boundary conditions for a field is embedded in the 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 [31,32,47]. To derive these boundary conditions, we will take an unconventional view: namely to see what sources can induce jump conditions 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.

4.3.1 Faraday's Law

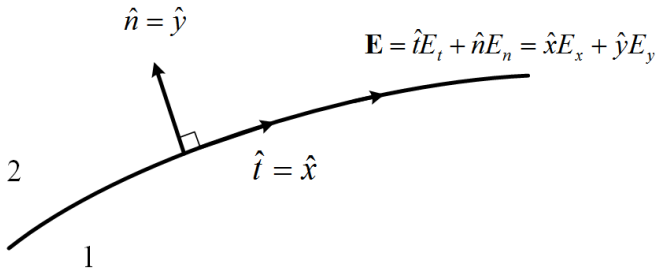


Figure 4.2: This figure is for the derivation of boundary condition induced by Faraday's law. A local coordinate system can be used to see the boundary condition more lucidly. Here, the normal $\hat{n} = \hat{y}$ and the tangential component $\hat{t} = \hat{x}$.

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, and it 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 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 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 bring out the tangential component of a vector $\mathbf{E}$ (convince yourself).

4.3.2 Gauss's Law for Electric Flux

From this Gauss's law, we have

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

where ϱ is the volume charge density.

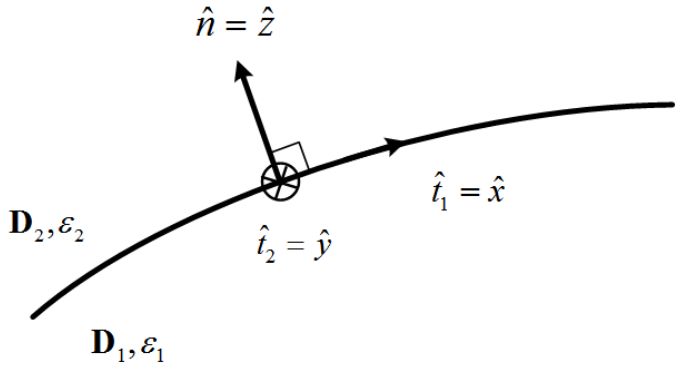


Figure 4.3: A figure showing the derivation of boundary condition for Gauss's law. Again, a local coordinate system can be introduced for simplicity.

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. By looking at the above expression, the only term that can produce a $\delta(z)$ is from $\frac{\partial}{\partial z} D_z$. In other words, D_z has 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 $D_z(0^+) = D_{2n}$, $D_z(0^-) = D_{1n}$, the above becomes

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

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

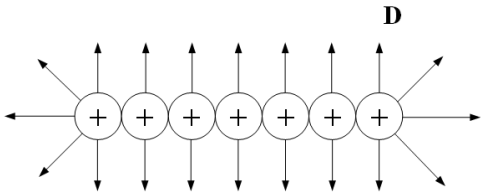


Figure 4.4: A figure intuitively showing why a sheet of charge gives rise to a jump discontinuiy in the normal component of the electric flux $\mathbf{D}$.

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 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 coordinates just as we did for Faraday's law.

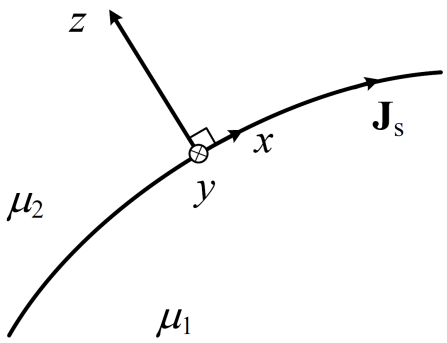


Figure 4.5: A figure showing the derivation of boundary condition for Ampere's law. A local coordinate system is used for simplicity.

To be general, we also include the presence of a current sheet at the interface. 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,

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

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, 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, 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

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

[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}.

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 combined as one, 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, $\hat{n} \times \mathbf{H}$. This is illustrated intuitively in Figure 4.6.

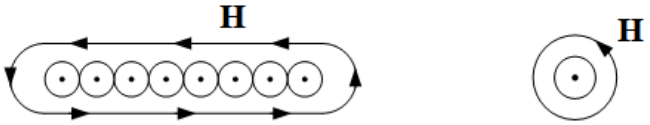


Figure 4.6: A figure intuitively showing that with the understanding of how a single line current source generates a magnetic field (right), a cluster of them forming a current sheet will generate a jump discontinuity in the tangential component of the magnetic field **H** (left).

4.3.4 Gauss's Law for Magnetic Flux

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

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

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. If there singular terms in Maxwell's equations, then via the differential operators, the boundary conditions can be deduced. These boundary conditions are also known as jump condition if a current or a source sheet is present.

Lecture 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 [31, 32, 48]. This is the cumulative work of Ampere, Oersted, Biot, and Savart. At this stage of the course, one has the mathematical tool to derive this law from Ampere's law and Gauss's law for magnetostatics. 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

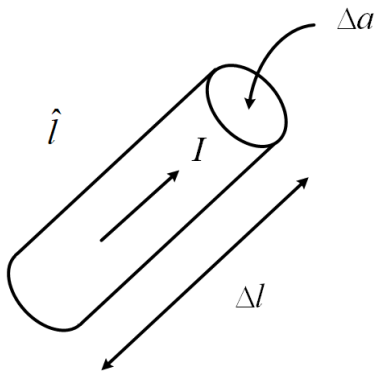


Figure 5.1: A current element used to illustrate the derivation of Biot-Savart law. The current element generates a magnetic field due to Ampere's law in the static limit. This law was established experimentally, but here, we will derive this law based on our mathematical knowledge so far.

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 $R = |\mathbf{r} - \mathbf{r}'|$. When the current element is small, and is carried by a wire of cross sectional area Δa as shown in Figure 5.1, we can approximate the integrand as

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

In the above, $\Delta V = (\Delta a)\Delta l$ and $\hat{l} I/\Delta a = \mathbf{J}(\mathbf{r}')$ 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. 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 $\mathbf{J}(\mathbf{r}')\Delta V'$ 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 the 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 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 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. We have also made use of the fact that $R = \sqrt{(x-x')^2 + (y-y')^2 + (z-z')^2}$. Consequently, assuming that the incremental length becomes infinitesimally small, or $\Delta\mathbf{l} \to \mathbf{dl}$, we have, after using (5.1.6) in (5.1.5), that

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

Since $\mathbf{B} = \mu\mathbf{H}$, we have

$$\mathbf{dH} = \frac{I\mathbf{dl}' \times \hat{R}}{4\pi R^2} \tag{5.1.8}$$

or

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

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

[1] This is best done by expressing the ∇ operator in spherical coordinates.

5.2 Shielding by Conductive Media

5.2.1 Boundary Conditions–Conductive Media Case

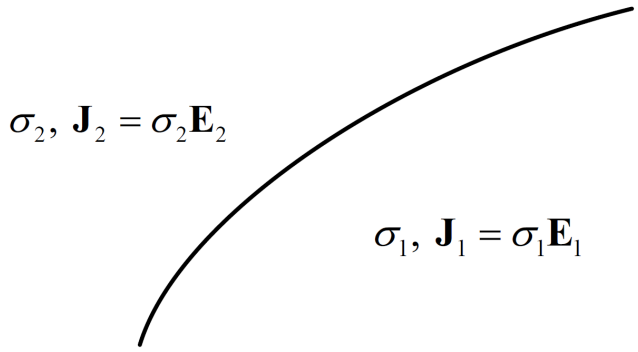


Figure 5.2: The schematics for deriving the boundary condition for the current density $\mathbf{J}$ at the interface of two conductive media.

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

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

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,

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

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

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

The above has to be always true in the static limit irrespective of the values of σ_1 and σ_2. 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 is incompatible with (5.2.3) unless $\varrho_s \neq 0$. Hence, surface charge density or charge accumulation is necessary at the 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

The electric field inside a perfect electric conductor (PEC) has to be zero by the explanation as follows. 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.

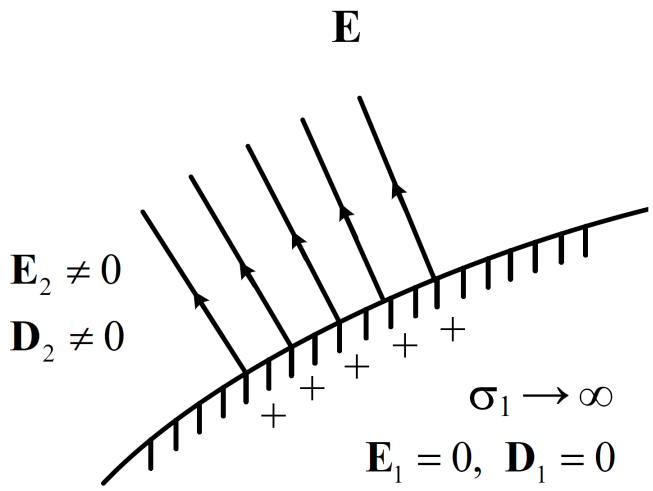


Figure 5.3: The behavior of the electric field and electric flux at the interface of a perfect electric conductor and free space (or air).

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 density has to be nonzero at a PEC/air interface. Moreover, normal $\mathbf{D}_2 \neq 0$, 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.

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. 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. Therefore, the boundary condition that the fields have to be normal to the conducting object surface is still true for elecrostatics. A sketch of the electric field between two conducting spheres is show in Figure 5.4.

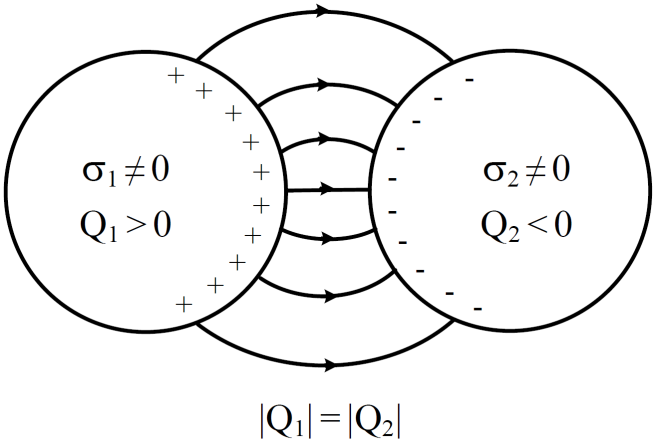


Figure 5.4: The behavior of the electric field and flux outside two conductors in the static limit. The two conductors need not be PEC, and yet, the fields are normal to the interface.

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 $\mathbf{E} = 0$ or $\partial_t \mathbf{E} = 0$. In a word, static **E** is zero inside a conductor.

But there are no magnetic charges nor magnetic conductors in this world. Thus this physical phenomenon does not happen for magnetic field: In other words, static magnetic field cannot be expelled from an 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 **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 **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 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, $\hat{n} \cdot \mathbf{B}$ is always continuous at an interface because of the absence of magnetic charges. The magnetic flux $\mathbf{B}_1$ is expelled from the perfect conductor making $\hat{n} \cdot \mathbf{B}_1 = 0$ zero both inside and outside the PEC (due to the continuity of normal $\mathbf{B}$ at an interface). And hence, there is no normal component of the $\mathbf{B}_1$ field at the interface. Therefore, the boundary condition for $\mathbf{B}_2$ becomes, for a PEC,

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

The $\mathbf{B}$ field in the vicinity of a perfect conductor surface is as shown in Figure 5.5.

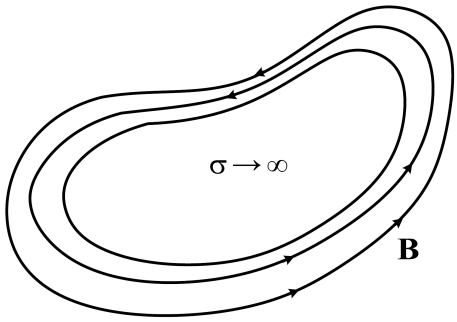


Figure 5.5: Sketch of the magnetic flux $\mathbf{B}$ around a perfect electric conductor. It is seen that $\hat{n} \cdot \mathbf{B} = 0$ at the surface of the perfect conductor.

When a superconductor cube is placed next to a static magnetic field near a permanent magnet, eddy current will be induced on the superconductor. The eddy current will expel the static 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, and cause the superconducting cube to levitate on the static magnetic field as shown in Figure 5.6.

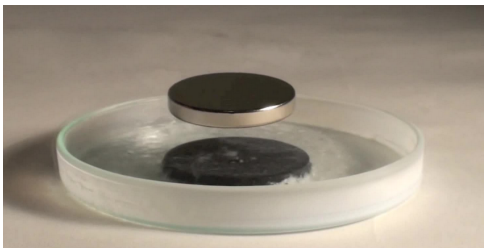


Figure 5.6: Levitation of a superconducting disk on top of a static magnetic field due to expulsion of the magnetic field from the superconductor.. This is also known as the Meissner effect (figure courtesy of Wikimedia).

5.3 Instantaneous Poynting's Theorem

Before we proceed further with studying energy and power, 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.[2] To this end, we have

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

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

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

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

Consider the first two of Maxwell's equations where fictitious magnetic current is included and that the medium is isotropic such that $\mathbf{B} = \mu\mathbf{H}$ and $\mathbf{D} = \varepsilon\mathbf{E}$. 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

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

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

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 [50].

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. We can show from By dot multiplying (5.3.5) with $\mathbf{H}$, and

[2]Even though magnetic current does not exist, electric current can be engineered to look like magnetic current as shall be learnt.

dot multiplying (5.3.6) with $\mathbf{E}$, we can show that

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

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

Using the identity, which is the same as the product rule for derivatives, we have[3]

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

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

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

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

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

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

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

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

where $|\mathbf{H}(\mathbf{r},t)|^2=\mathbf{H}(\mathbf{r},t)\cdot\mathbf{H}(\mathbf{r},t)$, and $|\mathbf{E}(\mathbf{r},t)|^2=\mathbf{E}(\mathbf{r},t)\cdot\mathbf{E}(\mathbf{r},t)$. 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

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

Equation (5.3.14) is reminiscent of the current continuity equation, namely that,

$$\nabla\cdot\mathbf{J}=-\frac{\partial\varrho}{\partial t} \tag{5.3.16}$$

which is a statement of charge conservation. In other words, time variation of current density at a point is due to charge density flow into or out of the point.

The vector quantity

$$\mathbf{S}_p=\mathbf{E}\times\mathbf{H} \tag{5.3.17}$$

[3]The cyclical identity 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.

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

$$\nabla \cdot \mathbf{S}_p = -\frac{\partial}{\partial t} W_t \tag{5.3.18}$$

where $W_t = W_e + W_m$ is the total energy density stored in the electric and magnetic fields. The above is similar to the current continuity equation in physical interpretation mentioned above. Analogous to that current density is charge density flow, power density is energy density flow. It is easy to show that $\mathbf{S}_p$ has a dimension of watts per meter square, or power density, and that W_t has a dimension of joules per meter cube, or energy density.

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 $\sigma \neq 0$, (5.3.18) becomes

$$\nabla \cdot \mathbf{S}_p = -\frac{\partial}{\partial t} W_t - \sigma |\mathbf{E}|^2 = -\frac{\partial}{\partial t} W_t - P_d \tag{5.3.19}$$

Here, $\nabla \cdot \mathbf{S}_p$ has 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, (5.3.19) is augmented by the last two terms in (5.3.10), or

$$\nabla \cdot \mathbf{S}_p = -\frac{\partial}{\partial t} W_t - P_d - \mathbf{H} \cdot \mathbf{M}_i - \mathbf{E} \cdot \mathbf{J}_i \tag{5.3.20}$$

The last two terms can be interpreted as the power density supplied by the impressed currents $\mathbf{M}_i$ and $\mathbf{J}_i$. Therefore, (5.3.20) becomes

$$\nabla \cdot \mathbf{S}_p = -\frac{\partial}{\partial t} W_t - P_d + P_s \tag{5.3.21}$$

where

$$P_s = -\mathbf{H} \cdot \mathbf{M}_i - \mathbf{E} \cdot \mathbf{J}_i \tag{5.3.22}$$

Here, P_s is the power supplied by the impressed current sources. These terms are positive if $\mathbf{H}$ and $\mathbf{M}_i$ have opposite signs, or if $\mathbf{E}$ and $\mathbf{J}_i$ have opposite signs. The last terms reminds us of what happens in a negative resistance device or a battery.[4] In a battery, positive charges move from a region of lower potential to a region of higher potential (see Figure 5.7). 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.

[4] A negative resistance has been made by Leo Esaki [51], winning him a share in the Nobel prize.

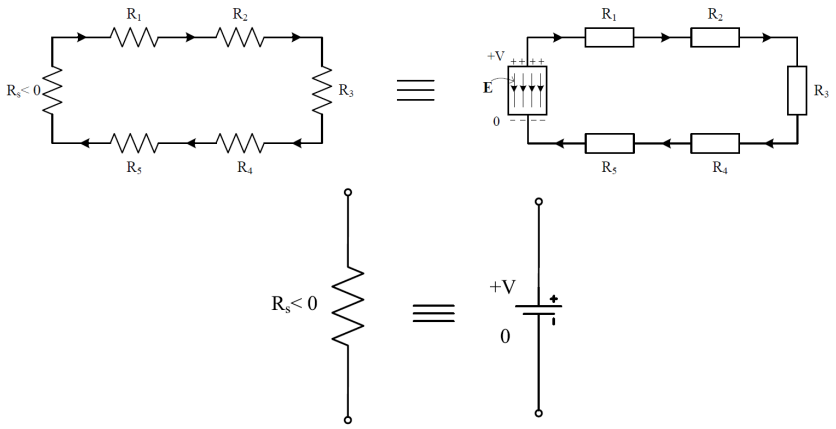


Figure 5.7: Figure showing the dissipation of energy as the current flows around a loop. A battery can be viewed as having a negative resistance.

In the above, one can easily work out that P_s has the unit of W m^{-3} which is power supplied density. One can also choose to rewrite (5.3.21) in integral form by integrating it over a volume V and invoking the divergence theorem yielding

$$\int_S d\mathbf{S} \cdot \mathbf{S}_p = -\frac{d}{dt} \int_V W_t dV - \int_V P_d dV + \int_V P_s dV \tag{5.3.23}$$

The left-hand side is

$$\int_S d\mathbf{S} \cdot \mathbf{S}_p = \int_S d\mathbf{S} \cdot (\mathbf{E} \times \mathbf{H}) \tag{5.3.24}$$

which represents the power flowing out of the surface S.

Lecture 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). Electrical engineers use a method called phasor technique [32, 52], to simplify equations involving time-harmonic signals. This is also a poor-man's Fourier transform [53]. That is one begets the benefits of Fourier transform technique without knowledge of Fourier transform. Since only one time-harmonic frequency is involved, this is also called frequency domain analysis.[1] Phasors are represented in complex numbers. 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.

6.1 Time-Harmonic Fields—Linear Systems

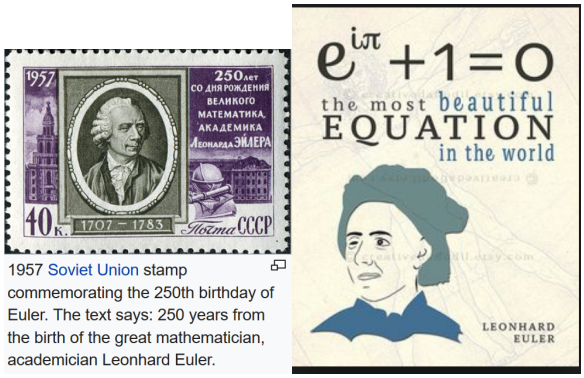


Figure 6.1: A commemorative stamp showing the contribution of Euler (courtesy of Wikipedia and Pinterest).

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

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

where $j = \sqrt{-1}$ is an imaginary number.[2]

From Euler's formula one gets

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

Hence, all time harmonic quantity 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.[3] Here, the phase $\alpha = \alpha(\mathbf{r})$ can also be a function of position $\mathbf{r}$, or x,y,z. Consequently, any component of a field can be expressed as

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

[2]But lo and behold, in other disciplines, $\sqrt{-1}$ 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.

[3]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.

The above can be repeated for y and z components. Compactly, 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. 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. Therefore, 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

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

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

using (6.1.11), and (6.1.14), into (6.1.12), one gets

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

and that

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

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

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

Hence, finding the phasor representation of an equation 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

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

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

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

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

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 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 of the phasors are different from the dimensions of the Fourier-transformed fields: $\underset{\sim}{\mathbf{E}}(\mathbf{r})$ and $\mathbf{E}(\mathbf{r}, \omega)$ do not have the same dimension on closer examination.

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. Now, we can use phasor technique to analyze the above. Assuming 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

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

Then 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

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

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

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

Hence, 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 . The power flow associated with it is termed complex power.

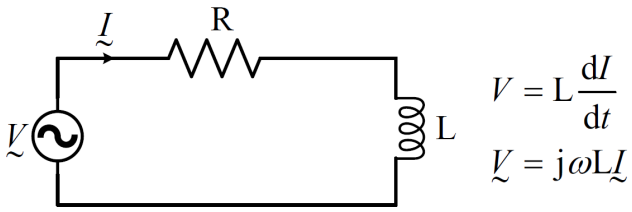


Figure 6.2: A simple circuit example to illustrate the concept of complex power in circuit theory.

To understand what complex power is , it is fruitful if we revisit complex power [50, 54] 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 is taken to be

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

But the instantaneous power is given by

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

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,

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

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

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

It can be seen that the first term does not time-average to zero, but the second term, by letting $\cos(\omega t)\sin(\omega t) = 0.5\sin(2\omega t)$, does time-average to zero. Now taking the time average of (6.3.15), we get

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

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

On the other hand, the reactive power

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

One sees that amplitude of the time-varying term in (6.3.15) is precisely proportional to $\Im m\left[\underset{\sim}{P}\right]$.[4]

The reason for the existence of imaginary part of $\underset{\sim}{P}$ is because $V(t)$ and $I(t)$ are out of phase or $\underset{\sim}{V} = V_0$, but $\underset{\sim}{I} = I_0 e^{j\alpha}$. The reason why they are out of phase is because the circuit has a reactive part to it. Hence the imaginary part of complex power is also called the reactive power [35,50,54]. 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. This means that 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

[4]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.

to our home, the power is complex because the current and voltage are not in phase. Even though the reactive power time-averages to zero, the power company still needs to deliver it to and from our home to run our washing machine, dish washer, fans, and air conditioner etc, and hence, charges us for it. Part of this power will be dissipated in the transmission lines that deliver power to our home. In other words, we have to pay to the use of imaginary power!

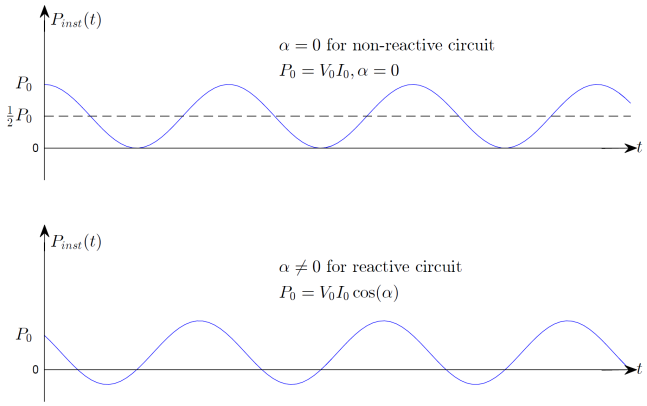


Figure 6.3: Plots of instantaneous power for when the voltage and the current is in phase ($\alpha = 0$), and when they are out of phase ($\alpha \neq 0$). In the out-of-phase case, there is an additional time-varying term that does not contribute to time-average power as shown in (6.3.15).

Lecture 7

More on Constitute Relations, Uniform Plane Wave

As mentioned before, constitutive relations are important for us to solve only the first two of four Maxwell's equations. Assuming that **J** is known or zero, then the first two vector equations have four vector unknowns: **E**, **H**, **D**, and **B**, which cannot be determined by solving only two equations. The addition two equations come from te 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, 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.,

The use of phasors allows us to study wave phenomena in these media quite easily. Hence, we will also study uniform plane wave in such media, including a 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 into depth more into various kinds of constitutive relations. Now that we have learned phasor technique, 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 [31], [32][p. 82], [48]

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

We have a simple model where

$$\mathbf{P} = \varepsilon_0 \chi_0 \mathbf{E} \tag{7.1.2}$$

where χ_0 is the electric susceptibility. When used in the generalized Ampere's law, $\mathbf{P}$, the polarization density, plays an important role for the flow of the displacement current through space. The polarization density is due to the presence of polar atoms or molecules that become little dipoles in the presence of an electric field. For instance, water, which 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 as capacitive coupling between the dipoles yielding polarization current flow 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 line up end to end, one can see a current 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 capacitor.

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 [52], or that at any given $\mathbf{r}$ [35][p. 212], [48][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

$$\mathbf{P}(\mathbf{r}, \omega) = \varepsilon_0 \chi_0(\mathbf{r}, \omega) \mathbf{E}(\mathbf{r}, \omega), \tag{7.1.5}$$

or

$$\mathbf{D}(\mathbf{r}, \omega) = \varepsilon_0 \left[1 + \chi_0(\mathbf{r}, \omega)\right] \mathbf{E}(\mathbf{r}, \omega) = \varepsilon(\mathbf{r}, \omega) \mathbf{E}(\mathbf{r}, \omega) \tag{7.1.6}$$

where $\varepsilon(\mathbf{r}, \omega) = \varepsilon_0 \left[1 + \chi_0(\mathbf{r}, \omega)\right]$ at any point $\mathbf{r}$ in space. There is a rich variety of ways at which $\chi_0(\omega)$ can manifest itself. Such a permittivity $\varepsilon(\mathbf{r}, \omega)$ is often called the effective permittivity. Such media where the effective permittivity is a function of frequency is termed dispersive media, or frequency dispersive media.

7.1.2 Anisotropic Media

For anisotropic media [32][p. 83]

$$\begin{aligned}\mathbf{D} &= \varepsilon_0\mathbf{E} + \varepsilon_0\overline{\boldsymbol{\chi}}_0(\omega)\cdot\mathbf{E}\\ &= \varepsilon_0\left[\overline{\mathbf{I}} + \overline{\boldsymbol{\chi}}_0(\omega)\right]\cdot\mathbf{E} = \overline{\boldsymbol{\varepsilon}}(\omega)\cdot\mathbf{E}\end{aligned} \tag{7.1.7}$$

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 often associated with a physical notion like the relation between two physical fields, whereas a matrix is not.)

Previously, we have assume that χ_0 to be frequency independent. This is not usually the case as all materials have χ_0'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 $\mathbf{D}$. This implies that in the time-domain, the above relation becomes a time-convolution relation as in (7.1.4).

Similarly for conductive media,

$$\mathbf{J} = \sigma\mathbf{E}, \tag{7.1.8}$$

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

$$\nabla\times\mathbf{H}(\mathbf{r}) = j\omega\varepsilon\mathbf{E}(\mathbf{r}) + \sigma\mathbf{E}(\mathbf{r}) = j\omega\underset{\sim}{\varepsilon}(\omega)\mathbf{E}(\mathbf{r}) \tag{7.1.9}$$

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 (7.1.9) 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. This is one of the strengths of phasor technique.

For anisotropic conductive media, one has

$$\mathbf{J} = \overline{\boldsymbol{\sigma}}(\omega)\cdot\mathbf{E}, \tag{7.1.10}$$

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 relationship between the input and the output of a linear system. This need not be so. In fact, the most general linear relationship between $\mathbf{P}(\mathbf{r},t)$ and $\mathbf{E}(\mathbf{r},t)$ is

$$\mathbf{P}(\mathbf{r},t) = \int_{-\infty}^{\infty}\iiint_{-\infty}^{\infty}\overline{\boldsymbol{\chi}}(\mathbf{r}-\mathbf{r}',t-t')\cdot\mathbf{E}(\mathbf{r}',t')d\mathbf{r}'dt' \tag{7.1.11}$$

The above is a general convolutional relationship in both space and time. In the Fourier transform space, by taking Fourier transform in both space and time, the above becomes

$$\mathbf{P}(\mathbf{k},\omega) = \overline{\boldsymbol{\chi}}(\mathbf{k},\omega)\cdot\mathbf{E}(\mathbf{k},\omega) \tag{7.1.12}$$

where

$$\overline{\boldsymbol{\chi}}(\mathbf{k},\omega) = \int_{-\infty}^{\infty} \overline{\boldsymbol{\chi}}(\mathbf{r},t)\exp(j\mathbf{k}\cdot\mathbf{r} - j\omega t)d\mathbf{r}dt \tag{7.1.13}$$

(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 [35][p. 6], [55]. In general[1]

$$\overline{\boldsymbol{\varepsilon}}(\mathbf{k},\omega) = 1 + \overline{\boldsymbol{\chi}}(\mathbf{k},\omega) \tag{7.1.14}$$

where

$$\mathbf{D}(\mathbf{k},\omega) = \overline{\boldsymbol{\varepsilon}}(\mathbf{k},\omega)\cdot\mathbf{E}(\mathbf{k},\omega) \tag{7.1.15}$$

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

$$\mathbf{B}(\mathbf{k},\omega) = \overline{\boldsymbol{\mu}}(\mathbf{k},\omega)\cdot\mathbf{H}(\mathbf{k},\omega) \tag{7.1.16}$$

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 on the electric field $\mathbf{E}$ and the magnetic flux $\mathbf{B}$, on the magnetic field $\mathbf{H}$. The concept of constitutive relation can be extended to where $\mathbf{D}$ and $\mathbf{B}$ depend on both $\mathbf{E}$ and $\mathbf{H}$. In general, one can write

$$\mathbf{D} = \overline{\boldsymbol{\varepsilon}}(\omega)\cdot\mathbf{E} + \overline{\boldsymbol{\xi}}(\omega)\cdot\mathbf{H} \tag{7.1.17}$$

$$\mathbf{B} = \overline{\boldsymbol{\zeta}}(\omega)\cdot\mathbf{E} + \overline{\boldsymbol{\mu}}(\omega)\cdot\mathbf{H} \tag{7.1.18}$$

A 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 [32][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 known as an inhomogeneous medium or a heterogeneous medium. There are usually no simple solutions to problems associated with such media [35].

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 [56–58]. They are optical materials where the permittivity tensor can be written as

$$\overline{\boldsymbol{\varepsilon}} = \begin{pmatrix} \varepsilon_1 & 0 & 0 \\ 0 & \varepsilon_2 & 0 \\ 0 & 0 & \varepsilon_3 \end{pmatrix} \tag{7.1.19}$$

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

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, 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 behavior through such a medium.

7.1.6 Nonlinear Media

In the previous cases, we have assumed that $\overline{\boldsymbol{\chi}}_0$ is independent of the field $\mathbf{E}$. The relationships between $\mathbf{P}$ and $\mathbf{E}$ can be written more generally as

$$\mathbf{P} = \varepsilon_0 \overline{\boldsymbol{\chi}}_0(\mathbf{E}) \cdot \mathbf{E} \tag{7.1.20}$$

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 [59–61]. Kerr optical nonlinearity, discovered in 1875, was one of the earliest nonlinear phenomena observed [32, 56, 59].

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

$$\mathbf{B} = \overline{\boldsymbol{\mu}}(\mathbf{H}) \cdot \mathbf{H} \tag{7.1.21}$$

This nonlinearity is important even at low frequencies, and in electric machinery designs [62, 63], and magnetic resonance imaging systems [64]. The large permeability in magnetic materials is usually due to the formation of magnetic domains which can only happen at low frequencies.

7.2 Wave Phenomenon in the Frequency Domain

We have seen the emergence of wave phenomenon in the time domain. Given the simplicity of the frequency domain method, it will be interesting to ask how this phenomenon presents itself for time-harmonic field or in the frequency domain. In the frequency domain, the source-free Maxwell's equations are [32][p. 429], [65][p. 107]

$$\nabla \times \mathbf{E}(\mathbf{r}) = -j\omega\mu_0 \mathbf{H}(\mathbf{r}) \tag{7.2.1}$$

$$\nabla \times \mathbf{H}(\mathbf{r}) = j\omega\varepsilon_0 \mathbf{E}(\mathbf{r}) \tag{7.2.2}$$

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

$$\nabla \times \nabla \times \mathbf{E}(\mathbf{r}) = -j\omega\mu_0 \nabla \times \mathbf{H}(\mathbf{r}) = \omega^2 \mu_0 \varepsilon_0 \mathbf{E}(\mathbf{r}) \tag{7.2.3}$$

Again, using the 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

$$(\nabla^2 + \omega^2 \mu_0 \varepsilon_0)\mathbf{E}(\mathbf{r}) = 0 \tag{7.2.5}$$

This is known as the Helmholtz wave equation or just the Helmholtz equation.[2]

For simplicity of seeing the wave phenomenon, we 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

$$\left(\frac{d^2}{dz^2} + k_0^2\right) E_x(z) = 0 \tag{7.2.6}$$

where $k_0^2 = \omega^2 \mu_0 \varepsilon_0 = \omega^2/c_0^2$ where $c_0 = 1/\sqrt{\mu_0 \varepsilon_0}$ is the velocity of light. The general solution to (7.2.6) is of the form

$$E_x(z) = E_{0+} e^{-jk_0 z} + E_{0-} e^{jk_0 z} \tag{7.2.7}$$

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

$$E_x(z,t) = |E_{0+}| \cos(\omega t - k_0 z + \alpha_+) + |E_{0-}| \cos(\omega t + k_0 z + \alpha_-) \tag{7.2.8}$$

where we have assumed 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

$$\cos(\omega t \mp k_0 z + \alpha_\pm) = \cos\left[\frac{\omega}{c_0}(c_0 t \mp z) + \alpha_\pm\right] \tag{7.2.10}$$

where we have used the fact that $k_0 = \frac{\omega}{c_0}$. 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_0. The above plane wave is uniform and a constant in the xy plane and 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 $\cos(\mp k_0 z + \alpha_\pm)$. This is a periodic function in z with period $2\pi/k_0$ which is the wavelength λ_0, or that

$$k_0 = \frac{2\pi}{\lambda_0} = \frac{\omega}{c_0} = \frac{2\pi f}{c_0} \tag{7.2.11}$$

One can see that because c_0 is a humongous number in electromagnetics, λ_0 can be very large. You can plug in the frequency of your local AM 920 station, operating at 920 KHz, to see how big λ_0 is.

The above analysis still holds true if ϵ_0 and μ_0 are replaced by ϵ and μ but are real numbers. In this case, the velocity c of the wave is now given by $c = 1/\sqrt{\mu\epsilon}$. This velocity is the velocity of the phase of a time-harmonic signal and hence, is known as phase velocity.

[2] For a comprehensive review of this topic, one may read the lecture notes [44].

7.3 Uniform Plane Waves in 3D

By repeating the previous derivation for a homogeneous lossless medium, the vector Helmholtz equation for a source-free medium is given by [44]

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

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. And upon substituting (7.3.3) into (7.3.2), it is seen that

$$k_x^2 + k_y^2 + k_z^2 = \omega^2 \mu\varepsilon \tag{7.3.4}$$

This is called the dispersion relation for a plane wave.

In general, k_x, k_y, and k_z can be arbitrary and even complex numbers as long as this relation 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 7.2). This flat plane represents the constant phase wave front. By increasing the constant, we obtain different planes for progressively changing phase fronts.[3]

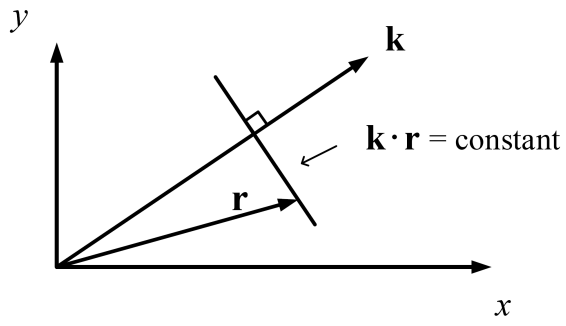


Figure 7.2: A figure showing the geometrical meaning of $\mathbf{k} \cdot \mathbf{r}$ equal to a constant. It is a flat plane that defines the wavefront of a plane wave.

[3]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.

Further, since $\nabla \cdot \mathbf{E} = 0$, we have

$$\begin{aligned}\nabla \cdot \mathbf{E} &= \nabla \cdot \mathbf{E}_0 e^{-jk_x x - jk_y y - jk_z z} = \nabla \cdot \mathbf{E}_0 e^{-j\mathbf{k}\cdot\mathbf{r}} \\ &= (-\hat{x}jk_x - \hat{y}jk_y - \hat{z}jk_z) \cdot \mathbf{E}_0 e^{-j\mathbf{k}\cdot\mathbf{r}} \\ &= -j(\hat{x}k_x + \hat{y}k_y + \hat{z}k_z) \cdot \mathbf{E} = 0 \end{aligned} \tag{7.3.5}$$

or that

$$\mathbf{k} \cdot \mathbf{E}_0 = \mathbf{k} \cdot \mathbf{E} = 0 \tag{7.3.6}$$

Thus, $\mathbf{E}$ is 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 do the substitution that $\nabla \to -j\mathbf{k}$. Hence, in a source-free homogenous medium,

$$\nabla \times \mathbf{E} = -j\omega\mu\mathbf{H} \tag{7.3.7}$$

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. Then $\mathbf{E}_0$, $\mathbf{H}_0$ and $\mathbf{k}$ form a right-handed 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 system.) Such a wave, where the electric field and magnetic field are transverse to the direction of propagation, is called a transverse electromagnetic (TEM) wave.

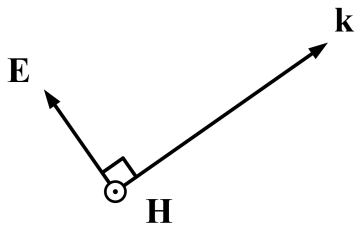


Figure 7.3: The $\mathbf{E}$, $\mathbf{H}$, and $\mathbf{k}$ together form a right-hand coordinate system, obeying the right-hand rule. Namely, $\mathbf{E} \times \mathbf{H}$ points in the direction of $\mathbf{k}$.

Figure 7.3 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 7.3 also shows, as $\mathbf{k}$, $\mathbf{E}$, and $\mathbf{H}$ are orthogonal to each other.

Since in general, $\mathbf{E}_0$ and $\mathbf{H}_0$ can be complex vectors, because they are phasors, we need to show the more general case. From (7.3.9), one can show, using the "back-of-the-cab" formula, that

$$\mathbf{E} \times \mathbf{H}^* = \mathbf{E} \cdot \mathbf{E}^* \frac{\mathbf{k}}{\omega\mu} = |\mathbf{E}|^2 \frac{\mathbf{k}}{\omega\mu} \tag{7.3.12}$$

But $\mathbf{E} \times \mathbf{H}^*$ is the direction of power flow, and it is in fact in the $\mathbf{k}$ direction assuming that $\mathbf{k}$ is a real vector. This is also required by the Poynting's theorem.

Furthermore, we can show in general that

$$|\mathbf{H}| = \frac{|\mathbf{k}||\mathbf{E}|}{\omega\mu} = \sqrt{\frac{\varepsilon}{\mu}}|\mathbf{E}| = \frac{1}{\eta}|\mathbf{E}| \tag{7.3.13}$$

where the quantity

$$\eta = \sqrt{\frac{\mu}{\varepsilon}} \tag{7.3.14}$$

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 is real, and it 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 $c = 1/\sqrt{\mu_0\varepsilon}$. 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.[4]

[4]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.

Lecture 8

Lossy Media, Lorentz Force Law, Drude-Lorentz-Sommerfeld Model

In the previous lecture, we see 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. In other words, in a conductive, 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 source-free Ampere's law in the frequency domain for a lossless medium where ε is completely real. Using the same method as before, a wave solution

$$\mathbf{E} = \mathbf{E}_0 e^{-j\mathbf{k}\cdot\mathbf{r}} \tag{8.1.3}$$

will have the dispersion relation which is now given by

$$k_x^2 + k_y^2 + k_z^2 = \omega^2 \mu \underset{\sim}{\varepsilon} \tag{8.1.4}$$

Since $\underset{\sim}{\varepsilon}$ is complex now, k_x, k_y, and k_z cannot be all real. Equation (8.1.4) 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 a complex $\mathbf{k}$ vector that is suppose to indicate the direciton with which the wave is propagating. Here, the wave can decay and oscillate in different directions.

So again, for physical insight, we look at the simplified case where

$$\mathbf{E} = \hat{x} E_x(z) \tag{8.1.5}$$

so that $\nabla \cdot \mathbf{E} = \partial_x E_x(z) = 0$, and let $\mathbf{k} = \hat{z}k = \hat{z}\omega\sqrt{\mu\underset{\sim}{\varepsilon}}$. 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,

$$\mathbf{E} = \hat{x}\mathbf{E}_x(z) = \hat{x} E_0 e^{-jkz} \tag{8.1.6}$$

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 gives rise to

$$\mathbf{H} = \frac{\mathbf{k}\times\mathbf{E}}{\omega\mu} = \hat{y}\frac{kE_x(z)}{\omega\mu} = \hat{y}\sqrt{\frac{\underset{\sim}{\varepsilon}}{\mu}}E_x(z) \tag{8.1.7}$$

where $\mathbf{k}$ vector is defined shortly after (8.1.5) above, and $k = \omega\sqrt{\mu\underset{\sim}{\varepsilon}}$, a complex number. It is seen that $\mathbf{H} = \hat{y}H_y$, and that

$$E_x/H_y = \sqrt{\frac{\mu}{\underset{\sim}{\varepsilon}}} \tag{8.1.8}$$

8.1.1 Highly Conductive Case

When the medium is highly conductive, $\sigma \to \infty$, and $\underset{\sim}{\varepsilon} = \epsilon - j\frac{\sigma}{\omega} \approx -j\frac{\sigma}{\omega}$. In other words, when $|\frac{\sigma}{\omega}| \gg \varepsilon$, the conduction current dominates over displacement current. Thus, the following approximation can be made, namely,

$$k = \omega\sqrt{\mu\underset{\sim}{\varepsilon}} \simeq \omega\sqrt{-\mu\frac{j\sigma}{\omega}} = \sqrt{-j\omega\mu\sigma} \tag{8.1.9}$$

Taking $\sqrt{-j} = \frac{1}{\sqrt{2}}(1-j)$, we have for a highly conductive medium that

$$k \simeq (1-j)\sqrt{\frac{\omega\mu\sigma}{2}} = k' - jk'' \tag{8.1.10}$$

For a plane wave, e^{-jkz}, it then becomes

$$e^{-jkz} = e^{-jk'z - k''z} \tag{8.1.11}$$

By converting the above phasor back to the time domain, this plane wave decays exponentially as well as oscillates in the z direction. 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

$$\delta = \frac{1}{k''} = \sqrt{\frac{2}{\omega\mu\sigma}} \tag{8.1.12}$$

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 Earth, or wave propagating in highly conductive metal, like your induction cooker.

8.1.2 Lowly Conductive Case

When the conductivity is low, namely, when the displacement current is larger than the conduction current, then $\frac{\sigma}{\omega\varepsilon} \ll 1$, we have

$$\begin{aligned} k &= \omega\sqrt{\mu\left(\varepsilon - j\frac{\sigma}{\omega}\right)} = \omega\sqrt{\mu\varepsilon\left(1 - \frac{j\sigma}{\omega\varepsilon}\right)} \\ &\approx \omega\sqrt{\mu\varepsilon}\left(1 - j\frac{1}{2}\frac{\sigma}{\omega\varepsilon}\right) = k' - jk'' \end{aligned} \tag{8.1.13}$$

The above is the approximation to $k = k' - jk''$ for a low conductivity medium where conduction current is much smaller than displacement current.[1] 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, $e^{-i\omega t}$ time convention is preferred. In that case, we need to do the switch $j \to -i$, and a loss medium is denoted by $\varepsilon = \varepsilon' + i\varepsilon''$.

[1] We have made use of the approximation that $(1+x)^n \approx 1 + nx$ when x is small, which can 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 $\mathbf{J} = \varrho\mathbf{v}$.

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 [66]. 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 mechanism can be easily introduced.

First, let us look at the simple constitutive relation where

$$\mathbf{D} = \varepsilon_0\mathbf{E} + \mathbf{P} \tag{8.3.1}$$

We have 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. 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 and Lorentz models, can be unified in one equation as shall be shown.

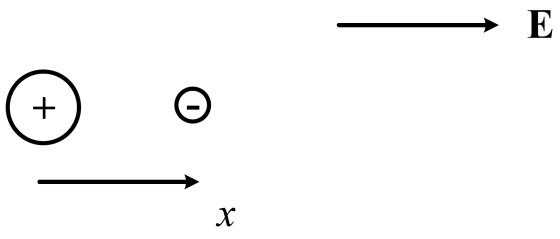


Figure 8.1: Polarization of an atom in the presence of an electric field. Here, it is assumed that the electron is weakly bound or unbound to the nucleus of the atom.

8.3.1 Cold Collisionless Plasma Medium

We can first start with a simple electron driven by an electric field $\mathbf{E}$ in the absence of a magnetic field $\mathbf{B}$.[2] 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). 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, 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 [67]. Hence, we assume that the positive ions are more massive, sluggish, and move very little compared to the 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, and we can write

$$\mathbf{P} = -\frac{Ne^2}{\omega^2 m_e}\mathbf{E} = -\frac{{\omega_p}^2}{\omega^2}\varepsilon_0 \mathbf{E} \tag{8.3.8}$$

[2]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.

where we have defined ${\omega_p}^2 = Ne^2/(m_e\varepsilon_0)$. 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 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 is negative, disallowing the propagation of a wave as we shall see.

Hence, $\varepsilon < 0$ if

$$\omega < \omega_p = \sqrt{N/(m_e\varepsilon_0)}e$$

Here, ω_p is the plasma frequency. 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 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 never 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 than once: the experiment also heralds in the age of wireless communications.

This experiment also stirred 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 [68]. 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 verses!

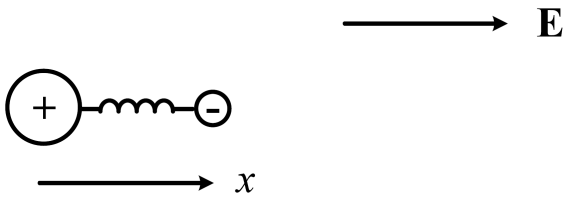


Figure 8.2: The electron is bound to the ion by an attractive force. This can be approximately modeled by a spring providing a restoring force to the electron.

8.3.2 Bound Electron Case

The above model 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 spring (see Figure 8.2), 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 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 then is

$$p = -ex = -\frac{e^2}{(\omega^2 - {\omega_0}^2) m_e} E \tag{8.3.13}$$

Therefore, when the frequency is low or $\omega = 0$, this dipole moment 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 $\mathbf{P}$.

8.3.3 Damping or Dissipation Case

Equation (8.3.11) can be generalized to the case when frictional, damping, or dissipation 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 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. It is proportional to the destruction (or change) of momentum ($m_e \frac{dx}{dt}$) of an electron. In the average sense, the destruction of the momentum is given by the product of the collision frequency 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.

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

$$P = \frac{-Ne^2}{(\omega^2 - j\omega\Gamma - {\omega_0}^2) m_e} E \tag{8.3.16}$$

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

$$\chi \approx +\frac{{\omega_p}^2}{j\omega\Gamma} = -j\frac{{\omega_p}^2}{\omega\Gamma} \tag{8.3.18}$$

χ exhibits a large negative imaginary part, the hallmark of a dissipative medium, as in the conducting medium we have previously studied.

8.3.4 Broad Applicability of Drude-Lorentz-Sommerfeld Model

The DLS model is a wonderful model because it can capture phenomenologically the essence of the physics of many electromagnetic media, even though it is a purely classical model.[3] 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 [34],[4] and the damped oscillation 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 [69] (see Figure 8.3).

[3] What we mean here is that only Newton's law has been used, and no quantum theory as yet.

[4] 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.

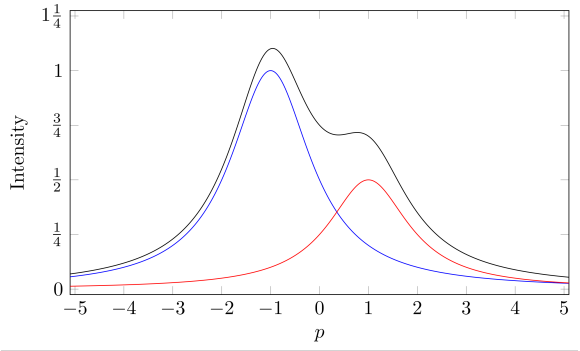


Figure 8.3: A Lorentzian has almost a bell-shape curve. By assuming multi-species of DLS oscillators in a medium, one can fit absorption spectra of almost any shape (courtesy of Wikipedia [69]).

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 [70]. 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.[5] Also, if there is no restoring force, then $\omega_0 = 0$. This is true for 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), again we have[6]

$$\chi \approx -j\frac{{\omega_p}^2}{\omega\Gamma} \tag{8.3.19}$$

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

$$\varepsilon = \varepsilon_0\left(1 - j\frac{\sigma}{\omega\varepsilon_0}\right) \tag{8.3.21}$$

[5]It is to be noted that electron has a different effective mass in a crystal lattice [71, 72], and hence, the electron mass has to be changed accordingly in the DLS model.

[6]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 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 [73].

As such, the DLS model is quite powerful: it can be used to explain a wide range of phenomena from very low frequency to optical frequency. The fact that $\varepsilon < 0$ can be used to explain many phenomena. The ionosphere is essentially a plasma medium described by

$$\varepsilon = \varepsilon_0 \left(1 - \frac{{\omega_p}^2}{\omega^2}\right) \tag{8.3.23}$$

with $\omega_0 = \Gamma = 0$ called a cold collisionless plasma. Radio wave or microwave can only penetrate through this ionosphere when $\omega > \omega_p$, so that $\varepsilon > 0$. 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.

8.3.5 Frequency Dispersive Media

The DLS model shows that, except for vacuum, all media are frequency dispersive. It is prudent to digress to 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

$$\mathbf{D}(\omega) = \varepsilon(\omega)\mathbf{E}(\omega) \tag{8.3.24}$$

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.,

$$\mathbf{D}(t) = \varepsilon(t) \circledast \mathbf{E}(t) \tag{8.3.25}$$

Since the above represents a linear time-invariant (LTI) system [52], it implies that an input is not followed by an instantaneous output. In other words, 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 mono-chromatic 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.

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. In other words, 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 8.4) 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 8.4: Effective masses of electron in different metals.

8.3.6 Plasmonic Nanoparticles

When a plasmonic nanoparticle made of gold is excited by light, its response is given by (see homework assignment)

$$\Phi_R = E_0 \frac{a^3 \cos\theta}{r^2} \frac{\varepsilon_s - \varepsilon_0}{\varepsilon_s + 2\varepsilon_0} \tag{8.3.26}$$

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.[7] 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 reflection of light. These nanoparticles were impregnated in the glass or lacquer ware. By impregnating these particles 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 [74]).

[7] 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 8.5: Ancient Roman goblets whose laquer coating glisten better due to the presence of gold nanoparticles. Gold or silver at optical frequencies behaves like plasma (courtesy of Smithsonian.com).

Lecture 9

Waves in Gyrotropic Media, Polarization

Gyrotropy is an important concept in electromagnetics. 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 the satellite by affecting the polarization of the wave. We will study this important topic in this lecture.

9.1 Gyrotropic Media and Faraday Rotation

This section derives the effective permittivity tensor of a gyrotropic medium in the ionsphere. Our ionosphere is always biased by a static magnetic field due to the Earth's magnetic field [75]. But in this derivation, to capture the salient feature 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

$$m_e \frac{d\mathbf{v}}{dt} = -e(\mathbf{E} + \mathbf{v} \times \mathbf{B}) \tag{9.1.1}$$

Next, let us assume that the electric field is polarized in the xy plane. The derivative of $\mathbf{v}$ 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 frequency domain 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 as follows: We can multiply (9.1.3) by $\pm j$ and add it to (9.1.2) to get two decoupled equations [76]:

$$m_e\omega^2(x+jy) = e[(E_x + jE_y) + \omega B_0(x+jy)] \tag{9.1.4}$$

$$m_e\omega^2(x-jy) = e[(E_x - jE_y) - \omega B_0(x-jy)] \tag{9.1.5}$$

In the above, if we take the new unknowns to be $x \pm jy$, the two equations are decoupled with respect to to these two unknowns. 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

$$C_\pm = \frac{e}{m_e\omega^2 \mp eB_0\omega} \tag{9.1.10}$$

(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 [77]. Thus, the new equation can be solved easily.)

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

$$P_\pm = \varepsilon_0\chi_\pm E_\pm \tag{9.1.11}$$

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[1] and plasma frequency, respectively.

$$\Omega = \frac{eB_0}{m_e}, \quad {\omega_p}^2 = \frac{Ne^2}{m_e\varepsilon_0} \tag{9.1.13}$$

At the cyclotron frequency, $|\chi_\pm| \to \infty$. 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

[1] This is also called the gyrofrequency.

zero. This is like what happens to an LC tank circuit at resonance whose current or voltage tends to infinity when the forcing term, like the voltage or current is nonzero.

Now, one can 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

$$\begin{aligned} P_x = \frac{P_+ + P_-}{2} = \frac{\varepsilon_0}{2}(\chi_+ E_+ + \chi_- E_-) &= \frac{\varepsilon_0}{2}[\chi_+(E_x + jE_y) + \chi_-(E_x - jE_y)] \\ &= \frac{\varepsilon_0}{2}[(\chi_+ + \chi_-)E_x + j(\chi_+ - \chi_-)E_y] \qquad (9.1.14) \\ P_y = \frac{P_+ - P_-}{2j} = \frac{\varepsilon_0}{2j}(\chi_+ E_+ - \chi_- E_-) &= \frac{\varepsilon_0}{2j}[\chi_+(E_x + jE_y) - \chi_-(E_x - jE_y)] \\ &= \frac{\varepsilon_0}{2j}[(\chi_+ - \chi_-)E_x + j(\chi_+ + \chi_-)E_y] \qquad (9.1.15) \end{aligned}$$

The above relationship in cartesian coordinates can be expressed using a tensor where

$$\mathbf{P} = \varepsilon_0 \overline{\boldsymbol{\chi}} \cdot \mathbf{E} \qquad (9.1.16)$$

where $\mathbf{P} = [P_x, P_y]$, and $\mathbf{E} = [E_x, E_y]$. From the above, $\overline{\boldsymbol{\chi}}$ is of the form

$$\overline{\boldsymbol{\chi}} = \frac{1}{2}\begin{pmatrix} (\chi_+ + \chi_-) & j(\chi_+ - \chi_-) \\ -j(\chi_+ - \chi_-) & (\chi_+ + \chi_-) \end{pmatrix} = \begin{pmatrix} -\frac{{\omega_p}^2}{\omega^2 - \Omega^2} & -j\frac{{\omega_p}^2 \Omega}{\omega(\omega^2 - \Omega^2)} \\ j\frac{{\omega_p}^2 \Omega}{\omega(\omega^2 - \Omega^2)} & -\frac{{\omega_p}^2}{\omega^2 - \Omega^2} \end{pmatrix} \qquad (9.1.17)$$

Notice that in the above, when the $\mathbf{B}$ field is turned off or $\Omega = 0$, the above resembles the solution of a collisionless, cold plasma again. For the $\mathbf{B} = 0$ case with electric field in the x (or y) direction, it will drive a motion of the electron to be in the x (or y) direction. In this case, $\mathbf{v} \times \mathbf{B}$ term is zero, and the electron motion is unaffected by the magnetic field as can be seen from the Lorentz force law or (9.1.1). Hence, it behaves like a simple collisionless plasma without a biasing magnetic field.

Consequently, for the $\mathbf{B} \neq 0$ case, the above can be generalized to 3D to give

$$\overline{\boldsymbol{\chi}} = \begin{bmatrix} \chi_0 & j\chi_1 & 0 \\ -j\chi_1 & \chi_0 & 0 \\ 0 & 0 & \chi_p \end{bmatrix} \qquad (9.1.18)$$

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 is like that of 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

$$\overline{\boldsymbol{\varepsilon}} = \varepsilon_0 \begin{bmatrix} 1 + \chi_0 & j\chi_1 & 0 \\ -j\chi_1 & 1 + \chi_0 & 0 \\ 0 & 0 & 1 + \chi_p \end{bmatrix} \qquad (9.1.19)$$

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 hallmark 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 [76], which we shall learn more later. This phenomenon poses a severe problem for Earth-to-satellite communication, using linearly polarized wave as it requires the alignment of the Earth-to-satellite antennas. 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 out 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 [78].

9.2 Wave Polarization

Studying wave polarization is very important for communication purposes [32]. A wave whose electric field is pointing in the x direction while propagating in the z direction is 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 suffers from 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 [79]. 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.

We can write a general uniform plane wave propagating in the z direction in the time domain 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 we have in general

$$E_x(z,t) = E_1 \cos(\omega t - \beta z) \tag{9.2.2}$$

$$E_y(z,t) = E_2 \cos(\omega t - \beta z + \alpha) \tag{9.2.3}$$

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. First, we set $z = 0$ to observe this field. 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 **E** traces out an ellipse as show in Figure 9.1. With the left-hand thumb pointing in the z direction, and the wave rotating in the direction of the fingers, such a wave is called left-hand elliptically polarized (LHEP) wave.

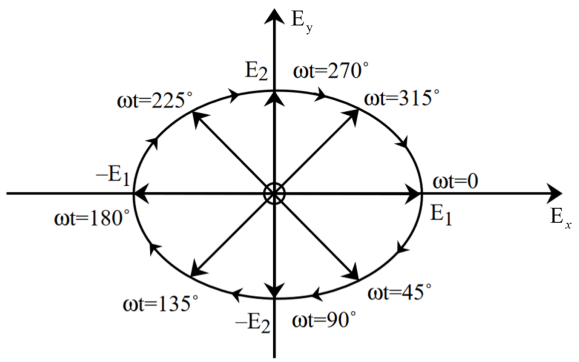


Figure 9.1: If one follows the tip of the electric field vector, it traces out an ellipse as a function of time t.

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 wave for different phase differences and amplitude ratio. Figure 9.3 shows a graphic picture of a CP wave propagating through space.

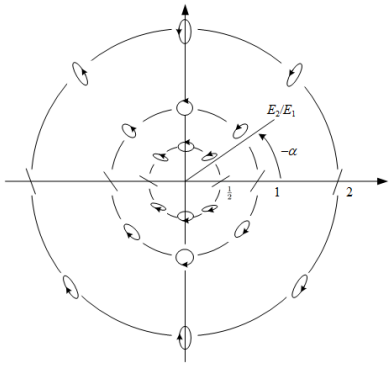


Figure 9.2: Due to different phase difference between the E_x and E_y components of the field, and their relative amplitudes E_2/E_1, different polarizations will ensure. The arrow indicates the direction of rotation of the field vector.

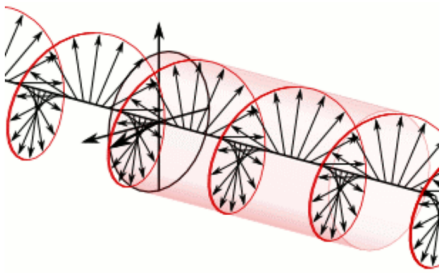


Figure 9.3: The rotation of the field vector of a right-hand circular polarization wave as it propagates in the right direction [80] (courtesy of Wikipedia).

9.2.1 Arbitrary Polarization Case and Axial Ratio[2]

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

$$E_x = E_1 \cos \omega t, \qquad E_y = E_2 \cos(\omega t + \alpha) = E_2(\cos \omega t \cos \alpha - \sin \omega t \sin \alpha) \tag{9.2.11}$$

[2]This section is mathematically complicated. It can be skipped on first reading.

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,

$$ax^2 - bxy + cy^2 = 1 \tag{9.2.15}$$

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 by letting

$$x' = x\cos\theta - y\sin\theta \tag{9.2.17}$$

$$y' = x\sin\theta + y\cos\theta \tag{9.2.18}$$

to get

$$x^2\left(\frac{\cos^2\theta}{A^2} + \frac{\sin^2\theta}{B^2}\right) - xy\sin 2\theta\left(\frac{1}{A^2} - \frac{1}{B^2}\right) + y^2\left(\frac{\sin^2\theta}{A^2} + \frac{\cos^2\theta}{B^2}\right) = 1 \tag{9.2.19}$$

Comparing (9.2.13) and (9.2.19), one gets

$$\theta = \frac{1}{2}\tan^{-1}\left(\frac{2\cos\alpha E_1E_2}{{E_2}^2 - {E_1}^2}\right) \tag{9.2.20}$$

$$\mathrm{AR} = \left(\frac{1+\Delta}{1-\Delta}\right)^{1/2} > 1 \tag{9.2.21}$$

where AR is the axial ratio and

$$\Delta = \left(1 - \frac{4{E_1}^2{E_2}^2\sin^2\alpha}{\left({E_1}^2 + {E_2}^2\right)^2}\right)^{1/2} \tag{9.2.22}$$

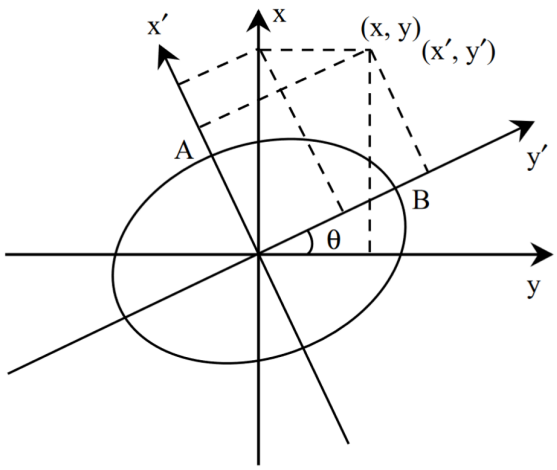


Figure 9.4: This figure shows the parameters used to derive the axial ratio (AR) of an elliptically polarized wave.

9.3 Polarization and Power Flow

For a linearly polarized wave in the time domain,

$$\mathbf{E} = \hat{x}E_0 \cos(\omega t - \beta z), \quad \mathbf{H} = \hat{y}\frac{E_0}{\eta}\cos(\omega t - \beta z) \tag{9.3.1}$$

Hence, the instantaneous power we have learnt previously in Section 5.3 becomes

$$\mathbf{S}(t) = \mathbf{E}(t) \times \mathbf{H}(t) = \hat{z}\frac{{E_0}^2}{\eta}\cos^2(\omega t - \beta z) \tag{9.3.2}$$

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.

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) becomes

$$\mathbf{E}(z,\omega) = \hat{x}E_1 e^{-j\beta z} + \hat{y}E_2 e^{-j\beta z + j\alpha} \tag{9.3.3}$$

For LHEP wave,

$$\mathbf{E}(z,\omega) = e^{-j\beta z}(\hat{x}E_1 + j\hat{y}E_2) \tag{9.3.4}$$

whereas for LHCP wave,

$$\mathbf{E}(z,\omega) = e^{-j\beta z}E_1(\hat{x} + j\hat{y}) \tag{9.3.5}$$

For RHEP wave, the above becomes

$$\mathbf{E}(z,\omega) = e^{-j\beta z}(\hat{x}E_1 - j\hat{y}E_2) \tag{9.3.6}$$

whereas for RHCP wave, it is

$$\mathbf{E}(z,\omega) = e^{-j\beta z}E_1(\hat{x} - j\hat{y}) \tag{9.3.7}$$

Focussing on the circularly polarized wave,

$$\mathbf{E} = (\hat{x} \pm j\hat{y})E_0 e^{-j\beta z} \tag{9.3.8}$$

Using that

$$\mathbf{H} = \frac{\boldsymbol{\beta} \times \mathbf{E}}{\omega\mu},$$

where $\boldsymbol{\beta} = \hat{z}\beta$, then

$$\mathbf{H} = (\mp\hat{x} - j\hat{y})j\frac{E_0}{\eta}e^{-j\beta z} \tag{9.3.9}$$

where $\eta = \sqrt{\mu/\varepsilon}$ is the intrinsic impedance of the medium. Therefore,

$$\mathbf{E}(t) = \hat{x}E_0\cos(\omega t - \beta z) \pm \hat{y}E_0\sin(\omega t - \beta z) \tag{9.3.10}$$

$$\mathbf{H}(t) = \mp\hat{x}\frac{E_0}{\eta}\sin(\omega t - \beta z) + \hat{y}\frac{E_0}{\eta}\cos(\omega t - \beta z) \tag{9.3.11}$$

Then the instantaneous power becomes

$$\mathbf{S}(t) = \mathbf{E}(t) \times \mathbf{H}(t) = \hat{z}\frac{{E_0}^2}{\eta}\cos^2(\omega t - \beta z) + \hat{z}\frac{{E_0}^2}{\eta}\sin^2(\omega t - \beta z) = \hat{z}\frac{{E_0}^2}{\eta} \tag{9.3.12}$$

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).

It is to be noted that the complex Poynting's vector for a lossless medium

$$\underset{\sim}{\mathbf{S}} = \mathbf{E} \times \mathbf{H}^* \tag{9.3.13}$$

is real and constant independent of space both for linearly, circularly, and elliptically polarized waves. 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!

Lecture 10

Spin Angular Momentum, Complex Poynting's Theorem, Lossless Condition, Energy Density

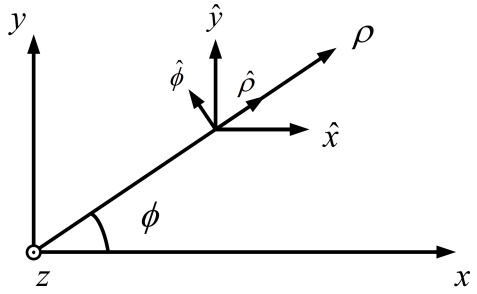


Figure 10.1: The local coordinates used to describe a circularly polarized wave: In cartesian and polar coordinates.

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 has a momentum: for a linearly polarized wave, it carries linear momention in the direction of the propagation of the traveling wave. But for a circularly polarized wave, it carries angular momentum as well.

We have studied the complex Poynting's theorem in the frequency domain with phasors.

Here, we will derive the lossless conditions for the permittivity and permeability tensors. Energy density is well defined for a lossless dispersionless medium, but it assumes a different formula when the medium is dispersive.

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, the $\hat{\rho}$ and $\hat{\phi}$ of a CP 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 writing

$$e^{-j\phi} = e^{-jk_\phi \rho\phi} \tag{10.1.2}$$

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 [81].

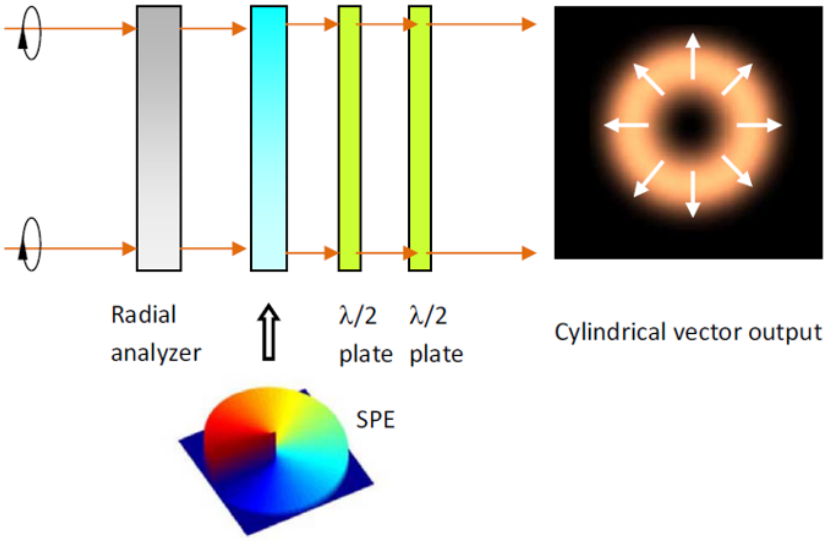


Figure 10.2: A cylindrical vector beam can be generated experimentally. The spiral phase element (SPE) compensates for the $\exp(\pm j\phi)$ phase shift (courtesy of Zhan, Q. [81]). The half-wave plate rotates the polarization of a wave by 90 degrees.

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 [32, p. 59] [48]. 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, associated with a packet of energy given by $\hbar\omega$. It is also well known that a photon has momentum given by $\hbar k$. 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 flow associated with these streaming photons is given by

$$\mathbf{E} \times \mathbf{H} = \hbar\omega N c\hat{z} \tag{10.2.2}$$

Assuming that the plane wave is propagating in the z direction. Using $k = \omega/c$, we can rewrite the above more suggestively as

$$\mathbf{E} \times \mathbf{H} = \hbar\omega N c\hat{z} = \hbar k N c^2 \hat{z} \tag{10.2.3}$$

where $k = \omega/c$. Defining the momentum density vector to be

$$\mathbf{G} = \hbar k N \hat{z} \tag{10.2.4}$$

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

$$\mathbf{G} = \mathbf{D} \times \mathbf{B} \tag{10.2.6}$$

where $\mathbf{D} = \varepsilon\mathbf{E}$, and $\mathbf{B} = \mu\mathbf{H}$.[1]

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 [32, 48]. 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 the phasors of time harmonic fields $\mathbf{E}(\mathbf{r}, t)$ and $\mathbf{H}(\mathbf{r}, t)$, namely, $\mathbf{E}(\mathbf{r}, \omega)$ and $\mathbf{H}(\mathbf{r}, \omega)$ respectively, we can show that

$$\langle \mathbf{E}(\mathbf{r}, t) \times \mathbf{H}(\mathbf{r}, t) \rangle = \frac{1}{2} \Re e\{\mathbf{E}(\mathbf{r}, \omega) \times \mathbf{H}^*(\mathbf{r}, \omega)\}. \tag{10.3.2}$$

Here, the vector $\mathbf{E}(\mathbf{r}, \omega) \times \mathbf{H}^*(\mathbf{r}, \omega)$, as previously discussed, is also known as the complex Poynting vector. If 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

$$\underset{\sim}{\mathbf{S}}(\mathbf{r}, \omega) = \mathbf{E}(\mathbf{r}, \omega) \times \mathbf{H}^*(\mathbf{r}, \omega) \tag{10.3.4}$$

[1] The author is indebted to Wei SHA for this simple derivation.

then for time-harmonic fields,

$$\langle \mathbf{S}(\mathbf{r},t)\rangle = \frac{1}{2}\Re e\underset{\sim}{\mathbf{S}}(\mathbf{r},\omega) \tag{10.3.5}$$

The above is often the source of confusion in the definition of Poynting's vector.

In the above definition of complex Poynting's vector and its aforementioned property, and its 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[2]

$$\nabla\cdot(\mathbf{E}\times\mathbf{H}^*) = \mathbf{H}^*\cdot\nabla\times\mathbf{E} - \mathbf{E}\cdot\nabla\times\mathbf{H}^*. \tag{10.3.6}$$

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.

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 energy conservation, the real part of the right-hand side of (10.3.11), without the $\mathbf{E}\cdot\mathbf{J}^*$ term, must be zero. In other words, the right-hand side of (10.3.11) should be purely imaginary. Thus

$$\int_V dV(\mathbf{H}^*\cdot\overline{\boldsymbol{\mu}}\cdot\mathbf{H} - \mathbf{E}\cdot\overline{\boldsymbol{\varepsilon}}^*\cdot\mathbf{E}^*) \tag{10.3.14}$$

[2]We will drop the argument $\mathbf{r},\omega$ for the phasors in our discussion next as they will be implied.

must be a 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.[3] 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 arbitraty $\mathbf{E}$ and $\mathbf{H}$, and (10.3.13) is true, implying a lossless region V. Notice that for an isotropic medium, 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 involved.

For a lossy medium which is conductive, 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

$$\int_S d\mathbf{S} \cdot (\mathbf{E} \times \mathbf{H}^*) = -j\omega \int_V dV \left[\mathbf{H}^* \cdot \overline{\boldsymbol{\mu}} \cdot \mathbf{H} - \mathbf{E} \cdot \left(\overline{\boldsymbol{\varepsilon}}^* + \frac{j\overline{\boldsymbol{\sigma}}^*}{\omega} \right) \cdot \mathbf{E}^* \right] \tag{10.3.16}$$

$$= -j\omega \int dV [\mathbf{H}^* \cdot \overline{\boldsymbol{\mu}} \cdot \mathbf{H} - \mathbf{E} \cdot \tilde{\overline{\boldsymbol{\varepsilon}}}^* \cdot \mathbf{E}^*], \tag{10.3.17}$$

where $\tilde{\overline{\boldsymbol{\varepsilon}}} = \overline{\boldsymbol{\varepsilon}} - \frac{j\overline{\boldsymbol{\sigma}}}{\omega}$ 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 $\tilde{\overline{\boldsymbol{\varepsilon}}}$ 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 will be derived 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, then the right-hand side of the

[3] $\mathbf{H}^* \cdot \overline{\boldsymbol{\mu}} \cdot \mathbf{H}$ 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.

above can be interpreted as stored energy density. Hence, 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.

10.4 Energy Density in Dispersive Media

A dispersive medium alters our concept of what energy density is.[4] To this end, we assume that the field has complex ω dependence in $e^{j\omega t}$, where $\omega = \omega' - j\omega''$, rather than real ω dependence. We take the divergence of the complex power for fields with such 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 since they are not truly time harmonic. In other words, we let

$$\mathbf{E}(\mathbf{r},t) = \underset{\sim}{\mathbf{E}}(\mathbf{r},\omega)e^{j\omega t}, \quad \mathbf{H}(\mathbf{r},t) = \underset{\sim}{\mathbf{H}}(\mathbf{r},\omega)e^{j\omega t} \tag{10.4.1}$$

The above, just like phasors, can be made to satisfy Maxwell's equations where the time derivative becomes $j\omega$ but with complex ω. We can study the quantity $\mathbf{E}(\mathbf{r},t) \times \mathbf{H}^*(\mathbf{r},t)$ which 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 not in the complex ω case. Hence,

$$\begin{aligned} \nabla \cdot [\mathbf{E}(t) \times \mathbf{H}^*(t)] &= \mathbf{H}^*(t) \cdot \nabla \times \mathbf{E}(t) - \mathbf{E}(t) \cdot \nabla \times \mathbf{H}^*(t) \\ &= -\mathbf{H}^*(t) \cdot j\omega\mu\mathbf{H}(t) + \mathbf{E}(t) \cdot j\omega^*\varepsilon^*\mathbf{E}^*(t) \end{aligned} \tag{10.4.2}$$

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, now $\mathbf{H}^*(t) \sim e^{-j\omega^* t}$, and the term like $\mathbf{E}(t) \times \mathbf{H}^*(t)$ is not truly time independent but becomes

$$\mathbf{E}(t) \times \mathbf{H}^*(t) \sim e^{j(\omega-\omega^*)t} = e^{2\omega'' t} \tag{10.4.3}$$

And each of the term above will have similar time dependence. Writing (10.4.2) more explicitly, by letting $\omega = \omega' - j\omega''$, we have

$$\nabla \cdot [\mathbf{E}(t) \times \mathbf{H}^*(t)] = -j(\omega' - j\omega'')\mu(\omega)|\mathbf{H}(t)|^2 + j(\omega' + j\omega'')\varepsilon^*(\omega)|\mathbf{E}(t)|^2 \tag{10.4.4}$$

Assuming that $\omega'' \ll \omega'$, or that the field is quasi-time-harmonic, we can let, after using Taylor series approximation, that

$$\mu(\omega' - j\omega'') \cong \mu(\omega') - j\omega'' \frac{\partial \mu(\omega')}{\partial \omega'}, \quad \varepsilon(\omega' - j\omega'') \cong \varepsilon(\omega') - j\omega'' \frac{\partial \varepsilon(\omega')}{\partial \omega'} \tag{10.4.5}$$

[4] The derivation here is inspired by H.A. Haus, Electromagnetic Noise and Quantum Optical Measurements [82]. Generalization to anisotropic media is given by W.C. Chew, Lectures on Theory of Microwave and Optical Waveguides [83].

Using (10.4.5) in (10.4.4), and collecting terms of the same order, and ignoring $(\omega'')^2$ terms, gives[5]

$$\begin{aligned}\nabla\cdot[\mathbf{E}(t)\times\mathbf{H}^*(t)] \cong & -j\omega'\mu(\omega')|\mathbf{H}(t)|^2+j\omega'\varepsilon^*(\omega')|\mathbf{E}(t)|^2\\ & -\omega''\mu(\omega')|\mathbf{H}(t)|^2-\omega'\omega''\frac{\partial\mu}{\partial\omega'}|\mathbf{H}(t)|^2\\ & -\omega''\varepsilon^*(\omega')|\mathbf{E}(t)|^2-\omega'\omega''\frac{\partial\varepsilon^*}{\partial\omega'}|\mathbf{E}(t)|^2\end{aligned}\tag{10.4.6}$$

The above can be rewritten as

$$\begin{aligned}\nabla\cdot[\mathbf{E}(t)\times\mathbf{H}^*(t)] \cong & -j\omega'\left[\mu(\omega')|\mathbf{H}(t)|^2-\varepsilon^*(\omega')|\mathbf{E}(t)|^2\right]\\ & -\omega''\left[\frac{\partial\omega'\mu(\omega')}{\partial\omega'}|\mathbf{H}(t)|^2+\frac{\partial\omega'\varepsilon^*(\omega')}{\partial\omega'}|\mathbf{E}(t)|^2\right]\end{aligned}\tag{10.4.7}$$

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. In the limit when $\omega''\to 0$, when half the imaginary part of the above equation is taken, we have

$$\nabla\cdot\frac{1}{2}\Im m\left[\mathbf{E}\times\mathbf{H}^*\right]=-\omega'\left[\frac{1}{2}\mu|\mathbf{H}|^2-\frac{1}{2}\varepsilon|\mathbf{E}|^2\right]\tag{10.4.8}$$

which has the physical interpretation of reactive power as has been previously discussed. When half the real part of (10.4.7) is taken, we obtain

$$\nabla\cdot\frac{1}{2}\Re e[\mathbf{E}\times\mathbf{H}^*]=-\frac{\omega''}{2}\left[\frac{\partial\omega'\mu}{\partial\omega'}|\mathbf{H}|^2+\frac{\partial\omega'\varepsilon}{\partial\omega'}|\mathbf{E}|^2\right]\tag{10.4.9}$$

Since the right-hand side has time dependence of $e^{2\omega''t}$, it can be written as

$$\nabla\cdot\frac{1}{2}\Re e[\mathbf{E}\times\mathbf{H}^*]=-\frac{\partial}{\partial t}\frac{1}{4}\left[\frac{\partial\omega'\mu}{\partial\omega'}|\mathbf{H}|^2+\frac{\partial\omega'\varepsilon}{\partial\omega'}|\mathbf{E}|^2\right]=-\frac{\partial}{\partial t}\langle W_T\rangle\tag{10.4.10}$$

Therefore, the time-average stored energy density can be identified as

$$\langle W_T\rangle=\frac{1}{4}\left[\frac{\partial\omega'\mu}{\partial\omega'}|\mathbf{H}|^2+\frac{\partial\omega'\varepsilon}{\partial\omega'}|\mathbf{E}|^2\right]\tag{10.4.11}$$

For a non-dispersive medium, the above reverts back to a familiar expression,

$$\langle W_T\rangle=\frac{1}{4}\left[\mu|\mathbf{H}|^2+\varepsilon|\mathbf{E}|^2\right]\tag{10.4.12}$$

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

[5]This is the general technique of perturbation expansion [45].

our usual interpretation of complex power. However, by assuming the frequency ω to have a 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, 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 [84], as energy density times group velocity should be power flow. More discussion on this topic can be found in Jackson [48].

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 accordingly modified if the capacitor or inductor is frequency dependent.

Lecture 11

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 start to emerge.

Even though circuit theory has played an indispensable role in the development of the computer chip industry, it 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 have the capability of including transmission line as an element in modeling.

11.1 Transmission Line Theory

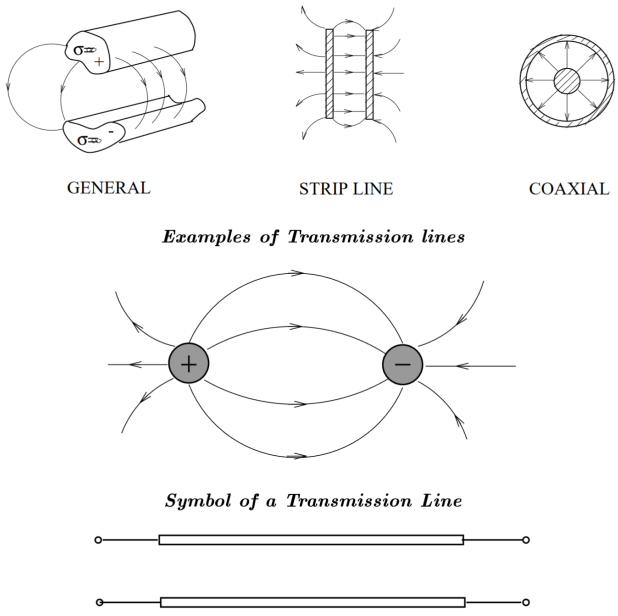


Figure 11.1: Various kinds of transmission lines. Schematically, all of them can be modeled by two parallel wires.

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 [31, 32, 44, 50, 54, 65, 79, 83, 85, 86].

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,[1] but approximate solutions can be obtained using circuit theory in the long-wavelength limit. We shall show that circuit theory is an approximation of electromagnetic field theory when the wavelength is very long: the longer the wavelength, the better is the approximation [50].

Examples of transmission lines are shown in Figure 11.1. The symbol for a transmission line is usually represented by two pieces of parallel wires, but in practice, these wires need not be parallel as showin in Figure 11.2.

[1] Usually called full-wave analysis.

Figure 11.2: 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 [87].)

Circuit theory also explains why waveguides 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 they still work well (see Figure 11.2). Hence, it is simplest to first explain the propagation of electromagnetic signal on a transmission line using circuit analysis.

11.1.1 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 capacitive coupling, with electric fields that start with positive charges and end with negative charges, and hence stored energy in the electric field. Moreover, a piece of wire carrying a current generates a magnetic field, and hence, yields stored energy in the magnetic field. 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. Hence, we can model two pieces of metal with a distributed lumped element model as shown in Figure 11.3. 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 lump-element model.

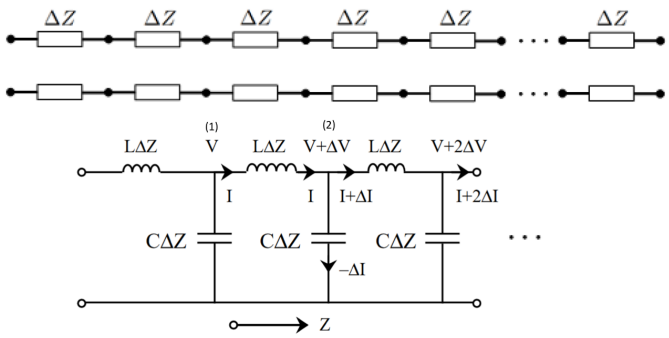


Figure 11.3: A long transmission line can be replaced by a concatenation of many short transmission lines. For each pair of short wires, there are capacitive coupling between them. Furtheremore, when current flows in the wire, magnetic field is generated making them behave like an inductor. Therefore, it can be replaced by a lumped-element approximation as shown.

Telegrapher's Equations

First, we recall that the V-I relation of an inductor is

$$V_0 = L_0 \frac{dI_0}{dt} \tag{11.1.1}$$

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, we have

$$V - (V + \Delta V) = L\Delta z \frac{\partial I}{\partial t} \tag{11.1.2}$$

The left-hand side is the voltage drop across the 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 can be simplified to

$$\Delta V = -L\Delta z \frac{\partial I}{\partial t} \tag{11.1.3}$$

Next, we make use of the V-I relation for a capacitor, which is

$$I_0 = C_0 \frac{dV_0}{dt} \tag{11.1.4}$$

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 gives

$$-\Delta I = C\Delta z \frac{\partial}{\partial t}(V + \Delta V) \approx C\Delta z \frac{\partial V}{\partial t} \tag{11.1.5}$$

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 capacitor is $-\Delta I$, where $\Delta I = I(z+\Delta z, t) - I(z,t)$. In the last approximation in (11.1.5), 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 (11.1.3) and (11.1.5) that

$$\frac{\partial V(z,t)}{\partial z} = -L\frac{\partial I(z,t)}{\partial t} \tag{11.1.6}$$

$$\frac{\partial I(z,t)}{\partial z} = -C\frac{\partial V(z,t)}{\partial t} \tag{11.1.7}$$

The above are the **telegrapher's equations**.[2] They are two coupled first-order equations, and can be converted into second-order equations easily. Therefore,

$$\frac{\partial^2 V}{\partial z^2} - LC\frac{\partial^2 V}{\partial t^2} = 0 \tag{11.1.8}$$

$$\frac{\partial^2 I}{\partial z^2} - LC\frac{\partial^2 I}{\partial t^2} = 0 \tag{11.1.9}$$

The above are wave equations that we have previously studied, where the velocity of the wave is given by

$$v = \frac{1}{\sqrt{LC}} \tag{11.1.10}$$

Furthermore, if we assume that

$$V(z,t) = V_0 f_+(z - vt), \qquad I(z,t) = I_0 f_+(z - vt) \tag{11.1.11}$$

corresponding to a right-traveling wave, they can be verified to satisfy (11.1.6) and (11.1.7) by back substitution.

Consequently, we can easily show that

$$\frac{V(z,t)}{I(z,t)} = \sqrt{\frac{L}{C}} = Z_0 \tag{11.1.12}$$

where Z_0 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, we can let,

$$V(z,t) = V_0 f_-(z + vt), \qquad I(z,t) = I_0 f_-(z + vt) \tag{11.1.13}$$

one can easily show that

$$\frac{V(z,t)}{I(z,t)} = -\sqrt{\frac{L}{C}} = -Z_0 \tag{11.1.14}$$

[2] They can be thought of as the distillation of the Faraday's law and Ampere's law from Maxwell's equations without the source term. Their simplicity gives them an important role in engineering electromagnetics.

Time-domain analysis is very useful for transient analysis of transmission lines, especially when nonlinear elements are coupled to the transmission line.[3] 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.[4]

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,

$$\lim_{c\to\infty} \nabla^2\Phi(\mathbf{r}) + \frac{\omega^2}{c^2}\Phi(\mathbf{r}) = 0 \quad \Longrightarrow \quad \nabla^2\Phi(\mathbf{r}) = 0 \tag{11.1.15}$$

Hence, events in Laplace's equation happen instantaneously. In other words, circuit theory assumes 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.

11.1.2 Frequency-Domain Analysis–the Power of Phasor Technique Again!

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 straight forward. 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.

The telegrapher's equations (11.1.6) and (11.1.7) then in frequency domain become

$$\frac{d}{dz}V(z,\omega) = -j\omega LI(z,\omega) \tag{11.1.16}$$

$$\frac{d}{dz}I(z,\omega) = -j\omega CV(z,\omega) \tag{11.1.17}$$

The corresponding Helmholtz equations are then

$$\frac{d^2V}{dz^2} + \omega^2 LCV = 0 \tag{11.1.18}$$

$$\frac{d^2I}{dz^2} + \omega^2 LCI = 0 \tag{11.1.19}$$

The above are second ordinary differential equations, and the general solutions to the above are

$$V(z) = V_+e^{-j\beta z} + V_-e^{j\beta z} \tag{11.1.20}$$

$$I(z) = I_+e^{-j\beta z} + I_-e^{j\beta z} \tag{11.1.21}$$

[3]Remember that we can only use frequency domain technique or Fourier transform for linear time-invariant systems.

[4]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.

where $\beta = \omega\sqrt{LC}$. This is similar to what we have seen previously for plane waves in the one-dimensional wave equation in free space, where

$$E_x(z) = E_{0+}e^{-jk_0z} + E_{0-}e^{jk_0z} \tag{11.1.22}$$

where $k_0 = \omega\sqrt{\mu_0\epsilon_0}$. We see much similarity between (11.1.20), (11.1.21), and (11.1.22).

To see the solution in the time domain, we let the phasor $V_\pm = |V_\pm|e^{j\phi_\pm}$ in (11.1.20), and the voltage signal above can then be converted back to the time domain using the key formula in phasor technique as

$$V(z,t) = \Re e\{V(z,\omega)e^{j\omega t}\} \tag{11.1.23}$$

$$= |V_+|\cos(\omega t - \beta z + \phi_+) + |V_-|\cos(\omega t + \beta z + \phi_-) \tag{11.1.24}$$

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 (11.1.16) or (11.1.17)

$$\frac{V(z)}{I(z)} = \frac{V_+}{I_+} = \sqrt{\frac{L}{C}} = Z_0 \tag{11.1.25}$$

where Z_0 is the characteristic impedance. Since Z_0 is real, it implies that $V(z)$ and $I(z)$ are in phase.

Similarly, applying the same process for a left-traveling wave only, by letting $V_+ = I_+ = 0$, then

$$\frac{V(z)}{I(z)} = \frac{V_-}{I_-} = -\sqrt{\frac{L}{C}} = -Z_0 \tag{11.1.26}$$

In other words, for the left-traveling waves, the voltage and current are 180 ° out of phase.

11.2 Lossy Transmission Line

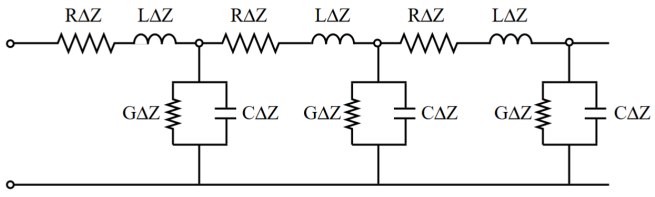


Figure 11.4: In a lossy transmission line, series resistance can be added to the series inductance, and the shunt conductance can be added to the shun susceptance of the capacitor. However, this problem is homomorphic to the lossless case in the frequency domain.

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 11.4. 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 (11.1.16) and (11.1.17) in terms of series line impedance and shunt line admittance to arrive at

$$\frac{d}{dz}V = -ZI \tag{11.2.1}$$

$$\frac{d}{dz}I = -YV \tag{11.2.2}$$

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 11.4 is topologically similar to, or homomorphic[5] to the lossless case in Figure 11.3. 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:

$$Z = j\omega L \rightarrow Z = j\omega L + R \tag{11.2.3}$$

$$Y = j\omega C \rightarrow Y = j\omega C + G \tag{11.2.4}$$

where R is the series line resistance, and G is the shunt line conductance, and now Z and Y are the series impedance and shunt admittance, (which are complex number rather than being pure imaginary), respectively. Then, the corresponding Helmholtz equations are

$$\frac{d^2V}{dz^2} - ZYV = 0 \tag{11.2.5}$$

$$\frac{d^2I}{dz^2} - ZYI = 0 \tag{11.2.6}$$

or

$$\frac{d^2V}{dz^2} - \gamma^2 V = 0 \tag{11.2.7}$$

$$\frac{d^2I}{dz^2} - \gamma^2 I = 0 \tag{11.2.8}$$

where $\gamma^2 = ZY$, or that one can also think of $\gamma^2 = -\beta^2$ by comparing with (11.1.18) and (11.1.19). Then the above is homomorphic to the lossless case except that now, β is a complex

[5]A math term for "similar in math structure". The term is even used in computer science describing a emerging field of homomorphic computing.

number, indicating that the field is decaying as it propagates. As before, the above are second order one-dimensional Helmholtz equations where the general solutions are

$$V(z) = V_+ e^{-\gamma z} + V_- e^{\gamma z} \tag{11.2.9}$$

$$I(z) = I_+ e^{-\gamma z} + I_- e^{\gamma z} \tag{11.2.10}$$

and

$$\gamma = \sqrt{ZY} = \sqrt{(j\omega L + R)(j\omega C + G)} = j\beta \tag{11.2.11}$$

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,

$$V(z) = V_+ e^{-\beta'' z - j\beta' z} + V_- e^{\beta'' z + j\beta' z} \tag{11.2.12}$$

Again, letting $V_\pm = |V_\pm| e^{j\phi_\pm}$, the above can be converted back to the time domain as

$$V(z,t) = \Re e\{V(z,\omega) e^{j\omega t}\} \tag{11.2.13}$$

$$= |V_+| e^{-\beta'' z} \cos(\omega t - \beta' z + \phi_+) + |V_-| e^{\beta'' z} \cos(\omega t + \beta' z + \phi_-) \tag{11.2.14}$$

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 (11.2.11), we retrieve the lossless case where $\beta'' = 0$ and $\gamma = j\beta = j\omega\sqrt{LC}$.

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:

$$Z_0 = \frac{V_+}{I_+} = -\frac{V_-}{I_-} = \sqrt{\frac{L}{C}} = \sqrt{\frac{j\omega L}{j\omega C}} \rightarrow Z_0 = \sqrt{\frac{Z}{Y}} = \sqrt{\frac{j\omega L + R}{j\omega C + G}} \tag{11.2.15}$$

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

$$Z_0 = \sqrt{\frac{L}{C}} \tag{11.2.16}$$

the characteristic impedance for the lossless case previously derived.

Lecture 12

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.

12.1 Terminated Transmission Lines

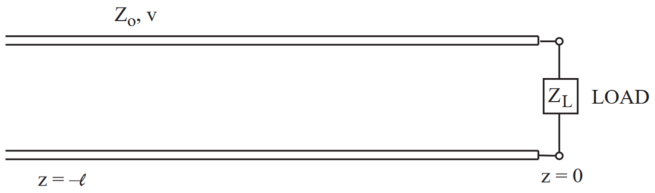


Figure 12.1: A schematic for a transmission line terminated with an impedance load Z_L at $z = 0$.

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

[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:)

by a load as shown in Figure 12.1, a right-traveling wave will be reflected by the load, 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. Thus, we assume that

$$V(z) = a_+ e^{-j\beta z} + a_- e^{j\beta z} = V_+(z) + V_-(z) \tag{12.1.1}$$

where $\beta = \omega\sqrt{LC}$. In the above, in general, $a_+ \neq a_-$. Besides, this is a linear system; hence, we can define the right-going wave $V_+(z)$ to be the input, and that the left-going wave $V_-(z)$ to be the output as due to the reflection of the right-going wave $V_+(z)$. 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

$$V_-(z = 0) = \Gamma_L V_+(z = 0) \tag{12.1.2}$$

thus, using the definition of $V_+(z)$ and $V_-(z)$ as implied in (12.1.1), we have

$$a_- = \Gamma_L a_+ \tag{12.1.3}$$

where Γ_L is the termed the reflection coefficient. Hence, (12.1.1) becomes

$$V(z) = a_+ e^{-j\beta z} + \Gamma_L a_+ e^{j\beta z} = a_+ \left(e^{-j\beta z} + \Gamma_L e^{j\beta z}\right) \tag{12.1.4}$$

The corresponding current $I(z)$ on the transmission line is given by using the telegrapher's equations as previously defined. By recalling that

$$\frac{\mathrm{d}V}{\mathrm{d}z} = -j\omega LI$$

then for the general case,

$$I(z) = \frac{a_+}{Z_0}\left(e^{-j\beta z} - \Gamma_L e^{j\beta z}\right) \tag{12.1.5}$$

Notice the sign change in the second term of the above expression.

Similar to Γ_L, a general reflection coefficient can be defined (which is a function of z) relating the left-traveling and right-traveling wave at location z such that

$$\Gamma(z) = \frac{V_-(z) = a_- e^{j\beta z}}{V_+(z) = a_+ e^{-j\beta z}} = \frac{a_- e^{j\beta z}}{a_+ e^{-j\beta z}} = \Gamma_L e^{2j\beta z} \tag{12.1.6}$$

Of course, $\Gamma(z = 0) = \Gamma_L$. Furthermore, due to the V-I relation at an impedance load, we must have[2]

$$\frac{V(z = 0)}{I(z = 0)} = Z_L \tag{12.1.7}$$

[2]One can also look at this from a differential equation viewpoint that this is a boundary condition.

or that using (12.1.4) and (12.1.5) with $z = 0$, the left-hand side of the above can be rewritten, and we have

$$\frac{1+\Gamma_L}{1-\Gamma_L} Z_0 = Z_L, \quad \text{or} \quad \frac{1+\Gamma_L}{1-\Gamma_L} = \frac{Z_L}{Z_0} = Z_{nL} \tag{12.1.8}$$

From the above, we can solve for Γ_L in terms of Z_{nL} to get

$$\Gamma_L = \frac{Z_{nL}-1}{Z_{nL}+1} = \frac{Z_L - Z_0}{Z_L + Z_0} \tag{12.1.9}$$

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

$$\begin{aligned} Z(z) = \frac{V(z)}{I(z)} &= \frac{a_+(e^{-j\beta z} + \Gamma_L e^{j\beta z})}{\frac{1}{Z_0} a_+(e^{-j\beta z} - \Gamma_L e^{j\beta z})} \\ &= Z_0 \frac{1+\Gamma_L e^{2j\beta z}}{1-\Gamma_L e^{2j\beta z}} = Z_0 \frac{1+\Gamma(z)}{1-\Gamma(z)} \end{aligned} \tag{12.1.10}$$

where $\Gamma(z)$ defined in (12.1.6) is used. The above can also be written as

$$Z_n(z) = Z(z)/Z_0 = \frac{1+\Gamma(z)}{1-\Gamma(z)} \tag{12.1.11}$$

where $Z_n(z)$ is the normalized generalized impedance. Conversely, one can write the above as

$$\Gamma(z) = \frac{Z_n(z)-1}{Z_n(z)+1} = \frac{Z(z)-Z_0}{Z(z)+Z_0} \tag{12.1.12}$$

From (12.1.10) above, one gets

$$Z(z) = Z_0 \frac{1+\Gamma_L e^{2j\beta z}}{1-\Gamma_L e^{2j\beta z}} \tag{12.1.13}$$

One can show that by setting $z = -l$, using (12.1.9), and after some algebra,

$$Z(-l) = Z_0 \frac{Z_L + jZ_0 \tan \beta l}{Z_0 + jZ_L \tan \beta l} \tag{12.1.14}$$

12.1.1 Shorted Terminations

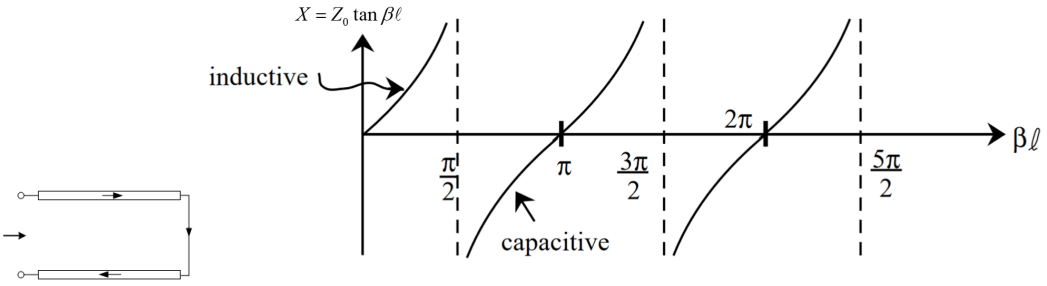


Figure 12.2: The input reactance (X) of a shorted transmission line as a function of its length l.

From (12.1.14) above, when we have a short such that $Z_L = 0$, then

$$Z(-l) = jZ_0 \tan(\beta l) = jX \tag{12.1.15}$$

When $\beta l \ll 1$, then $\tan \beta l \approx \beta l$, and (12.1.15) becomes

$$Z(-l) \cong jZ_0 \beta l \tag{12.1.16}$$

After using that $Z_0 = \sqrt{L/C}$ and that $\beta = \omega\sqrt{LC}$, (12.1.16) becomes

$$Z(-l) \cong j\omega L l \tag{12.1.17}$$

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.

On the other hand, when the length of the shorted line increases, due to the standing wave on the transmission line, certainly parts of the line will have charge accumulation giving rise to strong electric field, while other parts have current flow giving rise to strong magnetic field. Depending on this standing wave pattern, the line can become either capacitive or inductive.

12.1.2 Open Terminations

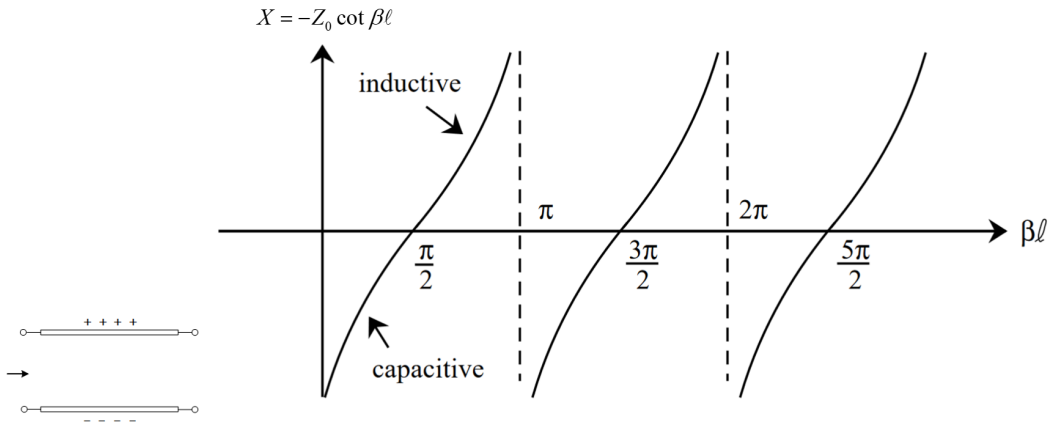


Figure 12.3: The input reactance (X) of an open transmission line as a function of its length l.

When we have an open circuit such that $Z_L = \infty$, then from (12.1.14) above

$$Z(-l) = -jZ_0 \cot(\beta l) = jX \tag{12.1.18}$$

Then, when $\beta l \ll 1$, $\cot(\beta l) \approx 1/\beta l$

$$Z(-l) \approx -j\frac{Z_0}{\beta l} \tag{12.1.19}$$

And then, again using $\beta = \omega\sqrt{LC}$, $Z_0 = \sqrt{L/C}$

$$Z(-l) \approx \frac{1}{j\omega C l} \tag{12.1.20}$$

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 line capacitance or capacitance per unit length.

Again, as shown in Figure 12.3, the impedance at $z = -l$ is purely reactive, 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 shorted or an open terminated line look like an inductor or a capacitor depending on its length l. This effect is shown in Figures 12.2 and 12.3. 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 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.

12.2 Smith Chart

In general, from (12.1.13) and (12.1.14), 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 (invented by P.H. Smith 1939 before the advent of the computer) [88]. 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 imposed a strong impression on these engineers: it also has become an indispensable visual aid for understanding and solving microwave engineering problems.

Equation (12.1.12) indicates that there is a unique 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 (12.1.10) and (12.1.12)

$$\Gamma = \frac{Z_n - 1}{Z_n + 1}, \qquad Z_n = \frac{1 + \Gamma}{1 - \Gamma} \tag{12.2.1}$$

Equations in (12.2.1) are related to a bilinear transform in complex variables [89]. It is a conformal map that maps circles to circles. Such a map is shown in Figure 12.4, 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 12.5 allows one to obtain the corresponding Γ given Z_n and vice versa as indicated in (12.2.1), but using a graphical calculator or the Smith chart.

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 that will absorb energy. Hence, such an impedance load will have reflection coefficient with amplitude less than one, which are points within the unit circle.

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 $|\Gamma| > 1$, and will be 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 it maps to the zero point on the complex Γ plane implying no reflection.

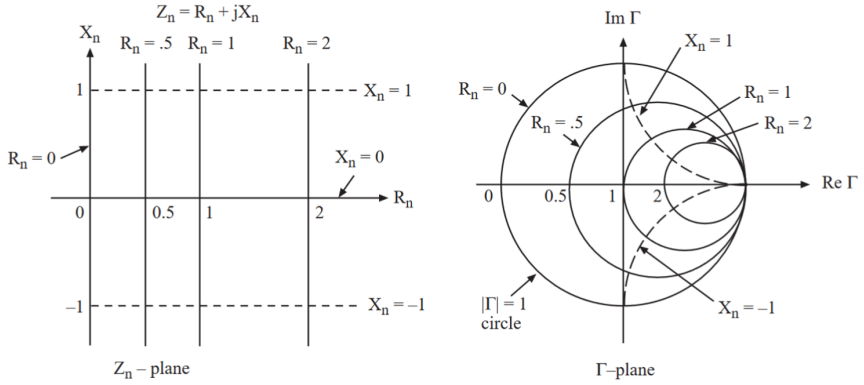


Figure 12.4: Bilinear map of the formulae $\Gamma = \frac{Z_n - 1}{Z_n + 1}$, and $Z_n = \frac{1+\Gamma}{1-\Gamma}$. The chart on the right, called the Smith chart, allows the values of Z_n to be determined quickly given Γ, and vice versa.

The Smith chart also allows one to quickly evaluate the expression

$$\Gamma(-l) = \Gamma_L e^{-2j\beta l} \tag{12.2.2}$$

and its corresponding Z_n not by using (12.2.1) 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

$$\Gamma(-l) = \Gamma_L e^{-4j\pi l/\lambda} \tag{12.2.3}$$

For increasing l, one moves away from the load to the generator (or source), l increases, and the phase is decreasing because of the negative sign. So given a point for Γ_L on the Smith chart, one has negative phase or 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 rotates a half circle around the chart. And when $l = \lambda/2$, the reflection coefficient will rotate a full circle, or back to the original point. Therefore, on the edge of the Smith chart, there are indication as to which direction one should rotate if one were to move toward the generator or toward the load.

Also, for two points diametrically opposite to each other on the Smith chart, Γ changes sign, and it can be shown easily from (12.2.1) that the normalized impedances are reciprocal of each other. Hence, the Smith chart can also be used to find the reciprocal of a complex number quickly. A full blown Smith chart is shown in Figure 12.5.

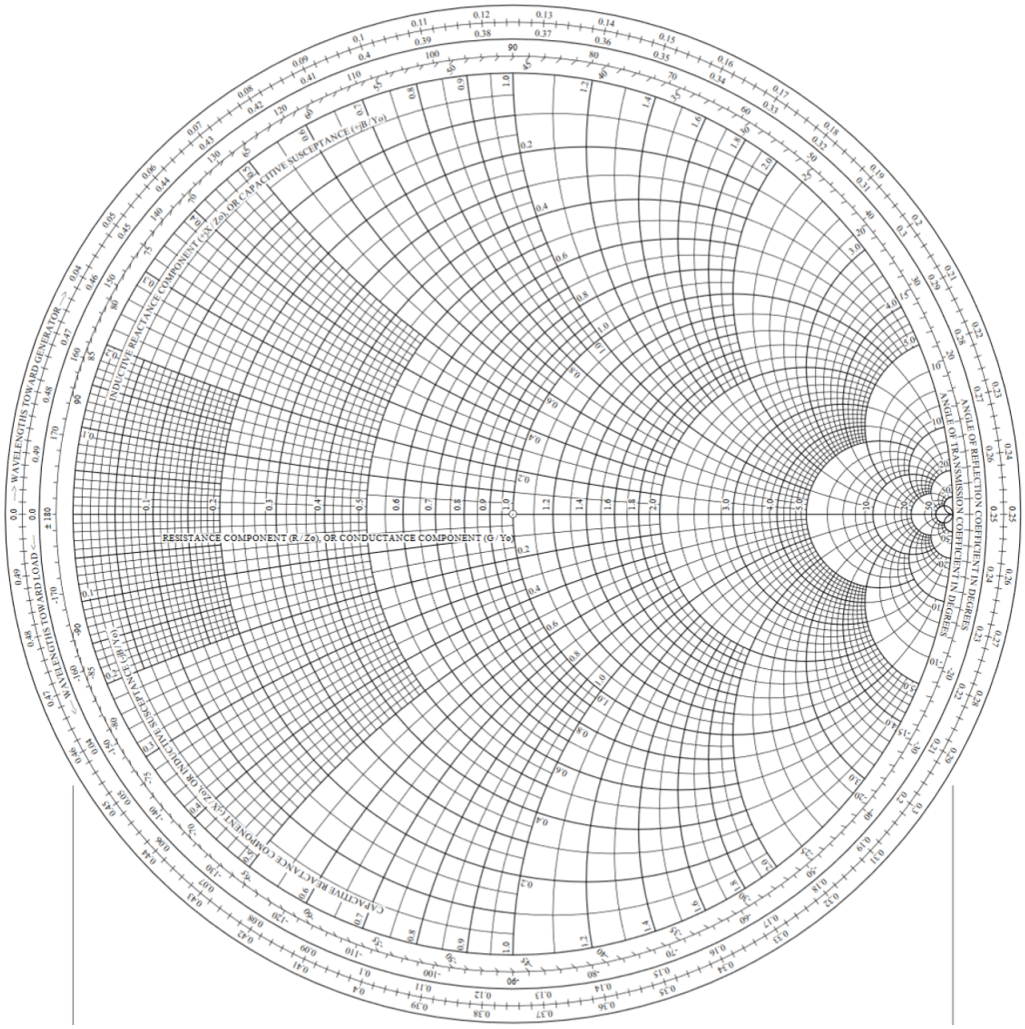


Figure 12.5: The Smith chart in its full glory. It was invented in 1939 before the age of digital computers, but it still allows engineers to do mental estimations and rough calculations with it, because of its simplicity.

12.3 VSWR (Voltage Standing Wave Ratio)

From the previous section, one sees that the voltage and current are not constant in 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 need to be designed as described below.

Due to the interference between that 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

$$\begin{aligned} V(z) &= V_0 e^{-j\beta z} + V_0 e^{j\beta z}\Gamma_L \\ &= V_0 e^{-j\beta z}\left(1 + \Gamma_L e^{2j\beta z}\right) \\ &= V_0 e^{-j\beta z}\left(1 + \Gamma(z)\right) \end{aligned} \tag{12.3.1}$$

where we have used (12.1.6) for $\Gamma(z)$ with $\gamma = j\beta$. Hence, $V(z)$ is not a constant but dependent on z, or

$$|V(z)| = |V_0||1 + \Gamma(z)| \tag{12.3.2}$$

For lack of a better name, this is called the standing wave, even though it is not truly a standing wave.

In Figure 12.6, the relationship between variation of $1 + \Gamma(z)$ as z varies is shown.

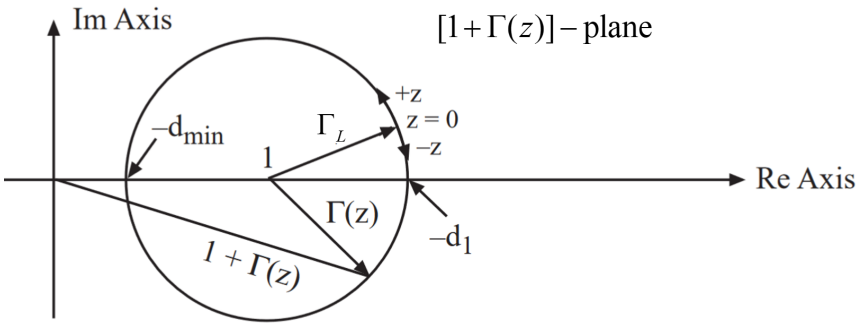


Figure 12.6: The voltage amplitude on a transmission line depends on $|V(z)|$, which is proportional to $|1 + \Gamma(z)|$ per equation (12.3.2). This figure shows how $|1 + \Gamma(z)|$ varies as z varies on a transmission line.

Using the triangular inequality, one gets the lower and upper bounds or that

$$|V_0|(1 - |\Gamma(z)|) \le |V(z)| \le |V_0|(1 + |\Gamma(z)|) \tag{12.3.3}$$

But from (12.1.6) and that β is pure real for a lossless line, then $|\Gamma(z)| = |\Gamma_L|$; hence

$$V_{\min} = |V_0|(1 - |\Gamma_L|) \le |V(z)| \le |V_0|(1 + |\Gamma_L|) = V_{\max} \tag{12.3.4}$$

The voltage standing wave ratio, VSWR is defined to be

$$\text{VSWR} = \frac{V_{\max}}{V_{\min}} = \frac{1 + |\Gamma_L|}{1 - |\Gamma_L|} \tag{12.3.5}$$

Conversely, one can invert the above to get

$$|\Gamma_L| = \frac{\text{VSWR} - 1}{\text{VSWR} + 1} \tag{12.3.6}$$

Hence, the knowledge of voltage standing wave pattern (VSWP), as shown in Figure 12.7, 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.

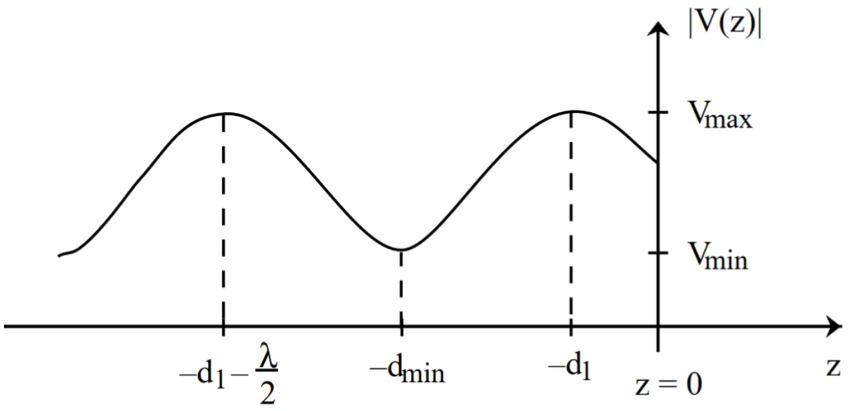


Figure 12.7: The voltage standing wave pattern (VSWP) as a function of z on a load-terminated transmission line.

The phase of Γ_L can also be determined from the measurement of the voltage standing wave pattern. The location of Γ_L in Figure 12.6 is determined by the phase of Γ_L. Hence, the value of d_1 in Figure 12.6 is determined by the phase of Γ_L as well. The length of the transmission line waveguide needed to null 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 12.7. Thus, d_1 is the value where the following equation is satisfied:

$$|\Gamma_L| e^{j\phi_L} e^{-4\pi j(d_1/\lambda)} = |\Gamma_L| \tag{12.3.7}$$

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 12.8: 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).

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 12.8. 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 12.9. 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 short 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, engineering and design are a lot simpler.

The above describes how the impedance of the device-under-test (DUT) can be measured at microwave frequencies. Nowadays, automated network analyzers make these measurements a lot simpler in a microwave laboratory. More resource on microwave measurements can be found on the web, such as in [90].

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. Hence, many experiments are easier to perform at microwave frequencies rather than at optical frequencies.

Many technologies are 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.

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 smaller components, this is an exciting and a largely untapped frontier in electromagnetic technology.

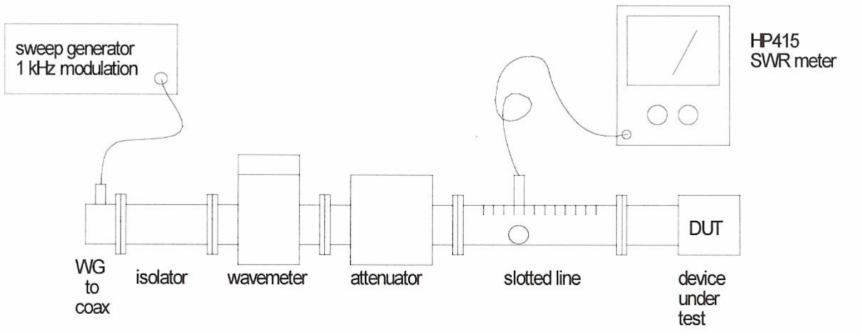


Figure 12.9: An experimental setup for a slotted line measurement (courtesy of Pozar and Knapp, U. Mass [91]).

Lecture 13

Multi-Junction Transmission Lines, Duality Principle

A simple extension of the transmission line theory of the previous lecture is to the case when tranmission 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.

Also, due to the symmetry of Maxwell's equations between the electric field and the magnetic field, once a set of solutions have been found for Maxwell's equations, new solutions can be found by symmetry arguments. This is known as duality principle in electromagnetics.

13.1 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 that we can explain with our textbook examples. Figure 13.1 shows many real world technologies, but they can be approximated with transmission line models as shall be seen.

Figure 13.1: 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 circuit (MIC) area and the monolithic microwave integrated circuit (MMIC) area. An MMIC circuit is shown in Figure 13.2. Many of the components can be approximated by multi-junction transmission lines. They are clear motivation for studying this topic.

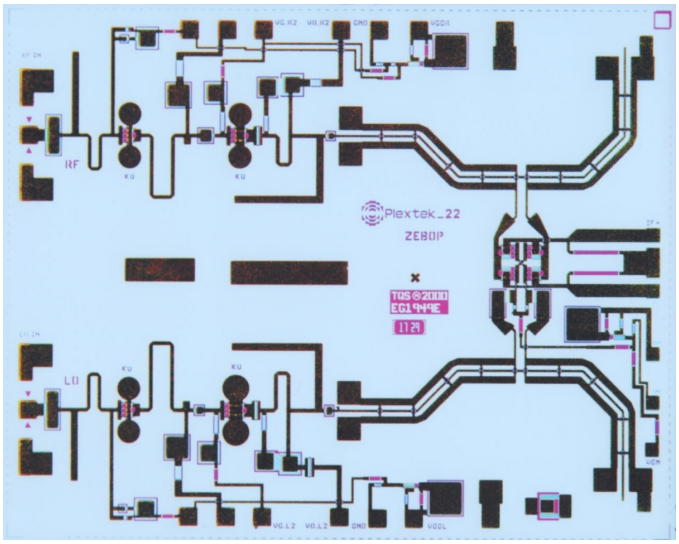


Figure 13.2: A generic GaAs monolithic microwave integrated circuit (MMIC). They are the motivation for studying multi-junction transmission lines (courtesy of Wikipedia).

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. A good reference for such problem is the book by Collin [92], but much of the treatment here is not found in any textbooks.

13.1.1 Single-Junction Transmission Lines

Consider two transmission lines connected at a single junction as shown in Figure 13.3. 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}.

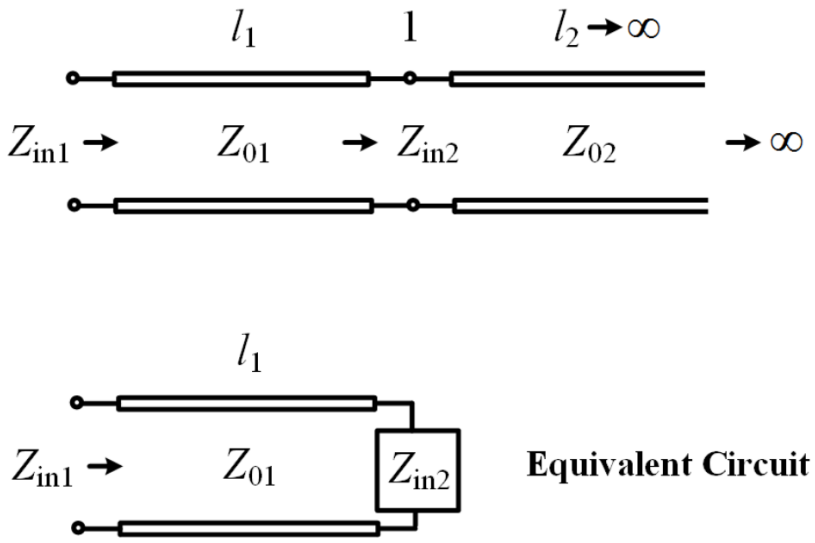


Figure 13.3: A single junction transmission line can be modeled by a equivalent transmission line terminated in a load Z_{in2}.

The impedance of the transmission line at junction 1 looking to the right,using the formula from previously derived,[1] is

$$Z_{in2} = Z_{02}\frac{1+\Gamma_{L,\infty}e^{-2j\beta_2 l_2}}{1-\Gamma_{L,\infty}e^{-2j\beta_2 l_2}} = Z_{02} \tag{13.1.1}$$

Since no reflected wave exists, $\Gamma_{L,\infty} = 0$, 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

[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 $Z_n = \frac{1+\Gamma}{1-\Gamma}$.

13.3 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

$$\Gamma_{12} = \frac{Z_L - Z_{01}}{Z_L + Z_{01}} = \frac{Z_{in2} - Z_{01}}{Z_{in2} + Z_{01}} = \frac{Z_{02} - Z_{01}}{Z_{02} + Z_{01}} \tag{13.1.2}$$

where we have assumed that $Z_L = Z_{in2} = Z_{02}$.

13.1.2 Two-Junction Transmission Lines

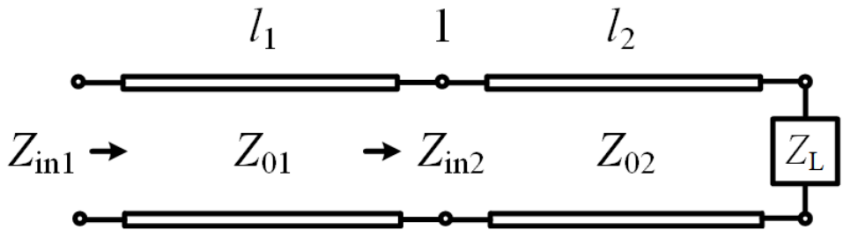


Figure 13.4: A single-junction transmission line with a load Z_L at the far (load) end of the second line. But it can be reduced to the equivalent circuit shown in the bottom of Figure 13.3.

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 13.4. Then, using the formula derived in the previous lecture,

$$Z_{in2} = Z_{02}\frac{1 + \Gamma(-l_2)}{1 - \Gamma(-l_2)} = Z_{02}\frac{1 + \Gamma_{L2}e^{-2j\beta_2 l_2}}{1 - \Gamma_{L2}e^{-2j\beta_2 l_2}} \tag{13.1.3}$$

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

$$\Gamma_{L2} = \frac{Z_L - Z_{02}}{Z_L + Z_{02}} \tag{13.1.4}$$

Now, line 1 sees a load of Z_{in2} hooked at its end. The equivalent circuit is the same as that shown in Figure 13.3. The generalized reflection coefficient at junction 1, which includes all the reflection of waves from its right, is now

$$\tilde{\Gamma}_{12} = \frac{Z_{in2} - Z_{01}}{Z_{in2} + Z_{01}} \tag{13.1.5}$$

Substituting (13.1.3) into (13.1.5), we have

$$\tilde{\Gamma}_{12} = \frac{Z_{02}(\frac{1+\Gamma}{1-\Gamma}) - Z_{01}}{Z_{02}(\frac{1+\Gamma}{1-\Gamma}) + Z_{01}} \tag{13.1.6}$$

where $\Gamma = \Gamma_{L2} e^{-2j\beta_2 l_2}$. The above can be rearranged to give

$$\tilde{\Gamma}_{12} = \frac{Z_{02}(1+\Gamma) - Z_{01}(1-\Gamma)}{Z_{02}(1+\Gamma) + Z_{01}(1-\Gamma)} \tag{13.1.7}$$

Finally, by further rearranging terms, after some somewhat tedious algebra, it can be shown that the above becomes

$$\tilde{\Gamma}_{12} = \frac{\Gamma_{12} + \Gamma}{1 + \Gamma_{12}\Gamma} = \frac{\Gamma_{12} + \Gamma_{L2} e^{-2j\beta_2 l_2}}{1 + \Gamma_{12}\Gamma_{L2} e^{-2j\beta_2 l_2}} \tag{13.1.8}$$

where Γ_{12}, the local reflection coefficient at the junction between line 1 and line 2, is given by (13.1.2), and $\Gamma = \Gamma_{L2} e^{-2j\beta_2 l_2}$ is the general reflection coefficient[2] at $z = -l_2$ due to the load Z_L. In other words,

$$\Gamma_{L2} = \frac{Z_L - Z_{02}}{Z_L + Z_{02}} \tag{13.1.9}$$

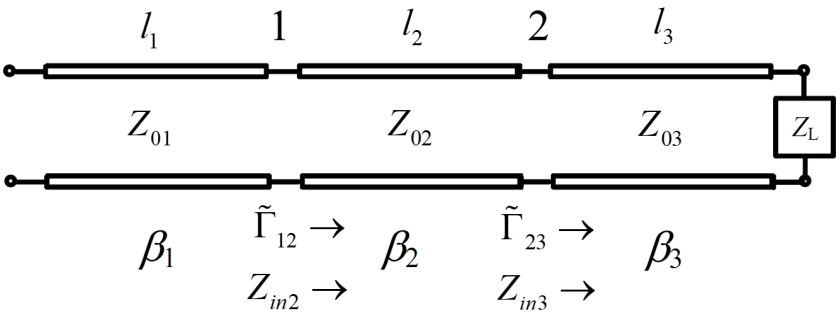


Figure 13.5: A two-junction transmission line with a load Z_L at the far end. The input impedance looking in from the far left can be found recursively using the formula (13.1.12) and (13.1.13).

Equation (13.1.8) is a **powerful formula** for multi-junction transmission lines. Imagine now that we add another section of transmission line as shown in Figure 13.5. We can use

[2]We will use the term "general 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.

the aforementioned method to first find $\tilde{\Gamma}_{23}$, the generalized reflection coefficient at junction 2. Using formula (13.1.8), it is given by

$$\tilde{\Gamma}_{23} = \frac{\Gamma_{23} + \Gamma_{L3} e^{-2j\beta_3 l_3}}{1 + \Gamma_{23}\Gamma_{L3} e^{-2j\beta_3 l_3}} \tag{13.1.10}$$

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 13.5. Here, it is given as

$$\Gamma_{L3} = \frac{Z_L - Z_{03}}{Z_L + Z_{03}} \tag{13.1.11}$$

Given the knowledge of $\tilde{\Gamma}_{23}$, we can use (13.1.8) again to find the new $\tilde{\Gamma}_{12}$ at junction 1. It is now

$$\tilde{\Gamma}_{12} = \frac{\Gamma_{12} + \tilde{\Gamma}_{23} e^{-2j\beta_2 l_2}}{1 + \Gamma_{12}\tilde{\Gamma}_{23} e^{-2j\beta_2 l_2}} \tag{13.1.12}$$

The equivalent circuit is again that shown in Figure 13.3. Therefore, we can use (13.1.12) recursively to find the generalized reflection coefficient for a multi-junction transmission line. Once the reflection coefficient is known, the impedance at that location can also be found. For instance, at junction 1, the impedance is now given by

$$Z_{in2} = Z_{01} \frac{1 + \tilde{\Gamma}_{12}}{1 - \tilde{\Gamma}_{12}} \tag{13.1.13}$$

instead of (13.1.3). In the above, Z_{01} is used because the generalized 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 (13.1.3) 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

$$Z_{in2} = Z_{02} \frac{1 + \tilde{\Gamma}_{23} e^{-2j\beta_2 l_2}}{1 - \tilde{\Gamma}_{23} e^{-2j\beta_2 l_2}} \tag{13.1.14}$$

With some algebraic manipulation, it can be shown that (13.1.13) and (13.1.14) are identical. But (13.1.13) is closer to an experimental scenario where one 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 13.6.

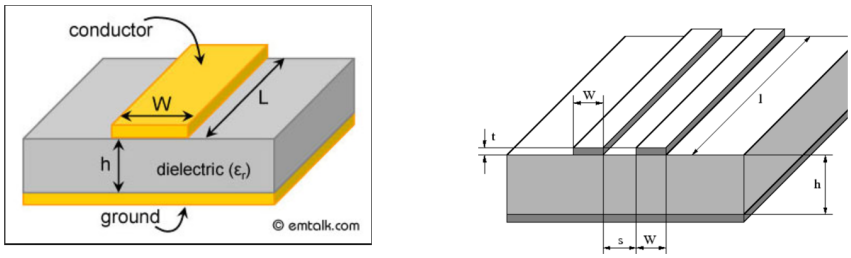


Figure 13.6: Schematic of a microstrip line with the signal line above, and a ground plane below (left). A strip line with each strip carrying currents of opposite polarity (right). A ground plane is not needed in the second case.

13.1.3 Stray Capacitance and Inductance

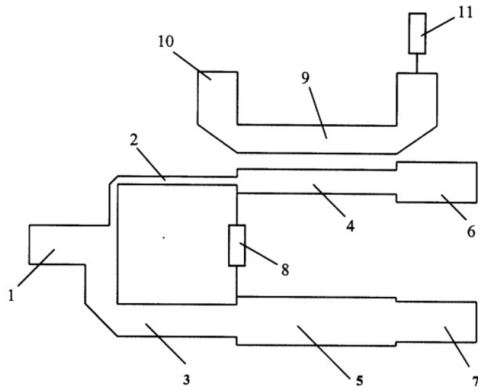


Figure 13.7: A general microwave integrated circuit with different kinds of elements.

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 cumulation or constricted current flow. Excess charge gives rise to excess electric field which corresponds to excess electric stored energy. This can be modeled by stray or parasitic capacitances.

Alternatively, there could be excess or constricted current flow that give rise to excess magnetic field. Excess magnetic field gives rise to excess magnetic stored energy. This can be modeled by stray or parasitic inductances. Hence, a junction can be approximated by a circuit model as shown in Figure 13.8 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.

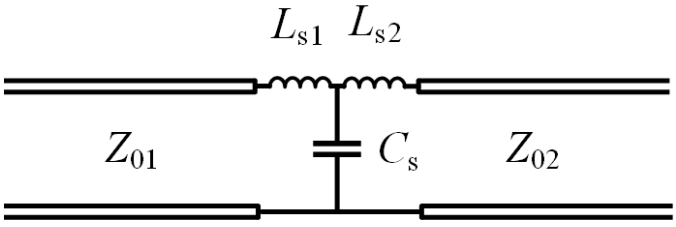


Figure 13.8: A junction between two microstrip lines can be modeled with stray junction capacitance and stray inductances. The capacitance is used to account for excess charges at the junction, while the inductances model the excess current at the junction. They are important as the frequency increases.

13.2 Duality Principle

Duality principle exploits the inherent symmetry of Maxwell's equations. Once a set of **E** and **H** fields has 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

$$\nabla \times \mathbf{E}(\mathbf{r},\omega) = -j\omega\mathbf{B}(\mathbf{r},\omega) - \mathbf{M}(\mathbf{r},\omega) \tag{13.2.1}$$

$$\nabla \times \mathbf{H}(\mathbf{r},\omega) = j\omega\mathbf{D}(\mathbf{r},\omega) + \mathbf{J}(\mathbf{r},\omega) \tag{13.2.2}$$

$$\nabla \cdot \mathbf{B}(\mathbf{r},\omega) = \varrho_m(\mathbf{r},\omega) \tag{13.2.3}$$

$$\nabla \cdot \mathbf{D}(\mathbf{r},\omega) = \varrho(\mathbf{r},\omega) \tag{13.2.4}$$

One way to make Maxwell's equations invariant is to do the following substitutions.

$$\mathbf{E} \to \mathbf{H}, \quad \mathbf{H} \to -\mathbf{E}, \quad \mathbf{D} \to \mathbf{B}, \quad \mathbf{B} \to -\mathbf{D} \tag{13.2.5}$$

$$\mathbf{M} \to -\mathbf{J}, \quad \mathbf{J} \to \mathbf{M}, \quad \varrho_m \to -\varrho, \quad \varrho \to \varrho_m \tag{13.2.6}$$

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

$$\nabla \times \mathbf{E} = -j\omega\overline{\mu} \cdot \mathbf{H} - \mathbf{M} \tag{13.2.7}$$

$$\nabla \times \mathbf{H} = j\omega\overline{\varepsilon} \cdot \mathbf{E} + \mathbf{J} \tag{13.2.8}$$

$$\nabla \cdot \overline{\boldsymbol{\mu}} \cdot \mathbf{H} = \varrho_m \tag{13.2.9}$$

$$\nabla \cdot \overline{\boldsymbol{\varepsilon}} \cdot \mathbf{E} = \varrho \tag{13.2.10}$$

In addition to the above swaps, one need further to swap the material parameters, namely,

$$\overline{\boldsymbol{\mu}} \to \overline{\boldsymbol{\varepsilon}}, \quad \overline{\boldsymbol{\varepsilon}} \to \overline{\boldsymbol{\mu}} \tag{13.2.11}$$

13.2.1 Unusual Swaps[3]

There are other swaps where seemingly the right-hand rule is not preserved, e.g.,

$$\mathbf{E} \to \mathbf{H}, \mathbf{H} \to \mathbf{E}, \mathbf{M} \to -\mathbf{J}, \mathbf{J} \to -\mathbf{M}, \tag{13.2.12}$$

$$\varrho_m \to -\varrho, \varrho \to -\varrho_m, \overline{\boldsymbol{\mu}} \to -\overline{\boldsymbol{\varepsilon}}, \overline{\boldsymbol{\varepsilon}} \to -\overline{\boldsymbol{\mu}} \tag{13.2.13}$$

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.

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

13.3 Fictitious Magnetic Currents

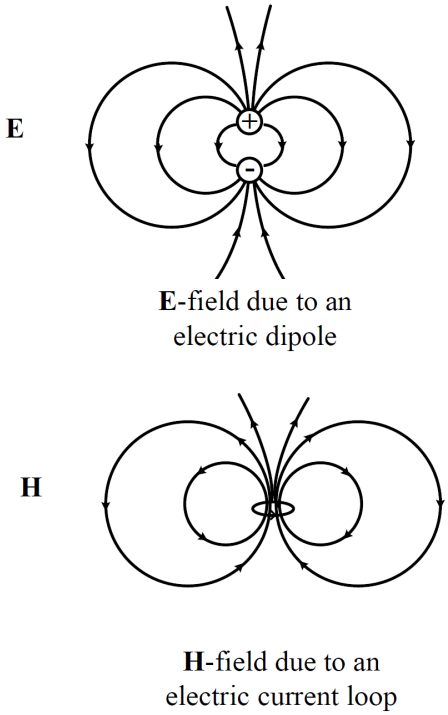


Figure 13.9: Sketches of the electric field due to an electric dipole and the magnetic field due to a electric current loop. The **E** and **H** fields have the same pattern, and can be described by the same formula.

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 made using a small electric current loop (see Figure 13.9). 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 and together. 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 13.10. 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 electric dipole. This is dual to the fact that an electric current loop produces a magnetic field.

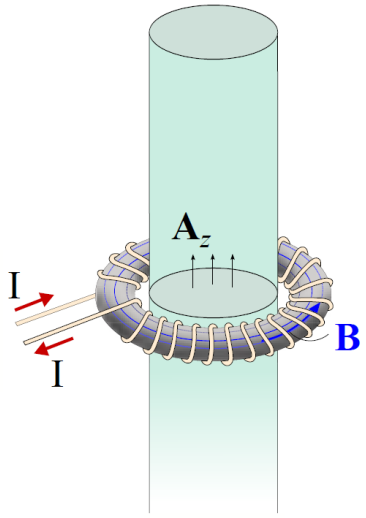


Figure 13.10: A toroidal antenna used to drive an electric current through a conducting cylinder of a dipole. It works similarly to a transformer: one can think of them as the primary and secondary turns (windings) of a transformer (courtesy of Q. S. Liu [93]).

Lecture 14

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 will give rise to the Fresnel reflection and transmission coefficients, and embedded in them are interesting physical phenomena. We will study these interesting phenomena as well.

(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 [31, 32, 44, 50, 54, 65, 79, 83, 85, 86].)

14.1 Reflection and Transmission—Single Interface Case

We will derive the plane-wave reflection coefficients for the single interface case between two dielectric media. These reflection coefficients are also called the Fresnel reflection coefficients because they were first derived by Austin-Jean Fresnel (1788-1827). Note that he lived before the completion of Maxwell's equations in 1865. But when Fresnel derived the reflection 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 [58]).

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. 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.

14.1.1 TE Polarization (Perpendicular or E Polarization)[1]

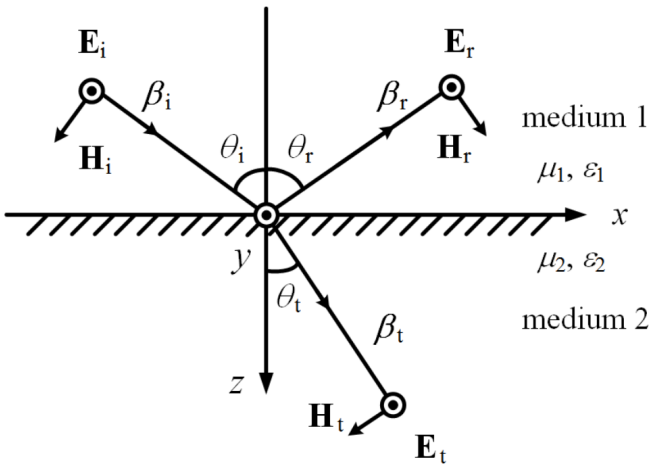


Figure 14.1: A schematic showing the reflection of the TE polarization wave impinging on a dielectric interface.

To set up the above problem, the wave in Region 1 can be written as the superposition or sum of the incident plane wave and reflected plane wave. Here, $\mathbf{E}_i(\mathbf{r})$ and $\mathbf{E}_r(\mathbf{r})$ are the incident and reflected plane waves, respectively. The total field is $\mathbf{E}_i(\mathbf{r}) + \mathbf{E}_r(\mathbf{r})$ which are the phasor representations of the fields. We assume plane wave polarized in the y direction where the wave vectors are $\boldsymbol{\beta}_i = \hat{x}\beta_{ix} + \hat{z}\beta_{iz}$, $\boldsymbol{\beta}_r = \hat{x}\beta_{rx} - \hat{z}\beta_{rz}$, $\boldsymbol{\beta}_t = \hat{x}\beta_{tx} + \hat{z}\beta_{tz}$, respectively for the incident, reflected, and transmitted waves. Then, from Section 7.3, we have

$$\mathbf{E}_i = \hat{y}E_0 e^{-j\boldsymbol{\beta}_i\cdot\mathbf{r}} = \hat{y}E_0 e^{-j\beta_{ix}x - j\beta_{iz}z} \tag{14.1.1}$$

which represents a uniform incident plane wave, and

$$\mathbf{E}_r = \hat{y}R^{TE}E_0 e^{-j\boldsymbol{\beta}_r\cdot\mathbf{r}} = \hat{y}R^{TE}E_0 e^{-j\beta_{rx}x + j\beta_{rz}z} \tag{14.1.2}$$

which is a uniform reflected wave. In Region 2, we only have transmitted plane wave; hence

$$\mathbf{E}_t = \hat{y}T^{TE}E_0 e^{-j\boldsymbol{\beta}_t\cdot\mathbf{r}} = \hat{y}T^{TE}E_0 e^{-j\beta_{tx}x - j\beta_{tz}z} \tag{14.1.3}$$

[1]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.

In the above, the incident wave is known and hence, E_0 is known. From (14.1.2) and (14.1.3), R^{TE} and T^{TE} are unknowns yet to be sought. To find them, we need two boundary conditions to yield two equations.[2] These boundary conditions are tangential $\mathbf{E}$ field continuous and tangential $\mathbf{H}$ field continuous, which are $\hat{n} \times \mathbf{E}$ continuous and $\hat{n} \times \mathbf{H}$ continuous conditions at the interface.

Imposing $\hat{n} \times \mathbf{E}$ continuous at $z = 0$, we get

$$E_0 e^{-j\beta_{ix}x} + R^{TE} E_0 e^{-j\beta_{rx}x} = T^{TE} E_0 e^{-j\beta_{tx}x}, \quad \forall x \tag{14.1.4}$$

where $\forall$ implies "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.[3] 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 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 that $\beta_1 \sin\theta_i = \beta_2 \sin\theta_t$. (This law of refraction that was also known in the Islamic world in the 900 AD. [94]).

The exponential terms or the phase terms on both sides of (14.1.4) are the same. Now, canceling common terms on both sides of the equation (14.1.4), the above simplifies to

$$1 + R^{TE} = T^{TE} \tag{14.1.5}$$

To impose $\hat{n} \times \mathbf{H}$ continuous, one needs to find the $\mathbf{H}$ field using $\nabla \times \mathbf{E} = -j\omega\mu\mathbf{H}$. By letting ∇ with $-j\beta$, then we have $\mathbf{H} = -j\boldsymbol{\beta} \times \mathbf{E}/(-j\omega\mu) = \boldsymbol{\beta} \times \mathbf{E}/(\omega\mu)$. By so doing[4]

$$\mathbf{H}_i = \frac{\boldsymbol{\beta}_i \times \mathbf{E}_i}{\omega\mu_1} = \frac{\boldsymbol{\beta}_i \times \hat{y}}{\omega\mu_1} E_0 e^{-j\boldsymbol{\beta}_i \cdot \mathbf{r}} = \frac{\hat{z}\beta_{ix} - \hat{x}\beta_{iz}}{\omega\mu_1} E_0 e^{-j\boldsymbol{\beta}_i \cdot \mathbf{r}} \tag{14.1.6}$$

$$\mathbf{H}_r = \frac{\boldsymbol{\beta}_r \times \mathbf{E}_r}{\omega\mu_1} = \frac{\boldsymbol{\beta}_r \times \hat{y}}{\omega\mu_1} R^{TE} E_0 e^{-j\boldsymbol{\beta}_r \cdot \mathbf{r}} = \frac{\hat{z}\beta_{rx} + \hat{x}\beta_{rz}}{\omega\mu_1} R^{TE} E_0 e^{-j\boldsymbol{\beta}_r \cdot \mathbf{r}} \tag{14.1.7}$$

$$\mathbf{H}_t = \frac{\boldsymbol{\beta}_t \times \mathbf{E}_t}{\omega\mu_2} = \frac{\boldsymbol{\beta}_t \times \hat{y}}{\omega\mu_2} T^{TE} E_0 e^{-j\boldsymbol{\beta}_t \cdot \mathbf{r}} = \frac{\hat{z}\beta_{tx} - \hat{x}\beta_{tz}}{\omega\mu_2} T^{TE} E_0 e^{-j\boldsymbol{\beta}_t \cdot \mathbf{r}} \tag{14.1.8}$$

Imposing $\hat{n} \times \mathbf{H}$ continuous or H_x continuous at $z = 0$, we have

$$-\frac{\beta_{iz}}{\omega\mu_1} E_0 e^{-j\beta_{ix}x} + \frac{\beta_{rz}}{\omega\mu_1} R^{TE} E_0 e^{-j\beta_{rx}x} = -\frac{\beta_{tz}}{\omega\mu_2} T^{TE} E_0 e^{-j\beta_{tx}x} \tag{14.1.9}$$

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,

$$\beta_{ix}^2 + \beta_{iz}^2 = \beta_{rx}^2 + \beta_{rz}^2 = \omega^2\mu_1\varepsilon_1 = \beta_1^2, \qquad \text{medium 1} \tag{14.1.10}$$

$$\beta_{tx}^2 + \beta_{tz}^2 = \omega^2\mu_2\varepsilon_2 = \beta_2^2, \qquad \text{medium 2} \tag{14.1.11}$$

[2]Here, we will treat this problem as a boundary value problem where the unknowns are sought from equations obtained from boundary conditions.

[3]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.

[4]We note here that field theory is a lot more complicated than transmission line theory. That is the triumph of transmission line theory as well.

Since

$$\beta_{ix} = \beta_{rx} = \beta_{tx} = \beta_x \tag{14.1.12}$$

the above implies that

$$\beta_{iz} = \beta_{rz} = \beta_{1z} \tag{14.1.13}$$

Moreover, $\beta_{tz} = \beta_{2z} \neq \beta_{1z}$ usually since $\beta_1 \neq \beta_2$. Then (14.1.9) simplifies to

$$\frac{\beta_{1z}}{\mu_1}(1 - R^{TE}) = \frac{\beta_{2z}}{\mu_2} T^{TE} \tag{14.1.14}$$

where $\beta_{1z} = \sqrt{\beta_1^2 - \beta_x^2}$, and $\beta_{2z} = \sqrt{\beta_2^2 - \beta_x^2}$.

Solving (14.1.5) and (14.1.14) for R^{TE} and T^{TE} yields

$$R^{TE} = \left(\frac{\beta_{1z}}{\mu_1} - \frac{\beta_{2z}}{\mu_2}\right) \Big/ \left(\frac{\beta_{1z}}{\mu_1} + \frac{\beta_{2z}}{\mu_2}\right) \tag{14.1.15}$$

$$T^{TE} = 2\left(\frac{\beta_{1z}}{\mu_1}\right) \Big/ \left(\frac{\beta_{1z}}{\mu_1} + \frac{\beta_{2z}}{\mu_2}\right) \tag{14.1.16}$$

14.1.2 TM Polarization (Parallel or H Polarization)[5]

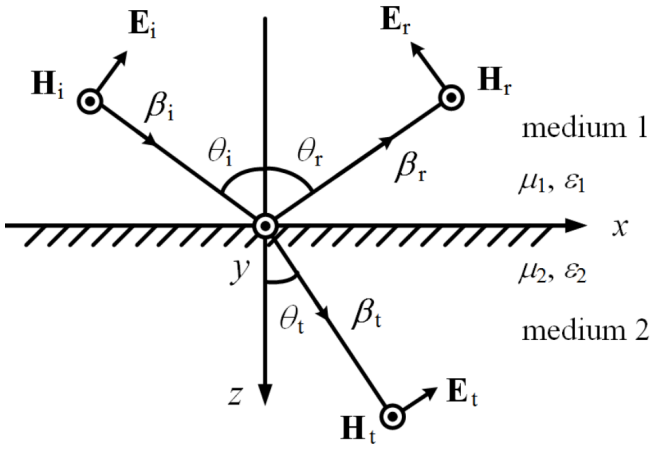


Figure 14.2: A similar schematic showing the reflection of the TM polarization wave impinging on a dielectric interface. The solution to this problem can be easily obtained by invoking duality principle.

[5]Also known as TM$_z$ polarization.

The solution to the TM polarization case can be obtained by invoking 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 14.2. The reflection coefficient for the TM magnetic field is then

$$R^{TM} = \left(\frac{\beta_{1z}}{\varepsilon_1} - \frac{\beta_{2z}}{\varepsilon_2}\right) \Big/ \left(\frac{\beta_{1z}}{\varepsilon_1} + \frac{\beta_{2z}}{\varepsilon_2}\right) \tag{14.1.17}$$

$$T^{TM} = 2\left(\frac{\beta_{1z}}{\varepsilon_1}\right) \Big/ \left(\frac{\beta_{1z}}{\varepsilon_1} + \frac{\beta_{2z}}{\varepsilon_2}\right) \tag{14.1.18}$$

Please remember 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.

14.2 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 look at them next.

14.2.1 Total Internal Reflection

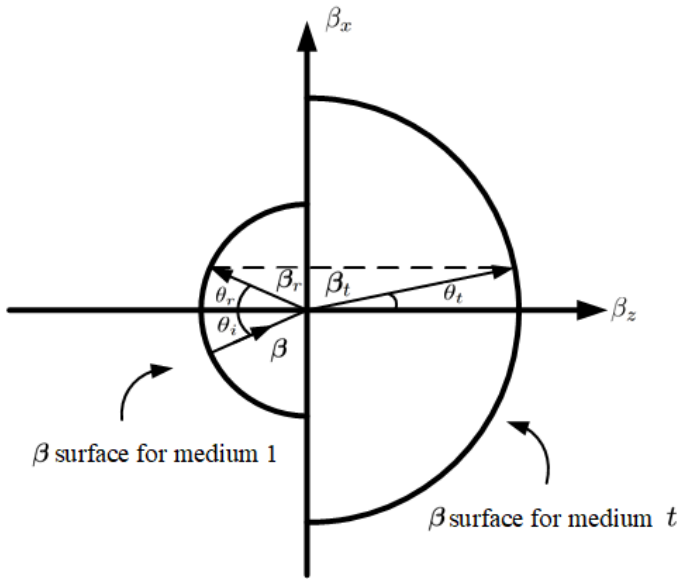


Figure 14.3: Wave-number surfaces in two regions showing the phase matching condition. The wave number in medium t is larger than the wave number in medium 0. The wave vectors for the incident wave, reflected wave, and transmitted wave have to be aligned in such a way that their components parallel to the interface are equal in order to satisfy the phase-matching condition. One can see that Snell's law is satisfied when the phase-matching condition is satisfied.

Total internal reflection comes about because of phase matching (also called momentum matching). This phase-matching condition can be illustrated using β-surfaces (same as k-surfaces in some literature), as shown in Figure 14.3. It turns out that because of phase matching, for certain interfaces, β_{2z} becomes pure imaginary.

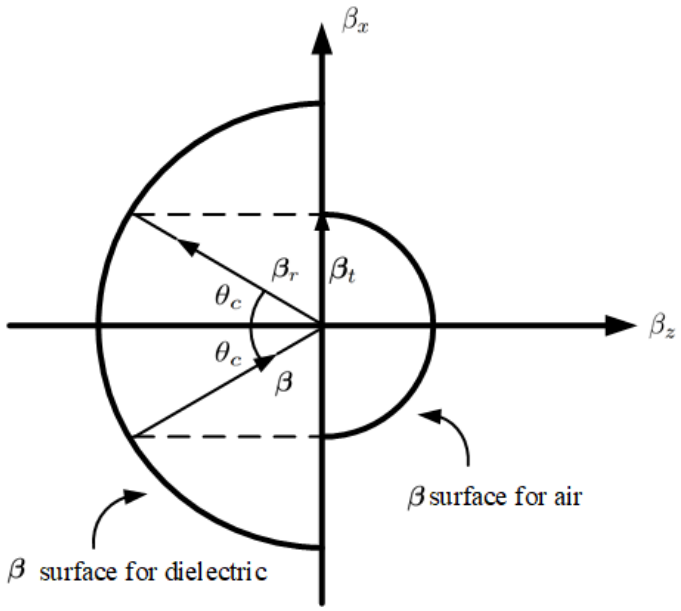


Figure 14.4: Wave-number surfaces in two regions showing the phase matching condition. The wave number in medium t is smaller than the wave number in medium 0. The figure shows in incident wave vector coming in at the critical angle. Then the transmitted wave vector is parallel to the interface as shown. When the incident angle is larger than the critical angle, β_z becomes an imaginary number the wave vector in Region t is complex and cannot be drawn.

As shown in Figures 14.3 and 14.4, 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 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 the $\boldsymbol{\beta}$ vectors are equal to each other as shown. One sees that $\theta_i = \theta_r$ in Figure 14.4, and also as θ_i increases, θ_t increases. For an optically less dense medium where $\beta_2 < \beta_1$, according to the Snell's law of refraction, the transmitted $\boldsymbol{\beta}$ will refract away from the normal, as seen in the figure (where k is synonymous with our β). 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 14.4).

Since $\beta_{ix} = \beta_1 \sin\theta_i = \beta_{rx} = \beta_1 \sin\theta_r = \beta_2$, or

$$\sin\theta_r = \sin\theta_i = \sin\theta_c = \frac{\beta_2}{\beta_1} = \frac{\sqrt{\mu_2\varepsilon_2}}{\sqrt{\mu_1\varepsilon_1}} = \frac{n_2}{n_1} \tag{14.2.1}$$

where n_1 is the reflective 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,

$$\theta_c = \sin^{-1}(n_2/n_1) \tag{14.2.2}$$

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 reflection coefficient (14.1.15) becomes of the form

$$R^{TE} = (A - jB)/(A + jB) \tag{14.2.3}$$

Since the numerator is the complex conjugate of the denominator. It is clear that $|R^{TE}| = 1$ 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 14.5 [58]. A wave that travels laterally along the surface of two media is also known as lateral waves [95, 96].

(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.)

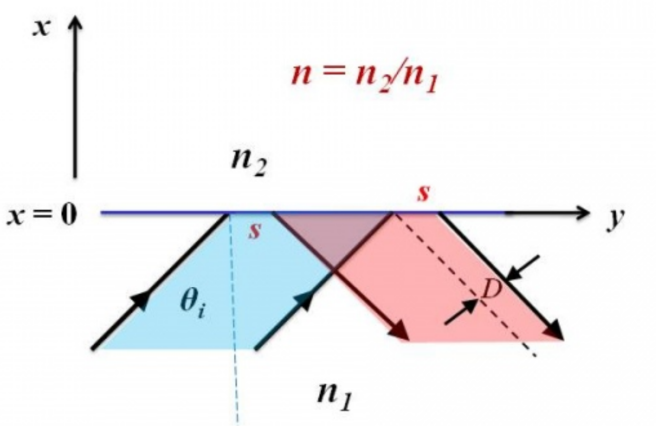


Figure 14.5: Goos-Hanschen Shift. A phase delay is equivalent to a time delay (courtesy of Paul R. Berman (2012), Scholarpedia, 7(3):11584 [97]).

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 14.6. The guidance of waves in an optical fiber works by similar mechanism of total internal reflection, as shown in Figure 14.7. 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 [98].

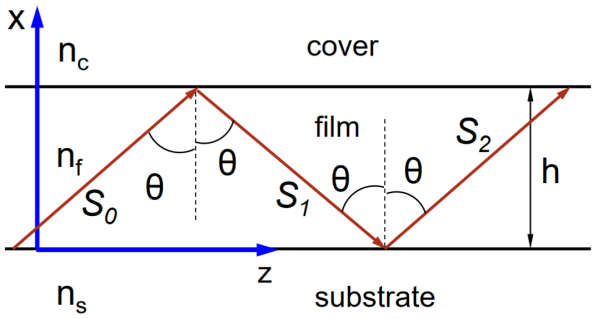


Figure 14.6: The total internal reflections at the two interfaces of a thin-film waveguide can be used to guide an optical wave (courtesy of E.N. Glytsis, NTUA, Greece [99]).

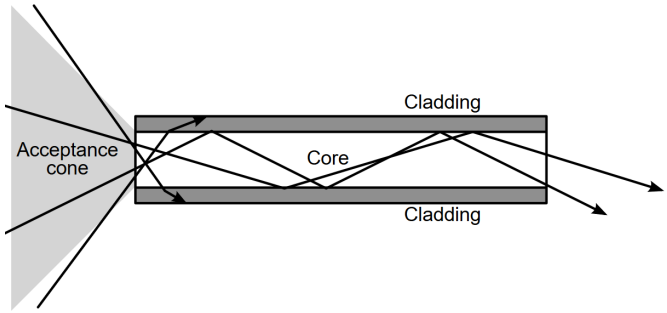


Figure 14.7: An optical fiber consists of a core and a cladding region. Total internal reflections occur at the core-cladding interface. They can guide an optical wave in the fiber (courtesy of Wikepedia [100]).

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 14.8 shows a submarine cable from 1869 using coaxial cable, and one used in the modern world using optical fiber.

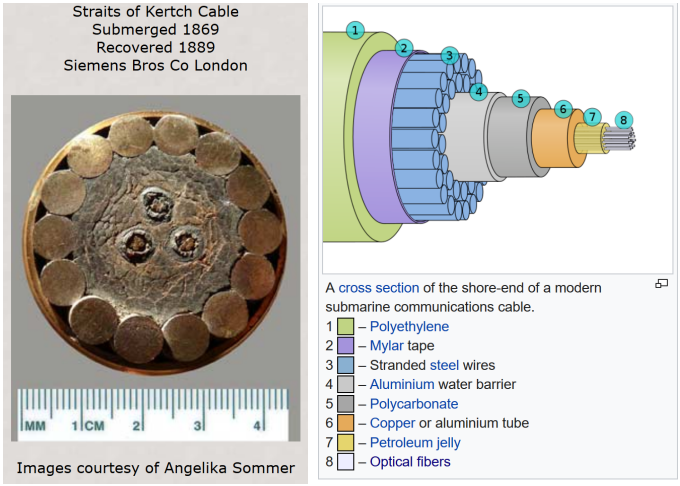


Figure 14.8: The picture of an old 1869 submarine cable made of coaxial cables (left), and modern submarine cable made of optical fibers (right) (courtesy of Atlantic-Cable [101], and Wikipedia [102]).

Lecture 15

More on Interesting Physical Phenomena, Homomorphism, Plane Waves, and 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 [103, 104] and that of surface plasmon resonance, or polariton [105, 106].

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.

15.1 Brewster's Angle

We will continue with understanding some interesting phenomena associated with the single-interface problem starting with the Brewster's angle.

[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 [107].

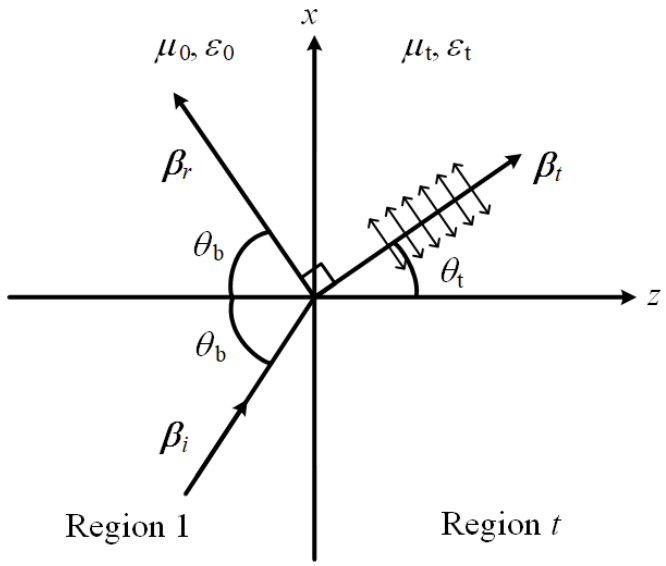


Figure 15.1: A figure showing a plane wave being reflected and transmitted at the Brewster's angle. In Region t, the polarization current or dipoles are all pointing in the β_r direction, and hence, there is no radiation in that direction.

Brewster angle was discovered in 1815 [103, 104]. 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, the TM polarization for light behaves differently from TE polarization. Hence, we shall focus on the reflection and transmission of the TM polarization of light, and we reproduce the previously derived TM reflection coefficient here:

$$R^{TM} = \left(\frac{\beta_{1z}}{\varepsilon_1} - \frac{\beta_{2z}}{\varepsilon_2}\right) \Big/ \left(\frac{\beta_{1z}}{\varepsilon_1} + \frac{\beta_{2z}}{\varepsilon_2}\right) \tag{15.1.1}$$

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

$$\varepsilon_2 \beta_{1z} = \varepsilon_1 \beta_{2z} \tag{15.1.2}$$

Squaring the above, making the note that $\beta_{iz} = \sqrt{\beta_i^2 - \beta_x^2}$, one gets

$$\varepsilon_2{}^2(\beta_1{}^2 - \beta_x{}^2) = \varepsilon_1{}^2(\beta_2{}^2 - \beta_x{}^2) \tag{15.1.3}$$

Solving the above, assuming $\mu_1 = \mu_2 = \mu$, gives

$$\beta_x = \omega\sqrt{\mu}\sqrt{\frac{\varepsilon_1\varepsilon_2}{\varepsilon_1 + \varepsilon_2}} = \beta_1 \sin\theta_1 = \beta_2 \sin\theta_2 \tag{15.1.4}$$

The latter two equalities come from phase matching at the interface or Snell's law. Therefore,

$$\sin\theta_1 = \sqrt{\frac{\varepsilon_2}{\varepsilon_1+\varepsilon_2}}, \quad \sin\theta_2 = \sqrt{\frac{\varepsilon_1}{\varepsilon_1+\varepsilon_2}} \tag{15.1.5}$$

or squaring the above and adding them,

$$\sin^2\theta_1 + \sin^2\theta_2 = 1, \tag{15.1.6}$$

Then, assuming that θ_1 and θ_2 are less than $\pi/2$, and using the identity that $\cos^2\theta_1 + \sin^2\theta_1 = 1$, we infer that $\cos^2\theta_1 = \sin^2\theta_2$. Then it follows that

$$\sin\theta_2 = \cos\theta_1 \tag{15.1.7}$$

In other words,

$$\theta_1 + \theta_2 = \pi/2 \tag{15.1.8}$$

The above formula can be used to explain why at Brewster angle, no light is reflected back to Region 1. Figure 15.1 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, which can be verified heuristically by field sketch and looking at the Poynting 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 $\theta_1 \cong \theta_2 \cong \pi/4$, and (15.1.8) is satisfied.

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, $|R^{TM}| \leq |R^{TE}|$ as shown in the plots in Figure 15.2. Then 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,[2] they will filter out and 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 15.3.

[2]Defined as one that will allow vertical polarization to pass through.

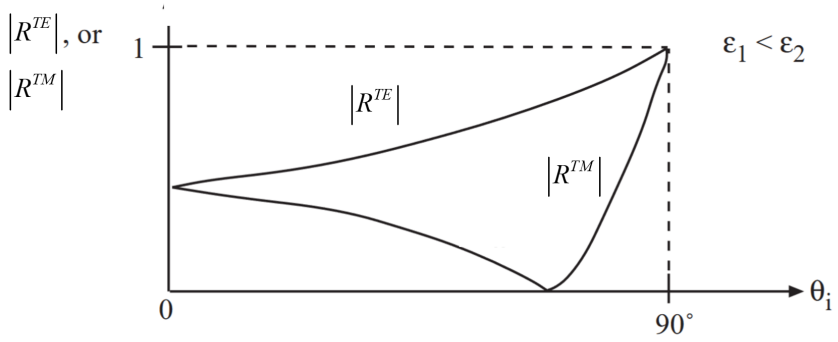


Figure 15.2: Because $|R^{TM}|$ has to through a null when $\theta_i = \theta_b$, therefore, plots of $|R^{TM}| \leq |R^{TE}|$ for all θ_i as shown above.

Figure 15.3: 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 [104].

15.1.1 Surface Plasmon Polariton

Surface plasmon polariton 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 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

$$-\varepsilon_2 \beta_{1z} = \varepsilon_1 \beta_{2z} \tag{15.1.9}$$

When R^{TM} becomes infinite, it implies that a reflected wave exists when there is no incident wave. Or when $H_{\text{ref}} = H_{\text{inc}} R^{TM}$, and $R^{TM} = \infty$, H_{inc} can be zero, and then H_{ref} can assume

any value.[3] 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, is a guided mode propagating in the x direction.

Solving (15.1.9) after squaring it, as in the Brewster angle case, yields

$$\beta_x = \omega\sqrt{\mu}\sqrt{\frac{\varepsilon_1\varepsilon_2}{\varepsilon_1+\varepsilon_2}} \tag{15.1.10}$$

This is the same equation for the Brewster angle except now that ε_2 is negative. Even if $\varepsilon_2 < 0$, but $\varepsilon_1 + \varepsilon_2 < 0$ is still possible so that the expression under the square root sign (15.1.10) is positive. Thus, β_x can be pure real. The corresponding β_{1z} and β_{2z} in (15.1.9) can be pure imaginary as explained below, and (15.1.9) can still be satisfied.

This corresponds to a guided wave propagating in the x direction. When this happens,

$$\beta_{1z} = \sqrt{{\beta_1}^2 - {\beta_x}^2} = \omega\sqrt{\mu}\left[\varepsilon_1\left(1 - \frac{\varepsilon_2}{\varepsilon_1+\varepsilon_2}\right)\right]^{1/2} = \omega\sqrt{\mu}\left[\frac{\varepsilon_1^2}{\varepsilon_1+\varepsilon_2}\right]^{1/2} \tag{15.1.11}$$

Since $\varepsilon_2 < 0$, $\varepsilon_2/(\varepsilon_1+\varepsilon_2) > 1$, then 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 evanescent waves. This mode is shown in Figure 15.4, and is the only case in electromagnetics where a single interface can guide a surface wave, while such phenomenon abounds for elastic waves.

When one operates close to the resonance of the mode so that the denominator in (15.1.10) 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. Hence, 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 offered by the total internal reflection, which is $\beta_z = \sqrt{\beta_t^2 - \beta_x^2}$ where β_x is no larger than β_1. It portends use in tightly packed optical components, and has caused some excitement in the optics community.

[3] In other words, infinity times zero is undefined. 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.

https://en.wikipedia.org/wiki/Surface_plasmon

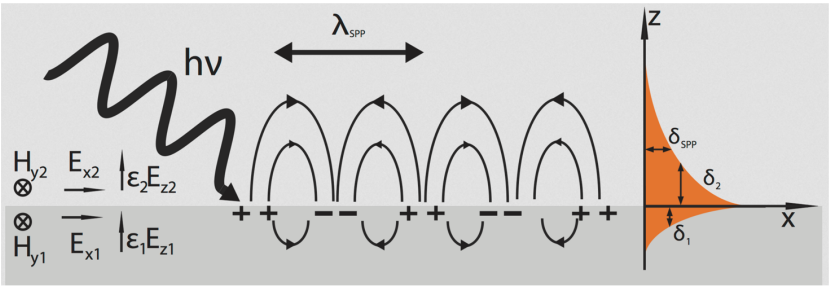


Figure 15.4: Figure showing a surface plasmonic mode propagating at an air-plasma interface. As in all resonant systems, a resonant mode entails the exchange of energies. In the case of surface plasmonic resonance, the energy is exchanged between the kinetic energy of the electrons and the energy store in the electic field (courtesy of Wikipedia [108]).

15.2 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 plane waves through layered medium can be mapped into the multi-section transmission line problem due to mathematical homomorphism between the two problems. Hence, we can kill two birds with one stone: apply all the transmission line techniques and equations that we have learnt to solve for the solutions of waves through layered medium problems.[4]

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 become

$$\boldsymbol{\beta}\times\mathbf{E} = \omega\mu\mathbf{H} \tag{15.2.1}$$

$$\boldsymbol{\beta}\times\mathbf{H} = -\omega\varepsilon\mathbf{E} \tag{15.2.2}$$

for a general isotropic homogeneous medium. We will specialize these equations for different polarizations.

[4]This treatment is not found elsewhere, and is peculiar to these lecture notes.

15.2.1 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 $\mathbf{E} = \hat{y}E_y$, $\mathbf{H} = \hat{x}H_x + \hat{z}H_z$. Then we have from (15.2.1)

$$\beta_z E_y = -\omega\mu H_x \tag{15.2.3}$$

$$\beta_x E_y = \omega\mu H_z \tag{15.2.4}$$

From (15.2.2), we have

$$\beta_z H_x - \beta_x H_z = -\omega\varepsilon E_y \tag{15.2.5}$$

The above equations involve three variables, E_y, H_x, and H_z. 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. Then, expressing H_z in terms of E_y from (15.2.4), we can show from (15.2.5) that

$$\begin{aligned}\beta_z H_x &= -\omega\varepsilon E_y + \beta_x H_z = -\omega\varepsilon E_y + \frac{\beta_x^2}{\omega\mu} E_y \\ &= -\omega\varepsilon(1 - \beta_x^2/\beta^2)E_y = -\omega\varepsilon\cos^2\theta E_y\end{aligned} \tag{15.2.6}$$

where $\beta_x = \beta\sin\theta$ has been used.

Eqns. (15.2.3) and (15.2.6) can be written to look like the telegrapher's equations by letting $-j\beta_z \to d/dz$ to get

$$\frac{d}{dz}E_y = j\omega\mu H_x \tag{15.2.7}$$

$$\frac{d}{dz}H_x = j\omega\varepsilon\cos^2\theta E_y \tag{15.2.8}$$

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

$$Z_0 = \sqrt{\frac{L}{C}} = \sqrt{\frac{\mu}{\varepsilon}}\frac{1}{\cos\theta} = \sqrt{\frac{\mu}{\varepsilon}}\frac{\beta}{\beta_z} = \frac{\omega\mu}{\beta_z} \tag{15.2.9}$$

The above $\omega\mu/\beta_z$ is the wave impedance for a propagating plane wave with propagation direction or the $\boldsymbol{\beta}$ inclined with an angle θ respect to the z axis. It is analogous to the characteristic impedance Z_0 of a transmission line. When $\theta = 0$, the wave impedance 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 13.1.1. The equivalent characteristic impedances of these two regions are then

$$Z_{01} = \frac{\omega\mu_1}{\beta_{1z}}, \quad Z_{02} = \frac{\omega\mu_2}{\beta_{2z}} \tag{15.2.10}$$

We can use the above to find Γ_{12} as given by

$$\Gamma_{12} = \frac{Z_{02} - Z_{01}}{Z_{02} + Z_{01}} = \frac{(\mu_2/\beta_{2z}) - (\mu_1/\beta_{1z})}{(\mu_2/\beta_{2z}) + (\mu_1/\beta_{1z})} \tag{15.2.11}$$

The above is the same as the Fresnel reflection coefficient found earlier for TE waves or R^{TE} after some simple re-arrangement.

Assuming that we have a single junction transmission line, one can define a transmission coefficient given by

$$T_{12} = 1 + \Gamma_{12} = \frac{2Z_{02}}{Z_{02} + Z_{01}} = \frac{2(\mu_2/\beta_{2z})}{(\mu_2/\beta_{2z}) + (\mu_1/\beta_{1z})} \tag{15.2.12}$$

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}.

15.2.2 TM or TM$_z$ Waves

For the TM polarization, by invoking duality principle, the corresponding equations are, from (15.2.7) and (15.2.8),

$$\frac{d}{dz} H_y = -j\omega\varepsilon E_x \tag{15.2.13}$$

$$\frac{d}{dz} E_x = -j\omega\mu \cos^2\theta H_y \tag{15.2.14}$$

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;

$$E_y \to V, \quad H_y \to I, \quad \mu\cos^2\theta \to L, \quad \varepsilon \to C \tag{15.2.15}$$

Then, the equivalent characteristic impedance is now

$$Z_0 = \sqrt{\frac{L}{C}} = \sqrt{\frac{\mu}{\varepsilon}}\cos\theta = \sqrt{\frac{\mu}{\varepsilon}}\frac{\beta_z}{\beta} = \frac{\beta_z}{\omega\varepsilon} \tag{15.2.16}$$

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 again that this wave impedance 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 (15.2.16), we arrive at

$$\Gamma_{12} = \frac{(\beta_{2z}/\varepsilon_2) - (\beta_{1z}/\varepsilon_1)}{(\beta_{2z}/\varepsilon_2) + (\beta_{1z}/\varepsilon_1)} \tag{15.2.17}$$

Notice that (15.2.17) has a sign difference from the definition of R^{TM} derived earlier in the last lecture. The reason is that R^{TM} 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.[5] A voltage transmission coefficient can be defined to be

$$T_{12} = 1 + \Gamma_{12} = \frac{2(\beta_{2z}/\varepsilon_2)}{(\beta_{2z}/\varepsilon_2) + (\beta_{1z}/\varepsilon_1)} \tag{15.2.18}$$

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.

[5]This is often the source of confusion for these reflection and transmission coefficients.

Lecture 16

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 [109]. The earth can be approximated by planarly layered media to capture the important physics behind the wave propagation. Many microwave components are made by planarly layered 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 or laminated materials.

16.1 Waves in Layered Media

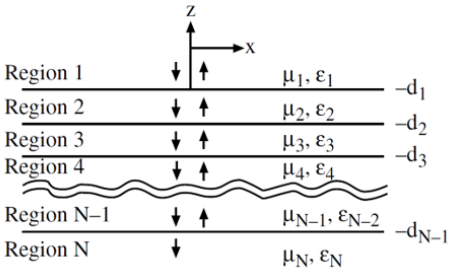


Figure 16.1: Waves in layered media. A wave entering the medium from above can be multiply reflected before emerging from the top again.

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. This treatment is not found in any other textbooks.

16.1.1 Generalized Reflection Coefficient for Layered Media

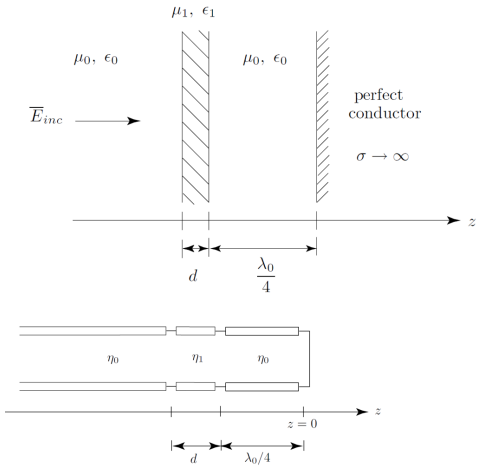


Figure 16.2: The equivalence of a layered medium problem to a transmission line problem. This equivalence is possible even for oblique incidence. For normal incidence, the wave impedance becomes intrinsic impedances (courtesy of J.A. Kong, Electromagnetic Wave Theory).

Because of the homomorphism between transmission line problems and plane waves in layered medium problems, one can capitalize on using the multi-section transmission line formulas for generalized reflection coefficient, which is

$$\tilde{\Gamma}_{12} = \frac{\Gamma_{12} + \tilde{\Gamma}_{23} e^{-2j\beta_2 l_2}}{1 + \Gamma_{12}\tilde{\Gamma}_{23} e^{-2j\beta_2 l_2}} \tag{16.1.1}$$

In the above, Γ_{12} is the local reflection at the 1,2 junction, whereas $\tilde{\Gamma}_{ij}$ are the generalized 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 16.1 and 16.2.

Using the result from the multi-junction transmission line, by analogy we can write down the generalized reflection coefficient for a layered medium with an incident wave at the 1,2 interface, including multiple reflections from behind the interface. It is given by

$$\tilde{R}_{12} = \frac{R_{12} + \tilde{R}_{23} e^{-2j\beta_{2z} l_2}}{1 + R_{12}\tilde{R}_{23} e^{-2j\beta_{2z} l_2}} \tag{16.1.2}$$

where R_{12} is the local Fresnel reflection coefficient and $\tilde{R}_{ij}$ is the generalized 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 generalized 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 generalized reflection coefficient $\tilde{R}_{23}$.

Figure 16.2 shows the case of a normally incident wave into a layered media. For this case, the wave impedance becomes the intrinsic impedance of homogeneous space.

16.1.2 Ray Series Interpretation of Generalized Reflection Coefficient

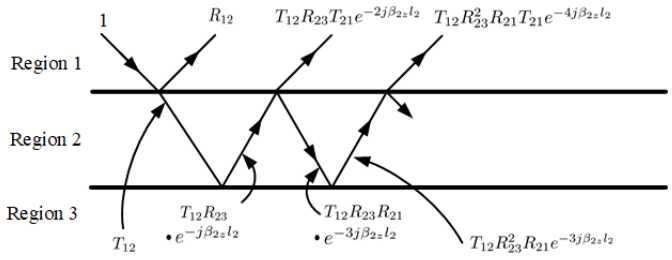


Figure 16.3: The expression of the generalized reflection coefficient into a ray series. Here, $l_2 = d_2 - d_1$ is the thickness of the slab (courtesy of [110]).

For simplicity, we will assume that $\tilde{R}_{23} = R_{23}$ in this section. By manipulation, one can convert the generalized reflection coefficient $\tilde{R}_{12}$ into a form that has a ray physics interpretation. By adding the term

$$R_{12}^2 R_{23} e^{-2j\beta_{2z}l_2}$$

to the numerator of (16.1.2), it can be shown to become

$$\tilde{R}_{12} = R_{12} + \frac{R_{23}e^{-2j\beta_{2z}l_2}(1 - R_{12}^2)}{1 + R_{12}R_{23}e^{-2j\beta_{2z}l_2}} \tag{16.1.3}$$

By using the fact that $R_{12} = -R_{21}$ and that $T_{ij} = 1 + R_{ij}$, the above can be rewritten as

$$\tilde{R}_{12} = R_{12} + \frac{T_{12}T_{21}R_{23}e^{-2j\beta_{2z}l_2}}{1 + R_{12}R_{23}e^{-2j\beta_{2z}l_2}} \tag{16.1.4}$$

Then using the fact that $(1 - x)^{-1} = 1 + x + x^2 + \ldots +$, the above can be rewritten as

$$\tilde{R}_{12} = R_{12} + T_{12}R_{23}T_{21}e^{-2j\beta_{2z}l_2} + T_{12}R_{23}^2R_{21}T_{21}e^{-4j\beta_{2z}l_2} + \cdots . \tag{16.1.5}$$

The above allows us to elucidate the physics of each of the terms. 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 16.3). Hence, the expansion of (16.1.2) into 16.1.5 renders a lucid interpretation for the generalized reflection coefficient. Consequently, the series in 16.1.5 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 also the consequence of expanding the denominator of the second term in (21). Hence, the denominator of the second term in (21) can be physically interpreted as a consequence of multiple reflections within region 2.

16.1.3 Guided Modes from Generalized Reflection Coefficients

We shall discuss finding guided waves in a layered medium next using the generalized reflection coefficient. For a general guided wave along the longitudinal direction parallel to the interfaces (x direction in our notation), the wave will propagate in the manner of

$$e^{-j\beta_x x}$$

For instance, the surface plasmon mode that we found previously can be thought of as a wave propagating in the x direction. The wave number in the x direction, β_x for the surface plasmon case can be found in closed form, but it is a rather complicated function of frequency. Hence, it has very interesting phase velocity, which is frequency dependent. From Section 15.1.1, we have derived that for a surface plasmon mode,

$$\beta_x = \omega\sqrt{\mu}\sqrt{\frac{\varepsilon_1\varepsilon_2}{\varepsilon_1+\varepsilon_2}} = \beta_1 \sin\theta_1 = \beta_2 \sin\theta_2 \tag{16.1.6}$$

Because ε_2 is a plasma medium with complex frequency dependence, β_x is in general, a complicated function of ω or frequency. Therefore, this wave has very interesting phase and group velocities, which is typical of waveguide modes. Hence, it is prudent to understand what phase and group velocities are before studying waveguide modes in greater detail.

16.2 Phase Velocity and Group Velocity

Now that we know how a medium can be frequency dispersive in a complicated fashion as in the Drude-Lorentz-Sommerfeld (DLS) model, we are ready to investigate the difference between the phase velocity and the group velocity

16.2.1 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

$$\begin{aligned} V(z,t) &= V_0 \cos(\omega t - kz + \alpha) \\ &= V_0 \cos\left[k\left(\frac{\omega}{k}t - z\right) + \alpha\right] \end{aligned} \tag{16.2.1}$$

This sinusoidal signal moves with a velocity

$$v_{ph} = \frac{\omega}{k} \tag{16.2.2}$$

where, for example, $k = \omega\sqrt{\mu\varepsilon}$, inside a simple coax. Hence,

$$v_{ph} = 1/\sqrt{\mu\varepsilon} \tag{16.2.3}$$

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. Therefore, signals with different ω's will travel with different phase velocities.

More bizarre still, what if the coax is filled with a plasma medium where

$$\varepsilon = \varepsilon_0 \left(1 - \frac{{\omega_p}^2}{\omega^2}\right) \tag{16.2.4}$$

Then, $\varepsilon < \varepsilon_0$ always meaning that the phase velocity given by (16.2.3) 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 (16.2.2), $v_{ph} = \infty$. These ludicrous observations can be justified or understood only if we can show that information can only be sent by using a wave packet.[1] 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 always less than the velocity of light.

[1] 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.

16.2.2 Group Velocity

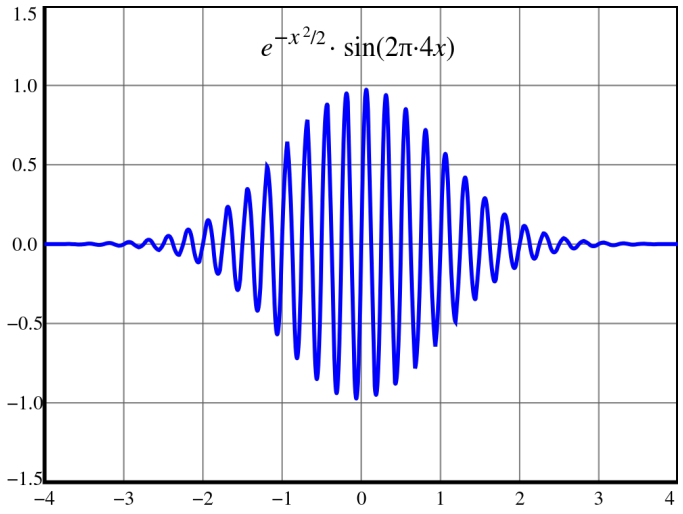


Figure 16.4: A Gaussian wave packet can be thought of as a linear superposition of monochromatic waves of slightly different frequencies. If one Fourier transforms the above signal, it will be a narrow-band signal centered about certain ω_0 (courtesy of Wikimedia [111]).

Now, consider a narrow band wave packet as shown in Figure 16.4. It cannot be monochromatic, 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[2]

$$V(z,t) = \int_{-\infty}^{\infty} d\omega \underset{\sim}{V}(z,\omega) e^{j\omega t} \tag{16.2.5}$$

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[3]

$$\frac{d^2}{dz^2}\underset{\sim}{V}(z,\omega) + k^2(\omega)\underset{\sim}{V}(z,\omega) = 0 \tag{16.2.6}$$

To derive this equation, 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 11.2 to the case of dispersive transmission lines. For instance, when the co-axial transmission

[2]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 $V(z,t)$.

[3]In this notes, we will use k and β interchangeably for wavenumber. The transmission line community tends to use β while the optics community uses k.

line is filled with a dispersive material, then $k^2 = \omega^2 \mu_0 \varepsilon(\omega)$. Thus, upon solving the above equation, one obtains that $V(z,\omega) = V_0(\omega) e^{-jkz}$, and

$$V(z,t) = \int_{-\infty}^{\infty} d\omega V_0(\omega) e^{j(\omega t - kz)} \tag{16.2.7}$$

In the general case, k is a complicated function of ω as shown in Figure 16.5.

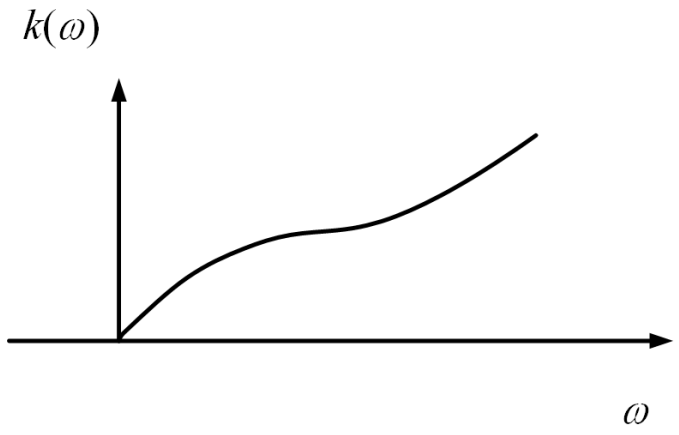


Figure 16.5: A typical frequency dependent $k(\omega)$ albeit the frequency dependence can be more complicated than shown here.

Since this is a wave packet, we assume that $V_0(\omega)$ is narrow band centered about a frequency ω_0, the carrier frequency as shown in Figure 16.6. Therefore, when the integral in (16.2.7) is performed, we need only sum over a narrow range of frequencies in the vicinity of ω_0.

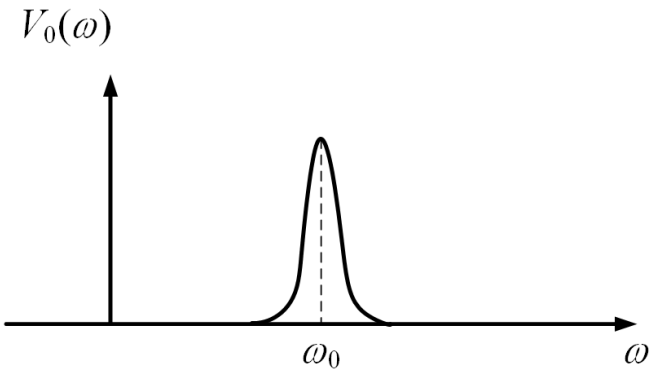


Figure 16.6: The frequency spectrum of $V_0(\omega)$ which is the Fourier transform of $V_0(t)$.

Thus, we can approximate the integrand in the vicinity of $\omega = \omega_0$, in particular, $k(\omega)$ by Taylor series expansion, and let

$$k(\omega) \cong k(\omega_0) + (\omega - \omega_0)\frac{dk(\omega_0)}{d\omega} + \frac{1}{2}(\omega - \omega_0)^2 \frac{d^2k(\omega_0)}{d\omega^2} + \cdots \tag{16.2.8}$$

To ensure the real-valuedness of (16.2.5), one ensures that $-\omega$ part of the integrand is exactly the complex conjugate of the $+\omega$ part, or $\underset{\sim}{V}(z, -\omega) = \underset{\sim}{V}^*(z, \omega)$. Thus the integral can be folded to integrate only over the positive frequency components. Another way is to sum over only the $+\omega$ part of the integral and take twice the real part of the integral. So, for simplicity, we rewrite (16.2.7) as

$$V(z,t) = 2\Re e \int_0^\infty d\omega V_0(\omega) e^{j(\omega t - kz)} \tag{16.2.9}$$

Thus the above follows from the reality condition of $V(z, t)$.

Since we need to integrate over $\omega \approx \omega_0$, we can substitute (16.2.8) into (16.2.9) and rewrite it as

$$V(z,t) \cong 2\Re e \left[e^{j[\omega_0 t - k(\omega_0)z]} \underbrace{\int_0^\infty d\omega V_0(\omega) e^{j(\omega-\omega_0)t} e^{-j(\omega-\omega_0)\frac{dk}{d\omega}z}}_{F\left(t - \frac{dk}{d\omega}z\right)} \right] \tag{16.2.10}$$

where more specifically,

$$F\left(t - \frac{dk}{d\omega}z\right) = \int_0^\infty d\omega V_0(\omega) e^{j(\omega-\omega_0)t} e^{-j(\omega-\omega_0)\frac{dk}{d\omega}z} \tag{16.2.11}$$

It can be seen that the above integral now involves the integral summation over a small range of ω in the vicinity of ω_0. By a change of variable by letting $\Omega = \omega - \omega_0$, it becomes

$$F\left(t - \frac{dk}{d\omega}z\right) = \int_{-\Delta}^{+\Delta} d\Omega V_0(\Omega + \omega_0) e^{j\Omega\left(t - \frac{dk}{d\omega}z\right)} \tag{16.2.12}$$

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, $V_0(\omega)$ is sufficiently small so that its value can be ignored.

The above itself is a Fourier transform integral that involves only the low frequencies of the Fourier spectrum where $e^{j\Omega\left(t - \frac{dk}{d\omega}z\right)}$ is evaluated over small Ω values. Hence, F is a slowly varying function. Moreover, this function F moves with a velocity

$$v_g = \frac{d\omega}{dk} \tag{16.2.13}$$

Here, $F(t - \frac{z}{v_g})$ in fact is the velocity of the envelope in Figure 16.4. In (16.2.10), the envelope function $F(t - \frac{z}{v_g})$ is multiplied by the rapidly varying function

$$e^{j[\omega_0 t - k(\omega_0)z]} \tag{16.2.14}$$

before one takes the real part of the entire function. Hence, this rapidly varying part represents the rapidly varying carrier frequency shown in Figure 16.4. More importantly, this carrier, the rapidly varying part of the signal, moves with the velocity

$$v_{ph} = \frac{\omega_0}{k(\omega_0)} \tag{16.2.15}$$

which is the phase velocity.

16.3 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 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 generalized reflection coefficient (16.1.2) will become infinite.[4]

16.3.1 Transverse Resonance Condition

Therefore, to have a guided mode exist in a layered medium due to multiple bounces, the generalized reflection coefficient becomes infinite, the denominator of (16.1.2) is zero, or that

$$1 + R_{12}\tilde{R}_{23}e^{-2j\beta_{2z}l_2} = 0 \tag{16.3.1}$$

[4]As mentioned previously in Section 15.1.1, 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.

where t is the thickness of the dielectric slab. Since $R_{12} = -R_{21}$, the above can be written as

$$1 = R_{21}\tilde{R}_{23}e^{-2j\beta_{2z}l_2} \tag{16.3.2}$$

The above has the physical meaning that the wave, after going through two reflections at the two interfaces, 21, and 23 interfaces, which are R_{21} and $\tilde{R}_{23}$, plus a phase delay given 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

$$1 = R_{21}R_{23}e^{-2j\beta_{2z}l_2} \tag{16.3.3}$$

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.

Lecture 17

Dielectric Waveguides

As mentioned before, the dielectric waveguide shares many salient features with the optical fiber waveguide, one of the most important waveguides of this century. Before we embark on the study of dielectric waveguides, we will revisit the transverse resonance again. The transverse resonance condition allows one to derive the guidance conditions for a dielectric waveguide easily without having to match the boundary conditions at the interface again: The boundary conditions are already used when deriving the Fresnel reflection coefficients, and hence they are embedded in these reflection coefficients and generalized reflection coefficients. Much of the materials in this lecture can be found in [32, 44, 83].

17.1 Generalized Transverse Resonance Condition

The guidance conditions, the transverse resonance condition given previously, can also be derived for the more general case. The generalized transverse resonance condition is a powerful condition that can be used to derive the guidance condition of a mode 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 17.1)[1] 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 at 5 must be the same as 1.

Assuming that the wave starts with amplitude 1 at position 1, 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 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

[1]The waveguide convention is to assume the direction of propagation to be z. 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.

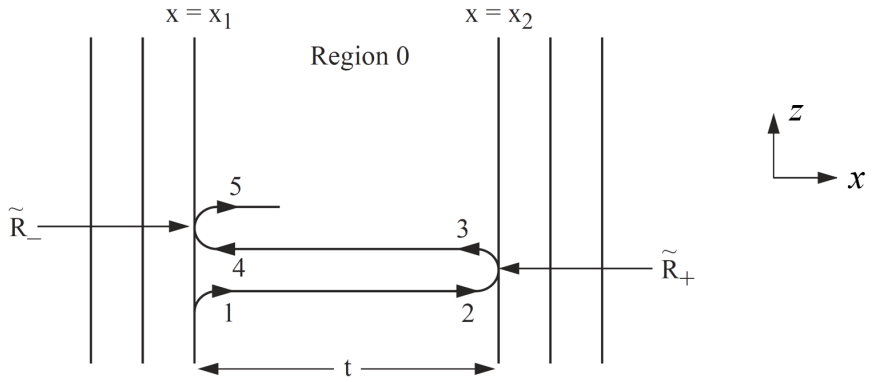


Figure 17.1: The transverse resonance condition for a layered medium. The phase of the wave at position 5 should be equal to the transverse phase at position 1 for constructive interference to occur.

reflection coefficient at the left interface of Region 0. For constructive interference to occur or for the mode to exist, we require that

$$\tilde{R}_{-}\tilde{R}_{+}e^{-2j\beta_{0x}t} = 1 \tag{17.1.1}$$

The above is the generalized transverse resonance condition for the guidance condition for a plane wave mode traveling in a layered medium.

In (17.1.1), a metallic wall has a reflection coefficient of 1 for a TM wave; hence if $\tilde{R}_{+}$ is 1, Equation (17.1.1) becomes

$$1 - \tilde{R}_{-}e^{-2j\beta_{0x}t} = 0. \tag{17.1.2}$$

On the other hand, in (17.1.1), a metallic wall has a reflection coefficient of -1, for TE wave, and Equation (17.1.1) becomes

$$1 + \tilde{R}_{-}e^{-2j\beta_{0x}t} = 0. \tag{17.1.3}$$

17.2 Dielectric Waveguide

The most important dielectric waveguide of the modern world is the optical fiber, whose invention was credited to Charles Kao [98]. He was awarded the Nobel prize in 2009 [112]. However, the 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 dielectric waveguides, one can study the slab dielectric waveguide, which shares many salient features with the optical fiber. This waveguide is also used as thin-film optical waveguides (see Figure 17.2). We start with analyzing the TE modes in this waveguide.

Optical Thin-Film
Waveguide

Figure 17.2: 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.

17.2.1 TE Case

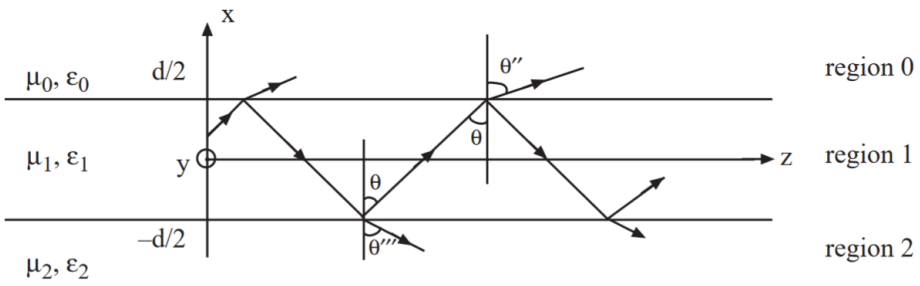


Figure 17.3: Schematic for the analysis of a guided mode in the dielectric waveguide. Total internal reflections occur at the top and bottom interfaces. If the waves add coherently, the wave is guided along the dielectric slab.

We shall look at the application of the transverse resonance condition to a TE wave guided in a dielectric waveguide. Again, we assume the direction of propagation of the guided mode to be in the z direciton in accordance with convention. Specializing the above equation to the dielectric waveguide shown in Figure 17.3, we have the guidance condition as

$$1 = R_{10}R_{12}e^{-2j\beta_{1x}d} \tag{17.2.1}$$

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)[2] than region 0 and 2.

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

$$1 = R_{10}^2 e^{-2j\beta_{1x}d} \tag{17.2.2}$$

[2]Optically more dense means higher optical refractive index, or higher dielectric constant.

By phase-matching, β_z is the same in all the three regions of Figure 17.3. By expressing all the β_{ix} in terms of the variable β_z, the above is an implicit equation 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 $\beta_{0x} = -j\alpha_{0x}$. Therefore, for TE polarization, the local, single-interface, or Fresnel reflection coefficient is

$$R_{10} = \frac{\mu_0\beta_{1x} - \mu_1\beta_{0x}}{\mu_0\beta_{1x} + \mu_1\beta_{0x}} = \frac{\mu_0\beta_{1x} + j\mu_1\alpha_{0x}}{\mu_0\beta_{1x} - j\mu_1\alpha_{0x}} = e^{j\theta_{TE}} \tag{17.2.3}$$

where θ_{TE} is the Goos-Hanschen shift for total internal reflection. It is given by

$$\theta_{TE} = 2\tan^{-1}\left(\frac{\mu_1\alpha_{0x}}{\mu_0\beta_{1x}}\right) \tag{17.2.4}$$

The guidance condition for constructive interference according to (17.2.1) is such that

$$2\theta_{TE} - 2\beta_{1x}d = 2n\pi \tag{17.2.5}$$

From the above, dividing it by four, and taking its tangent, we get

$$\tan\left(\frac{\theta_{TE}}{2}\right) = \tan\left(\frac{n\pi}{2} + \frac{\beta_{1x}d}{2}\right) \tag{17.2.6}$$

or using (17.2.4) for the left-hand side,

$$\frac{\mu_1\alpha_{0x}}{\mu_0\beta_{1x}} = \tan\left(\frac{n\pi}{2} + \frac{\beta_{1x}d}{2}\right) \tag{17.2.7}$$

The above gives rise to

$$\mu_1\alpha_{0x} = \mu_0\beta_{1x}\tan\left(\frac{\beta_{1x}d}{2}\right), \qquad n \text{ even} \tag{17.2.8}$$

$$-\mu_1\alpha_{0x} = \mu_0\beta_{1x}\cot\left(\frac{\beta_{1x}d}{2}\right), \qquad n \text{ odd} \tag{17.2.9}$$

It can be shown that when n is even, the mode profile is even, whereas when n is odd, the mode profile is odd. The above can also be rewritten as

$$\frac{\mu_0}{\mu_1}\frac{\beta_{1x}d}{2}\tan\left(\frac{\beta_{1x}d}{2}\right) = \frac{\alpha_{0x}d}{2}, \qquad \text{even modes} \tag{17.2.10}$$

$$-\frac{\mu_0}{\mu_1}\frac{\beta_{1x}d}{2}\cot\left(\frac{\beta_{1x}d}{2}\right) = \frac{\alpha_{0x}d}{2}, \qquad \text{odd modes} \tag{17.2.11}$$

Again, the above equations can be expressed in the β_z variable, but they do not have closed form solutions, save for graphical solutions (or numerical solutions). We shall discuss their graphical solutions.[3]

[3]This technique has been put together by the community of scholars in the optical waveguide area.

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. 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, one can show that

$$\alpha_{0x} = [\omega^2(\mu_1\epsilon_1 - \mu_0\epsilon_0) - \beta_{1x}^2]^{\frac{1}{2}} \tag{17.2.12}$$

Thus (17.2.10) and (17.2.11) become

$$\begin{aligned} \frac{\mu_0}{\mu_1}\frac{\beta_{1x}d}{2}\tan\left(\frac{\beta_{1x}d}{2}\right) &= \frac{\alpha_{0x}d}{2} \\ &= \sqrt{\omega^2(\mu_1\epsilon_1 - \mu_0\epsilon_0)\frac{d^2}{4} - \left(\frac{\beta_{1x}d}{2}\right)^2}, \quad \text{even modes} \end{aligned} \tag{17.2.13}$$

$$\begin{aligned} -\frac{\mu_0}{\mu_1}\frac{\beta_{1x}d}{2}\cot\left(\frac{\beta_{1x}d}{2}\right) &= \frac{\alpha_{0x}d}{2} \\ &= \sqrt{\omega^2(\mu_1\epsilon_1 - \mu_0\epsilon_0)\frac{d^2}{4} - \left(\frac{\beta_{1x}d}{2}\right)^2}, \quad \text{odd modes} \end{aligned} \tag{17.2.14}$$

We can solve the above graphically by plotting

$$y_1 = \frac{\mu_0}{\mu_1}\frac{\beta_{1x}d}{2}\tan\left(\frac{\beta_{1x}d}{2}\right) \quad \text{even modes} \tag{17.2.15}$$

$$y_2 = -\frac{\mu_0}{\mu_1}\frac{\beta_{1x}d}{2}\cot\left(\beta_{1x}\frac{d}{2}\right) \quad \text{odd modes} \tag{17.2.16}$$

$$y_3 = \left[\omega^2(\mu_1\epsilon_1 - \mu_0\epsilon_0)\frac{d^2}{4} - \left(\frac{\beta_{1x}d}{2}\right)^2\right]^{\frac{1}{2}} = \frac{\alpha_{0x}d}{2} \tag{17.2.17}$$

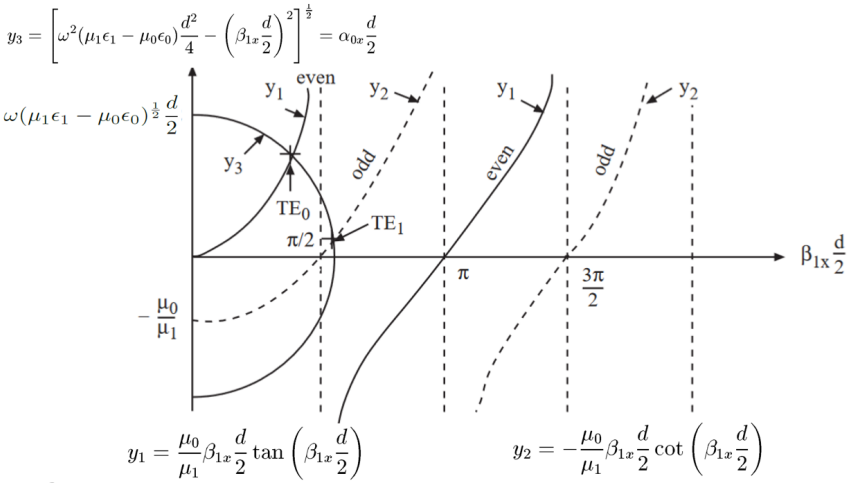


Figure 17.4: A way to solve (17.2.13) and (17.2.14) is via a graphical method. In this method, both the right-hand side and the left-hand side of the equations are plotted on the same plot. The solutions are at the intersection points of these plots.

In the above, y_3 is the equation of a circle; the radius of the circle is given by

$$\omega(\mu_1\epsilon_1 - \mu_0\epsilon_0)^{\frac{1}{2}}\frac{d}{2}. \tag{17.2.18}$$

The solutions to (17.2.13) and (17.2.14) are given by the intersections of y_3 with y_1 and y_2. We note from (17.2.1) that the radius of the circle can be increased in three ways: (i) by increasing the frequency, (ii) by increasing the contrast $\frac{\mu_1\epsilon_1}{\mu_0\epsilon_0}$, and (iii) by increasing the thickness d of the slab.[4] The mode profiles of the first two modes are shown in Figure 17.5.

[4]These area important salient features of a dielectric waveguide. These features are also shared by the optical fiber.

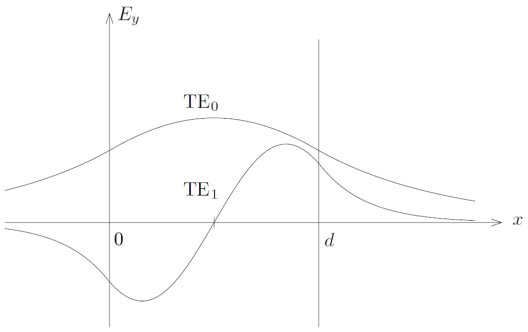


Figure 17.5: Mode profiles of the TE_0 and TE_1 modes of a dielectric slab waveguide (courtesy of J.A. Kong [32]).

When $\beta_{0x} = -j\alpha_{0x}$, the reflection coefficient for total internal reflection is

$$R_{10}^{TE} = \frac{\mu_0\beta_{1x} + j\mu_1\alpha_{0x}}{\mu_0\beta_{1x} - j\mu_1\alpha_{0x}} = \exp\left[+2j\tan^{-1}\left(\frac{\mu_1\alpha_{0x}}{\mu_0\beta_{1x}}\right)\right] \tag{17.2.19}$$

and $\left|R_{10}^{TE}\right| = 1$. Hence, the wave is guided by total internal reflections.

Cut-off occurs when the total internal reflection ceases to occur, i.e. when the frequency decreases such that $\alpha_{0x} = 0$. From Figure 17.4, we see that $\alpha_{0x} = 0$ when

$$\omega(\mu_1\epsilon_1 - \mu_0\epsilon_0)^{\frac{1}{2}}\frac{d}{2} = \frac{m\pi}{2}, \qquad m = 0, 1, 2, 3, \ldots \tag{17.2.20}$$

or

$$\omega_{mc} = \frac{m\pi}{d(\mu_1\epsilon_1 - \mu_0\epsilon_0)^{\frac{1}{2}}}, \qquad m = 0, 1, 2, 3, \ldots \tag{17.2.21}$$

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 17.6 where the TE mode profiles are similar since they are dual to each other. The boundary conditions at the dielectric interface is that the field and its normal derivative have to be continuous. The TE_0 or TM_0 mode can satisfy this boundary condition at all frequencies, but not the TE_1 or TM_1 mode. At the cut-off frequency, the field outside the slab has to become flat implying the $\alpha_{0x} = 0$ implying no guidance.

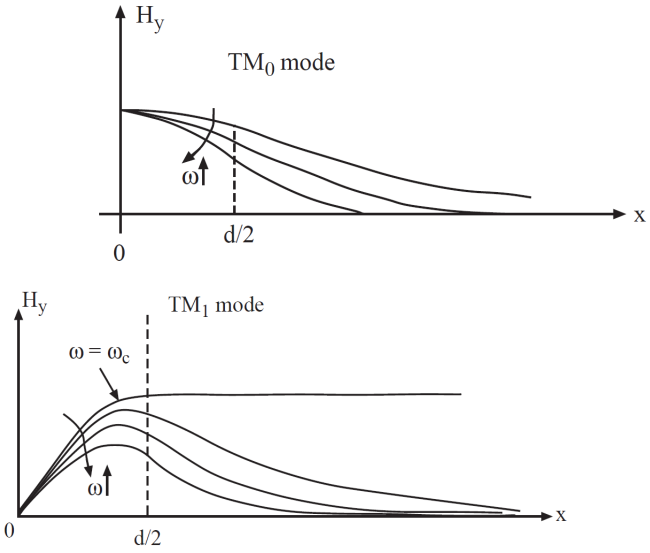


Figure 17.6: Mode profiles of the TM modes of a dielectric slab. The TE modes are dual to the TM modes and have similar mode profiles.

Next, we will elucidate more physics of the dielectric slab guided mode. At cut-off, $\alpha_{0x} = 0$, and from the dispersion relation that $\alpha_{0x}^2 = \beta_z^2 - \beta_0^2$,

$$\beta_z = \omega\sqrt{\mu_0\epsilon_0},$$

for all the modes. Hence, the phase velocity, ω/β_z, and the group velocity, $d\omega/d\beta_z$ are that 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 $\omega \to \infty$, the radius of the circle in the plot of y_3 becomes increasingly larger. As seen from Figure 17.4, the solution for $\beta_{1x} \to \frac{n\pi}{d}$ for all the modes. From the dispersion relation for Region 1,

$$\beta_z = \sqrt{\omega^2\mu_1\epsilon_1 - \beta_{1x}^2} \approx \omega\sqrt{\mu_1\epsilon_1}, \qquad \omega \to \infty \tag{17.2.22}$$

since $\omega^2\mu_1\epsilon_1 \gg \beta_{1x}^2 = (n\pi/d)^2$. 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 17.7. In this figure,[5] k_{c1}, k_{c2}, and k_{c3} are the cut-off wave number or frequency of the first three modes. Close to cut-off, the field is traveling mostly outside the waveguide, and $k_z \approx \omega\sqrt{\mu_0\varepsilon_0}$. Hence, both the phase and group velocities approach that of

[5]Please note again that in this course, we will use β and k interchangeably for wavenumbers.

the outer medium as shown in the figure. When the frequency increases, the mode is tightly confined in the dielectric slab, and hence, $k_z \approx \omega\sqrt{\mu_1\varepsilon_1}$. Both the phase and group velocities approach that of Region 1 as shown.

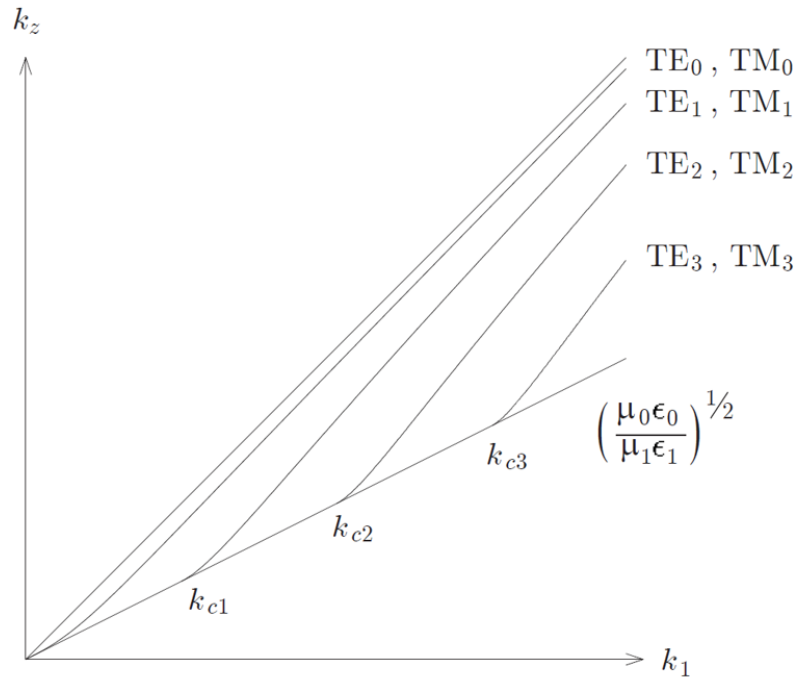


Figure 17.7: Here, we have k_z versus k_1 plots for dielectric slab waveguide. Near its cut-off, the energy of the mode is in the outer region, and hence, its group velocity is close to that of the outer region. At high frequencies, the mode is tightly bound to the slab, and its group velocity approaches that of the dielectric slab (courtesy of J.A. Kong [32]).

17.2.2 TM Case

For the TM case, a similar guidance condition analogous to (17.2.1) can be derived but with the understanding that the reflection coefficients in (17.2.1) are now TM reflection coefficients. Similar derivations show that the above guidance conditions, for $\epsilon_2 = \epsilon_0$, $\mu_2 = \mu_0$, reduce to

$$\frac{\epsilon_0}{\epsilon_1}\beta_{1x}\frac{d}{2}\tan\beta_{1x}\frac{d}{2} = \sqrt{\omega^2(\mu_1\epsilon_1 - \mu_0\epsilon_0)\frac{d^2}{4} - \left(\beta_{1x}\frac{d}{2}\right)^2}, \qquad \text{even modes} \tag{17.2.23}$$

$$-\frac{\epsilon_0}{\epsilon_1}\beta_{1x}\frac{d}{2}\cot\beta_{1x}\frac{d}{2} = \sqrt{\omega^2(\mu_1\epsilon_1 - \mu_0\epsilon_0)\frac{d^2}{4} - \left(\beta_{1x}\frac{d}{2}\right)^2}, \qquad \text{odd modes} \tag{17.2.24}$$

Note that for equation (17.2.1), when we have two parallel metallic plates, $R^{TM} = 1$, and $R^{TE} = -1$, and the guidance condition becomes

$$1 = e^{-2j\beta_{1x}d} \quad \Rightarrow \beta_{1x} = \frac{m\pi}{d}, \qquad m = 0, 1, 2, \ldots, \tag{17.2.25}$$

These are just the guidance conditions for parallel plate waveguides.

17.2.3 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 inside, in this case, 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 one of the Region i's.

Lecture 18

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. 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. A low loss waveguide is also a low noise waveguide.[1]

18.1 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. But 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 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, for instance. However, when the waveguide is hollow or is filled completely with a homogeneous medium, without a center conductor, it cannot support a TEM mode as we shall prove next.

[1] The fluctuation dissipation theorem [113, 114] 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 receives energy back from the environment in terms of thermal noise.

Much of the materials of this lecture can be found in [32, 83, 92].

18.1.1 Absence of TEM Mode in a Hollow Waveguide

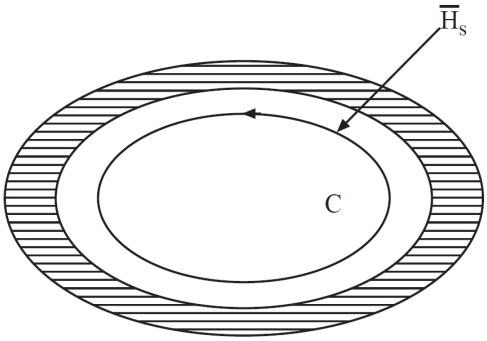


Figure 18.1: Absence of TEM mode in a hollow waveguide enclosed by a PEC wall. The magnetic field lines form a closed loop due to the absence of magnetic charges.

We would like to prove by contradiction (**reductio ad absurdum**) that a hollow waveguide as shown in Figure 18.1 (i.e. without a center conductor) cannot support a TEM mode 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. 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

$$\oint_C \mathbf{H}_s \cdot d\mathbf{l} = j\omega\varepsilon \int_S \mathbf{E} \cdot d\mathbf{S} + \int_S \mathbf{J} \cdot d\mathbf{S} \tag{18.1.1}$$

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

$$\oint_C \mathbf{H}_s \cdot d\mathbf{l} = j\omega\varepsilon \int_S \mathbf{E} \cdot d\mathbf{S} \tag{18.1.2}$$

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 without a center conductor cannot exist.

By the above argument, in a hollow waveguide filled with homogeneous medium, only TE_z (TE to z) or TM_z (TM to z) modes can exist like the case of a layered medium. 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, 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.

We begin by studying some simple closed form solutions to hollow waveguides, such as the rectangular waveguides. These closed form solutions offer physical insight into the propagation of waves in a hollow 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; the preferred method of solving these complicated problems is via numerical method.

18.1.2 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

$$\mathbf{E}(\mathbf{r}) = \nabla \times \hat{z}\Psi_h(\mathbf{r}) \tag{18.1.3}$$

By construction, equation (18.1.3) will guarantee that $E_z = 0$. Here, $\Psi_h(\mathbf{r})$ is a scalar potential and $\hat{z}$ is called the pilot vector. [2]

The waveguide is assumed source free and filled with a lossless, homogeneous material. Eq. (18.1.3) also satisfies the source-free condition since, clearly, $\nabla \cdot \mathbf{E} = 0$. And hence, from Maxwell's equations that

$$\nabla \times \mathbf{E} = j\omega\mu\mathbf{H} \tag{18.1.4}$$

$$\nabla \times \mathbf{H} = -j\omega\varepsilon\mathbf{E} \tag{18.1.5}$$

it can be shown that

$$\nabla \times \nabla \times \mathbf{E} - \omega^2\mu\varepsilon\mathbf{E} = 0 \tag{18.1.6}$$

Furthermore, using the appropriate vector identiy, such as the back-of-the-cab formula, it can be shown that the electric field $\mathbf{E}(\mathbf{r})$ satisfies the following Helmholtz wave equation (or partial differential equation) that

$$(\nabla^2 + \beta^2)\mathbf{E}(\mathbf{r}) = 0 \tag{18.1.7}$$

where $\beta^2 = \omega^2\mu\varepsilon$. Substituting (18.1.3) into (18.1.7), we get

$$(\nabla^2 + \beta^2)\nabla \times \hat{z}\Psi_h(\mathbf{r}) = 0 \tag{18.1.8}$$

In the above, we assume that $\nabla^2\nabla \times \hat{z}\Psi = \nabla \times \hat{z}(\nabla^2\Psi)$, or that these operators commute.[3] Then it follows that

$$\nabla \times \hat{z}(\nabla^2 + \beta^2)\Psi_h(\mathbf{r}) = 0 \tag{18.1.9}$$

[2] It "pilots" the field so that it is transverse to z.

[3] 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 $[A, B] = 0$.

Thus, if $\Psi_h(\mathbf{r})$ satisfies the following Helmholtz wave equation or partial differential equation

$$(\nabla^2 + \beta^2)\Psi_h(\mathbf{r}) = 0 \tag{18.1.10}$$

then (18.1.9) is satisfied, and so is (18.1.7). Hence, the $\mathbf{E}$ field constructed with (18.1.3) satisfies Maxwell's equations, if $\Psi_h(\mathbf{r})$ satisfies (18.1.10).

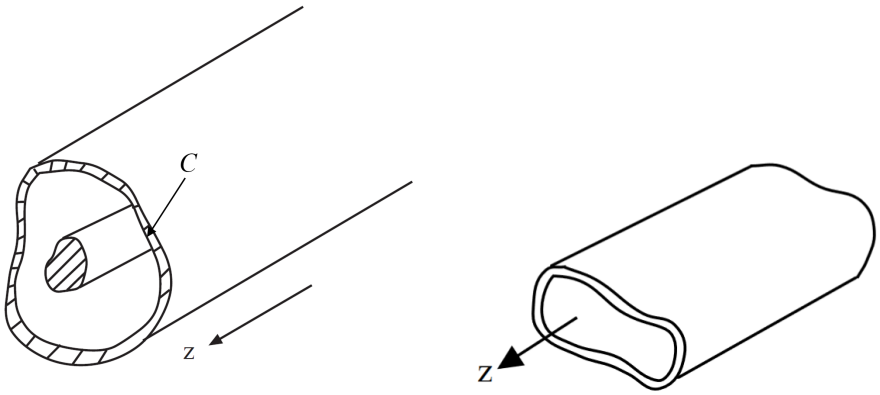


Figure 18.2: A hollow metallic waveguide with a center conductor (left), and without a center conductor (right).

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 (18.1.3), using the back-of-the-cab (BOTC) formula,

$$\hat{n} \times \mathbf{E} = \hat{n} \times (\nabla \times \hat{z}\Psi_h) = -\hat{n} \cdot \nabla\Psi_h = 0 \tag{18.1.11}$$

In applying the BOTC 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.

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 (18.1.11) then becomes

$$\hat{n} \cdot \nabla_s \Psi_h = \partial_n \Psi_h = 0 \text{ on } C \tag{18.1.12}$$

where C is the waveguide wall where ∂_n is a shorthand notation for $\hat{n} \cdot \nabla_s$ operator which is a scalar operator. 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 (18.1.10), or that

$$\Psi_h(\mathbf{r}) = \Psi_{hs}(\mathbf{r}_s)e^{\mp j\beta_z z} \tag{18.1.13}$$

where $\mathbf{r}_s = \hat{x}x + \hat{y}y$, or in short, $\Psi_{hs}(\mathbf{r}_s) = \Psi_{hs}(x, y)$ is a 2D function. Thus, $\partial_n \Psi_h = 0$ implies that $\partial_n \Psi_{hs} = 0$. With this assumption, $\frac{\partial^2}{\partial z^2} \to -{\beta_z}^2$, and (18.1.10) becomes even simpler, namely that,

$$({\nabla_s}^2 + \beta^2 - {\beta_z}^2)\Psi_{hs}(\mathbf{r}_s) = ({\nabla_s}^2 + \beta_s^2)\Psi_{hs}(\mathbf{r}_s) = 0 \ , \quad \partial_n \Psi_{hs}(\mathbf{r}_s) = 0, \text{ on } C \tag{18.1.14}$$

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 is also called the reduced wave equation.

18.1.3 TM Case ($E_z \neq 0$, $H_z = 0$, TM$_z$ Case)

Repeating similar treatment for TM waves, the TM magnetic field is then

$$\mathbf{H} = \nabla \times \hat{z}\Psi_e(\mathbf{r}) \tag{18.1.15}$$

where

$$(\nabla^2 + \beta^2)\Psi_e(\mathbf{r}) = 0 \tag{18.1.16}$$

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 waveguide wall. To this end, we find the corresponding $\mathbf{E}$ field by taking the curl of the magnetic field in (18.1.15), and thus the $\mathbf{E}$ field is proportional to

$$\mathbf{E} \sim \nabla \times \nabla \times \hat{z}\Psi_e(\mathbf{r}) = \nabla\nabla \cdot (\hat{z}\Psi_e) - \nabla^2 \hat{z}\Psi_e = \nabla \frac{\partial}{\partial z}\Psi_e + \hat{z}\beta^2 \Psi_e \tag{18.1.17}$$

where we have used the BOTC 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 the above will be zero if $\Psi_e = 0$ on the waveguide wall. Therefore, if

$$\Psi_e(\mathbf{r}) = 0 \text{ on } C, \tag{18.1.18}$$

where C is the waveguide wall, then,

$$\hat{n} \times \mathbf{E}(\mathbf{r}) = 0 \text{ on } C \tag{18.1.19}$$

Equation (18.1.18) is also called the homogeneous Dirichlet boundary condition.

Next, we assume that

$$\Psi_e(\mathbf{r}) = \Psi_{es}(\mathbf{r}_s)e^{\mp j\beta_z z} \tag{18.1.20}$$

This will allow us to replace $\partial^2/\partial z^2 = -\beta_z^2$. Thus, with some manipulation, the boundary value problem (BVP) related to equation (18.1.16) reduces to a simpler 2D problem, i.e.,

$$({\nabla_s}^2 + \beta_s^2)\Psi_{es}(\mathbf{r}_s) = 0 \tag{18.1.21}$$

with the homogeneous Dirichlet boundary condition that

$$\Psi_{es}(\mathbf{r}_s) = 0, \, \mathbf{r}_s \text{ on } C \tag{18.1.22}$$

To illustrate the above theory, we can solve some simple waveguides problems.

18.2 Rectangular Waveguides

Rectangular waveguides are among the simplest waveguides to analyze because closed form solutions exist in cartesian coordinates. One can imagine traveling waves in the xy directions bouncing off the walls of the waveguide causing standing waves to exist inside the waveguide.

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 waveguide. When a waveguide mode cannot propagate in a waveguide, that mode is known to be cut-off. The concept of cut-off for hollow waveguide is quite different from that of a dielectric waveguide we have studied previously.

18.2.1 TE Modes ($H_z \neq 0$, H Modes or TE$_z$ Modes)

For this mode, the scalar potential $\Psi_{hs}(\mathbf{r}_s)$ satisfies

$$(\nabla_s{}^2 + \beta_s{}^2)\Psi_{hs}(\mathbf{r}_s) = 0, \quad \frac{\partial}{\partial n}\Psi_{hs}(\mathbf{r}_s) = 0 \quad \text{on } C \tag{18.2.1}$$

where $\beta_s{}^2 = \beta^2 - \beta_z{}^2$. A viable solution using separation of variables[4] for $\Psi_{hs}(x, y)$ is then

$$\Psi_{hs}(x, y) = A\cos(\beta_x x)\cos(\beta_y y) \tag{18.2.2}$$

where $\beta_x{}^2 + \beta_y^2 = \beta_s^2$. One can see that the above is the representation of standing waves in the xy directions. It is quite clear that $\Psi_{hs}(x, y)$ satisfies the BVP and boundary conditions defined by equation (18.2.1). 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.

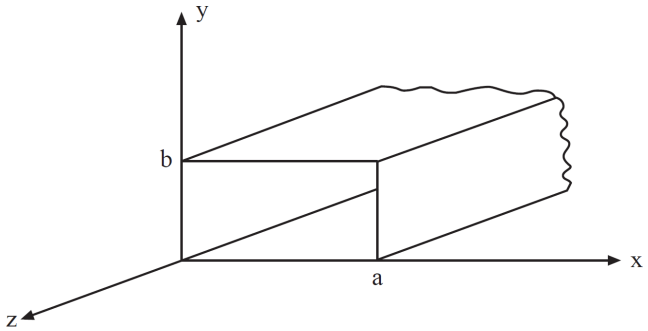


Figure 18.3: The schematic of a rectangular waveguide. By convention, the length of the longer side is usually named a.

[4]For those who are not familiar with this topic, please consult p. 385 of Kong [32].

To further satisfy the boundary condition at $x = a$, and $y = b$ surfaces, it is necessary that the boundary condition for eq. (18.1.12) is satisfied or that

$$\partial_x \Psi_{hs}(x, y)|_{x=a} \sim \sin(\beta_x a) \cos(\beta_y y) = 0, \tag{18.2.3}$$

$$\partial_y \Psi_{hs}(x, y)|_{y=b} \sim \cos(\beta_x x) \sin(\beta_y b) = 0, \tag{18.2.4}$$

The above puts constraints on β_x and β_y, implying that $\beta_x a = m\pi$, $\beta_y b = n\pi$ where m and n are integers. Hence, (18.2.2) becomes

$$\Psi_{hs}(x, y) = A \cos\left(\frac{m\pi}{a} x\right) \cos\left(\frac{n\pi}{b} y\right) \tag{18.2.5}$$

where

$$\beta_x^2 + \beta_y^2 = \left(\frac{m\pi}{a}\right)^2 + \left(\frac{n\pi}{b}\right)^2 = \beta_s^2 = \beta^2 - {\beta_z}^2 \tag{18.2.6}$$

Clearly, (18.2.5) satisfies the requisite homogeneous Neumann boundary condition at the entire waveguide wall.

At this point, it is prudent to stop and ponder on what we have done. Equation (18.2.1) is homomorphic to a matrix eigenvalue problem

$$\overline{\mathbf{A}} \cdot \mathbf{x}_i = \lambda_i \mathbf{x}_i \tag{18.2.7}$$

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 (18.2.5). The corresponding eigenmode is also called the TE_{mn} mode.

The above condition on β_s^2 expressed by (18.2.6) is also known as the guidance condition for the modes in the waveguide. Furthermore, from (18.2.6),

$$\beta_z = \sqrt{\beta^2 - \beta_s^2} = \sqrt{\beta^2 - \left(\frac{m\pi}{a}\right)^2 - \left(\frac{n\pi}{b}\right)^2} \tag{18.2.8}$$

And from (18.2.8), when the frequency is low enough, then

$$\beta_s^2 = \left(\frac{m\pi}{a}\right)^2 + \left(\frac{n\pi}{b}\right)^2 > \beta^2 = \omega^2 \mu\varepsilon \tag{18.2.9}$$

and β_z becomes pure imaginary and the mode cannot propagate or becomes evanescent in the z direction.[5] For fixed m and n, the frequency at which the above happens is called the cutoff frequency of the TE_{mn} mode of the waveguide. It is given by

$$\omega_{mn,c} = \frac{1}{\sqrt{\mu\varepsilon}} \sqrt{\left(\frac{m\pi}{a}\right)^2 + \left(\frac{n\pi}{b}\right)^2} \tag{18.2.10}$$

[5] We have seen this happening in a plasma medium earlier and also in total internal reflection.

When $\omega < \omega_{mn,c}$, or the wavelength is longer than a certain value, the TE_{mn} mode is evanescent and cannot propagate inside the waveguide. A corresponding cutoff wavelength is then

$$\lambda_{mn,c} = \frac{2}{[\left(\frac{m}{a}\right)^2 + \left(\frac{n}{b}\right)^2]^{1/2}} \tag{18.2.11}$$

So when $\lambda > \lambda_{mn,c}$, the mode cannot propagate inside the waveguide.

Lowest Guided Mode in a Rectangular Waveguide

When $m = n = 0$, then $\Psi_h(\mathbf{r}) = \Psi_{hs}(x, y) \exp(\mp j\beta_z z)$ 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 cutoff frequency or longest cutoff wavelength. Only when the frequency is above this cutoff frequency and the wavelength is shorter than this cutoff wavelength, the TE_{10} mode can propagate.

For the TE_{10} mode, for the mode to propagate, from (18.2.11), it is needed that

$$\lambda < \lambda_{10,c} = 2a \tag{18.2.12}$$

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 cutoff frequency or longest cutoff 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, a single guided mode only is desirable. Hence, the knowledge of the cutoff frequencies of the fundamental mode (the mode with the lowest cutoff 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 cutoff, the field is evanescent, and there is no real power flow down the waveguide: Only reactive power is carried by such a mode.

Lecture 19

More on Hollow Waveguides

We have seen that the hollow waveguide is the simplest of waveguides. 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 exist also. 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 eye of the beholder. Once one is familiarized 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 19.1)!

Figure 19.1: The ripple wave in your tea cup is describable by a Bessel function (courtesy of dreamstime.com).

19.1 Rectangular Waveguides, Contd.

We have seen the mathematics for the TE modes of a rectangular waveguide. We shall study the TM modes and the modes of a circular waveguide in this lecture.

19.1.1 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

$$\Psi_{es}(x, y) = A \sin\left(\frac{m\pi}{a}x\right) \sin\left(\frac{n\pi}{b}y\right) \tag{19.1.1}$$

Here, sine functions are chosen for the standing waves, and the chosen values of β_x and β_y ensure that the 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 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, then the eigenvalue is

$$\beta_s^2 = \beta_x^2 + \beta_y^2 = \left(\frac{m\pi}{a}\right)^2 + \left(\frac{n\pi}{b}\right)^2 \tag{19.1.2}$$

which is the same as the TE case. Therefore, the corresponding cutoff frequencies and cutoff 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. Furthermore, the lowest modes, TE$_{11}$ and TM$_{11}$ modes have the same cutoff frequency. Figure 19.2 shows the dispersion curves for different modes of a rectangular waveguide. Notice that the group velocities of all the modes are zero at cutoff, and then the group velocities approach that of the waveguide medium as frequency becomes large. These observations can be explained physically.

[1] Again, "homogeneous" here means "zero".

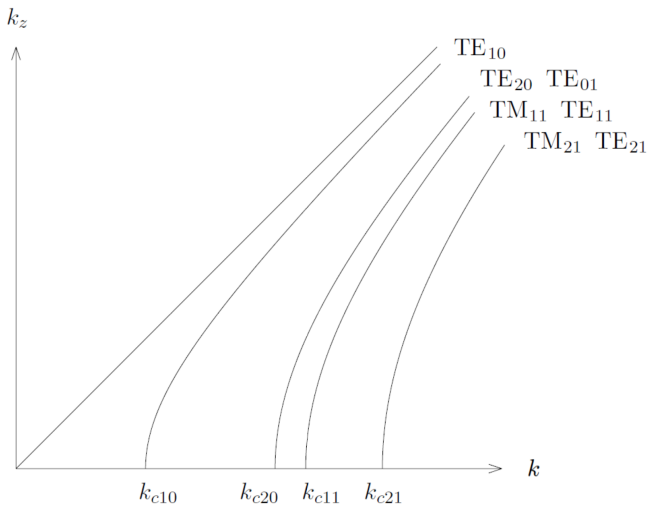


Figure 19.2: Dispersion curves for a rectangular waveguide (courtesy of J.A. Kong [32]). Notice that the lowest TM mode is the TM_{11} mode, and k is equivalent to β in this course. At cutoff, the guided mode does not propagate in the z direction, and here, the group velocity is zero. But when $\omega \to \infty$, the mode propagates in direction almost parallel to the axis of the waveguide, and hence, the group velocity approaches that of the waveguide medium.

19.1.2 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 that they are

$$[\exp(-j\beta_x x) \pm \exp(j\beta_x x)]\,[\exp(-j\beta_y y) \pm \exp(j\beta_y y)]$$

When the above is expanded and together with the $\exp(-j\beta_z z)$ the mode propagating in the z direction in addition to the standing waves in the tranverse direction. Or we see four waves bouncing around in the xy directions and propagating in the z direction. The picture of this bouncing wave can be depicted in Figure 19.3.

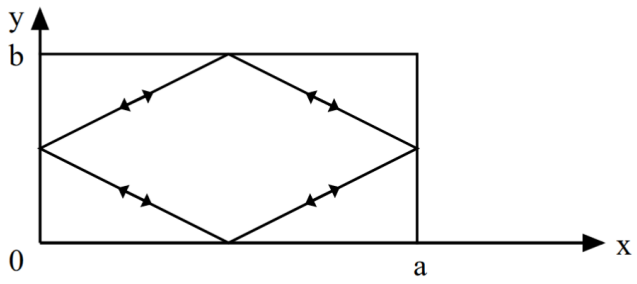


Figure 19.3: The waves in a rectangular waveguide can be thought of as bouncing waves off the four walls as they propagate in the z direction.

19.1.3 Field Plots

Given the knowledge of the scalar piloting 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

$$\mathbf{E} = \nabla \times \hat{z}\Psi_h(\mathbf{r}), \qquad \mathbf{H} = -\nabla \times \mathbf{E}/(j\omega\mu) \tag{19.1.3}$$

Then all the electromagnetic field of a waveguide mode can be found, and similarly for TM modes.

Plots of the fields of different rectangular waveguide modes are shown in Figure 19.4. 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 $\beta_z = \sqrt{\beta^2 - \beta_x^2 - \beta_y^2}$, higher frequencies are needed to make β_z real in order to propagate the higher order modes or the high m and n modes.

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" but with a $\pi/2$ phase delay due to the $j\omega$ factor. Therefore, the $\mathbf{E}$ and $\mathbf{H}$ fields do not curl around each other at one location, but at a displaced location due to the $\pi/2$ phase difference. This is shown in Figure 19.5.

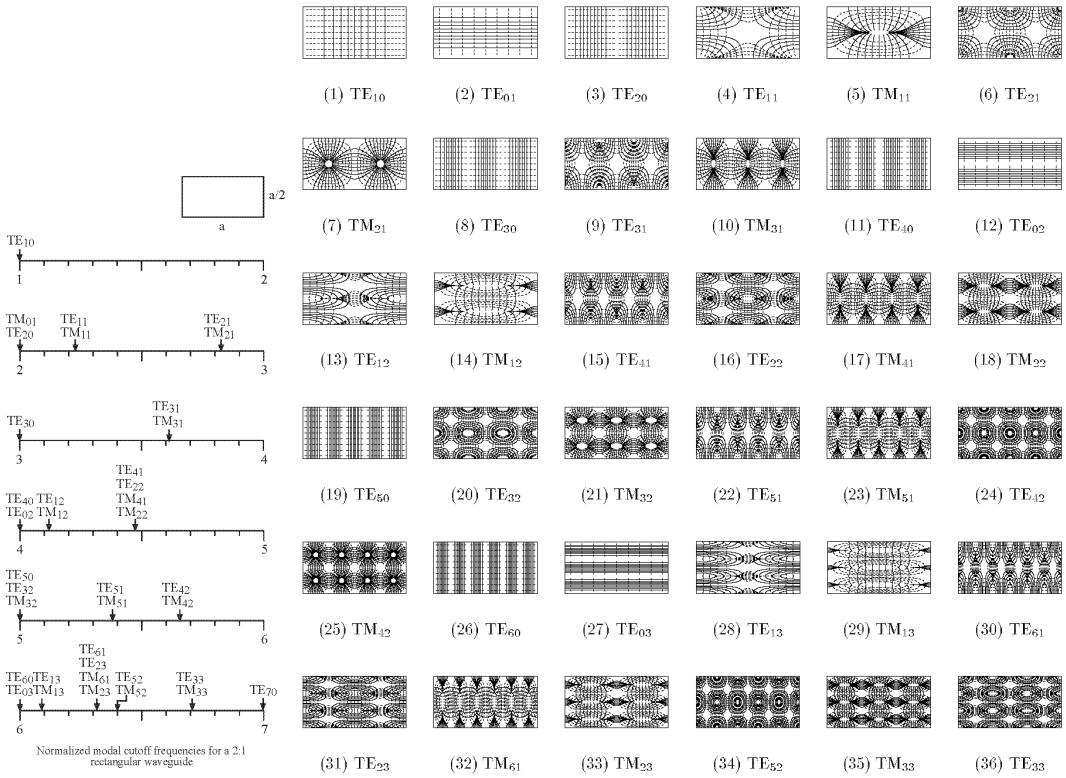


Figure 19.4: Transverse field plots of different modes in a rectangular waveguide (courtesy of Andy Greenwood. Original plots published in Lee, Lee, and Chuang, IEEE T-MTT, 33.3 (1985): pp. 271-274. [115]).

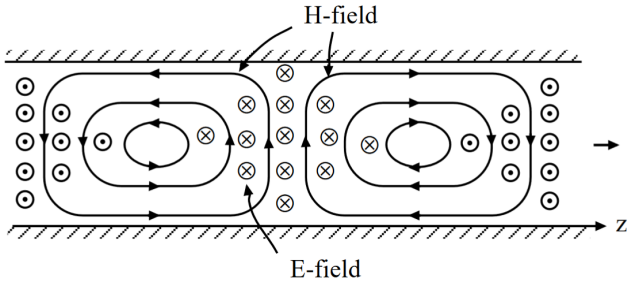


Figure 19.5: Field plot of a mode propagating in the z direction of a rectangular waveguide. Notice that the E and H fields do not exactly curl around each other.

19.2 Circular Waveguides

Another waveguide where closed-form solutions can be easily obtained is the circular hollow waveguide as shown in Figure 19.6, but they involve the use of Bessel functions.

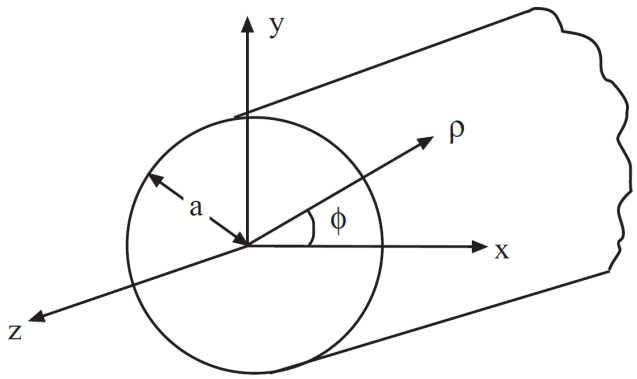


Figure 19.6: Schematic of a circular waveguide in cylindrical coordinates. It is one of the separable coordinate systems.

19.2.1 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 first ∇_s is a gradient operator while the second $\nabla_s \cdot$ is a divergence operator: they have different physical meanings. Formulas for grad and div operators are given in many text books [32, 116]. 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

$$\left({\nabla_s}^2 + {\beta_s}^2\right) \Psi_{hs} = \left(\frac{1}{\rho} \frac{\partial}{\partial \rho} \rho \frac{\partial}{\partial \rho} + \frac{1}{\rho^2} \frac{\partial^2}{\partial \phi^2} + {\beta_s}^2 \right) \Psi_{hs}(\rho, \phi) = 0 \tag{19.2.1}$$

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). Using separation of variables, we let

$$\Psi_{hs}(\rho, \phi) = B_n(\beta_s \rho) e^{\pm jn\phi} \tag{19.2.2}$$

Then $\frac{\partial^2}{\partial\phi^2} \to -n^2$, and (19.2.1) simplifies to an ordinary differential equation which is

$$\left(\frac{1}{\rho}\frac{d}{d\rho}\rho\frac{d}{d\rho} - \frac{n^2}{\rho^2} + {\beta_s}^2\right) B_n(\beta_s\rho) = 0 \tag{19.2.3}$$

Here, dividing the above equation by β_s^2, we can let $\beta_s\rho$ in (19.2.2) and (19.2.3) be x. Then the above can be rewritten as

$$\left(\frac{1}{x}\frac{d}{dx}x\frac{d}{dx} - \frac{n^2}{x^2} + 1\right) B_n(x) = 0 \tag{19.2.4}$$

The above is known as the Bessel equation whose solutions are special functions denoted as $B_n(x)$.[2]

These special 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 the order, and x is the argument.[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]

Bessel, $$J_n(x) = \frac{1}{2}[{H_n}^{(1)}(x) + {H_n}^{(2)}(x)] \tag{19.2.5}$$

Neumann, $$N_n(x) = \frac{1}{2j}[{H_n}^{(1)}(x) - {H_n}^{(2)}(x)] \tag{19.2.6}$$

Hankel–First kind, $${H_n}^{(1)}(x) = J_n(x) + jN_n(x) \tag{19.2.7}$$

Hankel–Second kind, $${H_n}^{(2)}(x) = J_n(x) - jN_n(x) \tag{19.2.8}$$

It can be shown that

$${H_n}^{(1)}(x) \sim \sqrt{\frac{2}{\pi x}}e^{jx - j(n+\frac{1}{2})\frac{\pi}{2}}, \quad x \to \infty \tag{19.2.9}$$

$${H_n}^{(2)}(x) \sim \sqrt{\frac{2}{\pi x}}e^{-jx + j(n+\frac{1}{2})\frac{\pi}{2}}, \quad x \to \infty \tag{19.2.10}$$

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 ${H_n}^{(2)}(x) \to \infty$ when $x \to 0$. Since the field has to be regular when $\rho \to 0$ at the center of the waveguide shown in Figure 19.6, the only viable solution for the hollow waveguide, to be chosen from (19.2.5) to (19.2.9), is that $B_n(\beta_s\rho) = AJ_n(\beta_s\rho)$. Thus for a circular hollow waveguide, the eigenfunction or mode is of the form

$$\Psi_{hs}(\rho,\phi) = AJ_n(\beta_s\rho)e^{\pm jn\phi} \tag{19.2.11}$$

[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 $\exp(-jx)$, $\sin(x)$, and $\cos(x)$.

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

$$\frac{d}{d\rho} J_n(\beta_s \rho) = 0, \quad \rho = a \tag{19.2.12}$$

Defining ${J_n}'(x) = \frac{d}{dx} J_n(x)$,[6] the above is the same as

$${J_n}'(\beta_s a) = 0 \tag{19.2.13}$$

The above are the zeros of the derivative of Bessel function and they are tabulated in many textbooks and handbooks. The m-th zero of ${J_n}'(x)$ is denoted to be β_{nm} in many books.[7] Plots of Bessel functions and their derivatives are shown in Figure 19.8, and some zeros of Bessel function and its derivative are also shown in Figure 19.9. With this knowledge, the guidance condition for a waveguide mode is then

$$\beta_s = \beta_{nm}/a \tag{19.2.14}$$

for the TE_{nm} mode. From the above, β_s^2 can be obtained which is the eigenvalue of (19.2.1) and (19.2.3). 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 β^2 is small enough (or the frequency low enough) so that $\beta^2 < \beta_s^2$ or $\beta < \beta_s$. From this, the corresponding cutoff frequency (the frequency below which β_z becomes pure imaginary) of the TE_{nm} mode is

$$\omega_{nm,c} = \frac{1}{\sqrt{\mu\varepsilon}} \frac{\beta_{nm}}{a} \tag{19.2.15}$$

When $\omega < \omega_{nm,c}$, the corresponding mode cannot propagate in the waveguide as β_z becomes pure imaginary. The corresponding cutoff wavelength is

$$\lambda_{nm,c} = \frac{2\pi}{\beta_{nm}} a \tag{19.2.16}$$

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.

19.2.2 TM Case

The corresponding partial differential equation and boundary value problem for this case is

$$\left(\frac{1}{\rho} \frac{\partial}{\partial \rho} \rho \frac{\partial}{\partial \rho} + \frac{1}{\rho^2} \frac{\partial^2}{\partial \phi^2} + {\beta_s}^2 \right) \Psi_{es}(\rho, \phi) = 0 \tag{19.2.17}$$

[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 [117]. An online version is available at [118].

with the homogeneous Dirichlet boundary condition, $\Psi_{es}(a,\phi) = 0$, on the waveguide wall. The eigenfunction solution is

$$\Psi_{es}(\rho,\phi) = AJ_n(\beta_s\rho)e^{\pm jn\phi} \tag{19.2.18}$$

with the boundary condition that $J_n(\beta_s a) = 0$. The zeros of $J_n(x)$ are labeled as α_{nm} is many textbooks, as well as in Figure 19.9; and hence, the guidance condition is that for the TM_{nm} mode is that

$$\beta_s = \frac{\alpha_{nm}}{a} \tag{19.2.19}$$

where the eigenvalue for (19.2.17) is β_s^2 which is a constant independent of frequency. With $\beta_z = \sqrt{\beta^2 - \beta_s^2}$, the corresponding cutoff frequency is

$$\omega_{nm,c} = \frac{1}{\sqrt{\mu\varepsilon}}\frac{\alpha_{nm}}{a} \tag{19.2.20}$$

or when $\omega < \omega_{nm,c}$, the mode cannot be guided. The cutoff wavelength is

$$\lambda_{nm,c} = \frac{2\pi}{\alpha_{nm}}a \tag{19.2.21}$$

with the notion that when $\lambda > \lambda_{nm,c}$, the mode cannot be guided.

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.

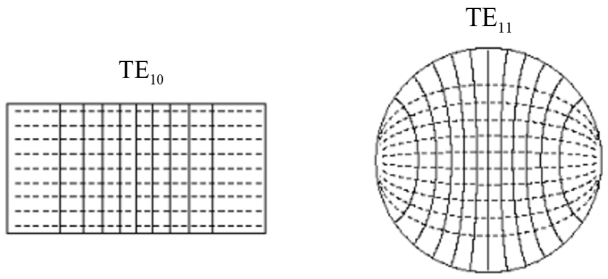


Figure 19.7: Side-by-side comparison of the field plots of the TE_{10} mode of a rectangular waveguide versus that of the TE_{11} mode of a circular waveguide. If one is imaginative enough, one can see that the field plot of one mode morphs into that of the other mode. Electric fields are those that have to end on the waveguide walls.

Figure 19.8 shows the plots of Bessel function $J_n(x)$ and its derivative $J_n'(x)$. Tables in Figure 19.9 show the roots of $J_n'(x)$ and $J_n(x)$ which are important for determining the cutoff frequencies of the TE and TM modes of circular waveguides.

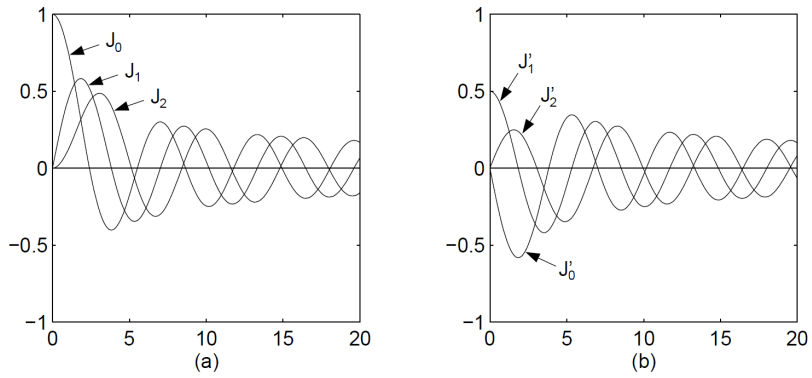


Figure 19.8: Plots of the Bessel function, $J_n(x)$, and its derivatives $J'_n(x)$. The zeros of these functions are used to find the eigenvalue β_s^2 of the problem, and hence, the guidance condition. The left figure is for TM modes, while the right figure is for TE modes. Here, $J'_n(x) = dJ_n(x)/dx$.

Table 2.3.1. Roots of $J'_n(x) = 0$.

n	β_{n1}	β_{n2}	β_{n3}	β_{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 2.3.2. 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 19.9: Table 2.3.1 shows the zeros of $J'_n(x)$, which are useful for determining the guidance conditions of the TE_{mn} mode of a circular waveguide. On the other hand, Table 2.3.2 shows the zeros of $J_n(x)$, which are useful for determining the guidance conditions of the TM_{mn} mode of a circular waveguide.

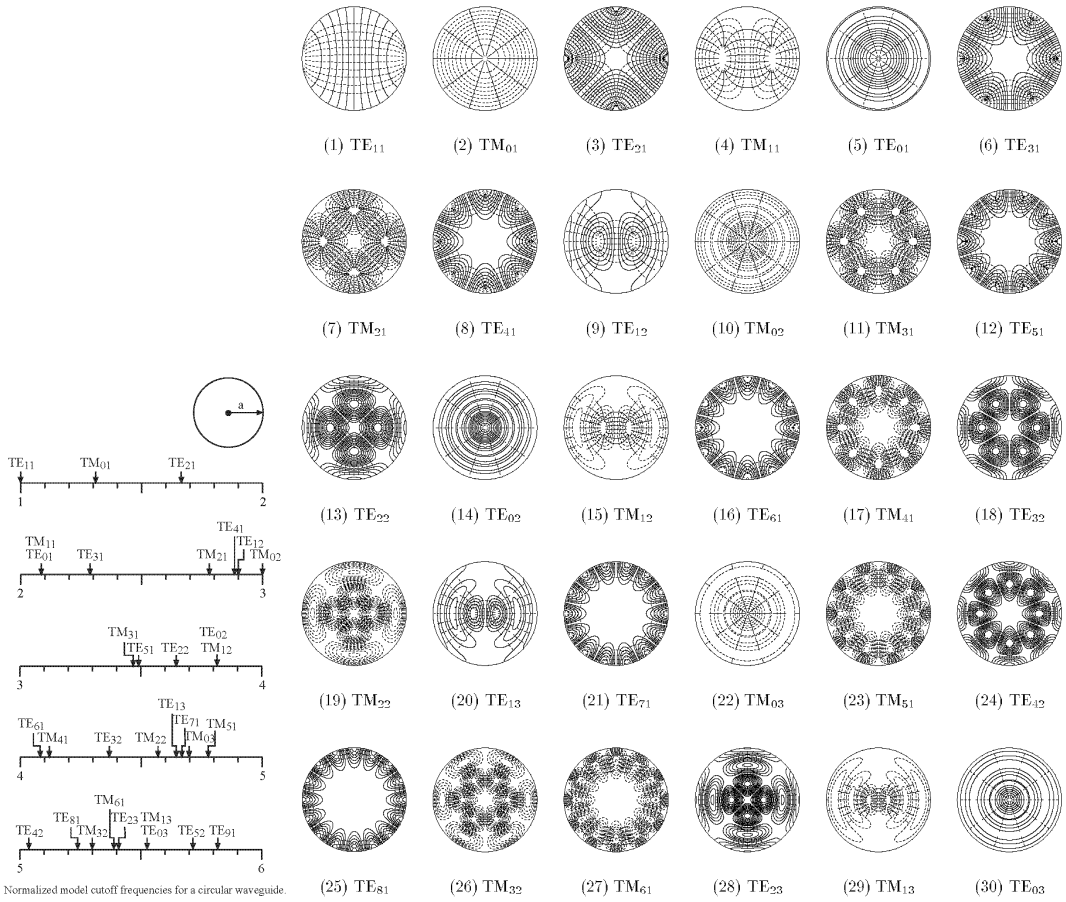


Figure 19.10: Transverse field plots of different modes in a circular waveguide (courtesy of Andy Greenwood. Original plots published in Lee, Lee, and Chuang [115]). The axially symmetric TE_{01} mode has the lowest loss, and finds a number of real-world applications as in radio astronomy.

Lecture 20

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. We will use heuristics to understand some systems whose designs follow from physical insight of simpler systems.

Also, 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 encountered in microwave and optical engineering by adding junction capacitances and inductances.

20.1 Circular Waveguides, Contd.

As in the rectangular waveguide case, the guidance of the wave in a circular waveguide can be viewed as bouncing waves in the radial direction. But these bouncing waves give rise to standing waves expressible in terms of Bessel functions. The scalar potential (or pilot potential) for the modes in the waveguide is expressible as

$$\Psi_{\alpha s}(\rho, \phi) = AJ_n(\beta_s \rho)e^{\pm jn\phi} \tag{20.1.1}$$

where $\alpha = h$ for TE waves and $\alpha = e$ for TM waves. The Bessel function or wave is expressible in terms of Hankel functions as in (19.2.5). Since Hankel functions are traveling waves, Bessel functions represent standing waves. Therefore, the Bessel waves can be thought of as bouncing traveling waves as 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.

20.1.1 An Application of Circular Waveguide

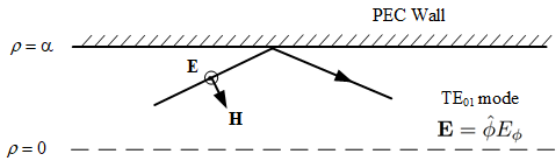


Figure 20.1: Bouncing wave picture of the Bessel wave inside a circular waveguide for the TE_{01} mode.

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 wall implying that $\hat{n} \cdot (\mathbf{E} \times \mathbf{H}^*) \neq 0$. Thus energy can dissipate 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 including rectangular waveguide modes. Hence, this waveguide mode is of interest to astronomers who are interested in building low-loss and low-noise systems.[1]

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 flowing in the ϕ direction. Moreover, by looking at a bouncing wave picture of the guided waveguide mode, this mode has a small component of tangential magnetic field on a waveguide wall: It becomes increasingly smaller as the frequency increases (see Figure 20.1). 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.[2] 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. The wall loss (or copper loss or eddy current loss) of the waveguide, hence, becomes smaller for higher frequencies. In fact, for high frequencies, the TE_{01} mode has the smallest copper loss of the waveguide modes: It becomes the mode of choice (see Figure 20.2). 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 [120] as shown in Figure 20.3.

[1] Low-loss systems are also low-noise due to energy conservation and the fluctuation dissipation theorem [113, 114, 119].

[2] Recall that for a fixed mode, β_s is independent of frequency.

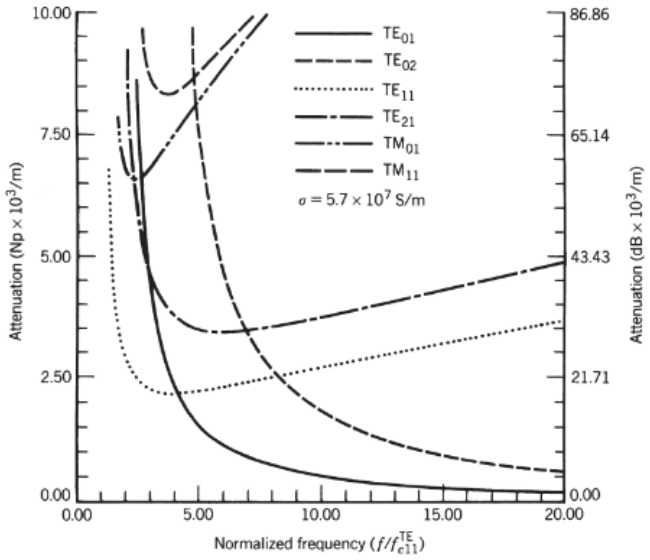


Figure 20.2: Losses of different modes in a circular waveguide or radius 1.5 cm . It is seen that at high frequencies, the TE_{01} mode has the lowest loss (courtesy of [121]).

Figure 20.3: Picture of the Very Large Array in New Mexico, USA (courtesy of [120]).

Figure 20.4 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 but not the TE_{01} mode, and (ii), by designing corrugated waveguide wall to discourage the flow of axial current and hence, the propagation of the non-TE_{01} mode. More details of circular waveguides can be found in [121]. Typical loss of a circular waveguide can be as low as 2 dB/km.[3]

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. Such antenna is called a horn antenna. 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. Ways to enhance the TE_{01} mode are also desirable [122] as shown in Figure 20.5.

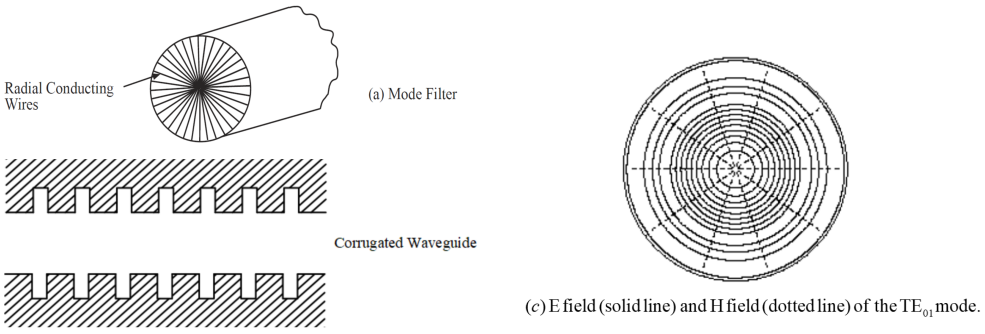


Figure 20.4: Ways to enhance the TE_{01} mode in a circular waveguide: (a) Using mode filter that only allows the mode to go through. (b) Use corrugated waveguide to discourage axial current flow. The field plot of the mode is shown in (c). Such waveguide is used in radio astronomy to design the communication links between antennas in a very large array (VLA [120]), or it is used in a circular horn antenna [122].

[3]For optical fiber, this figure of merit (FOM) can be lower than 1 dB/km making the optical fiber a darling for long-distance communication.

Figure 20.5: Picture of a circular horn antenna where corrugated wall is used to enhance the TE_{01} mode (courtesy of [123]).

20.2 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 into how waves are guided, and how they are cutoff 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.

20.2.1 Quasi-TEM Modes

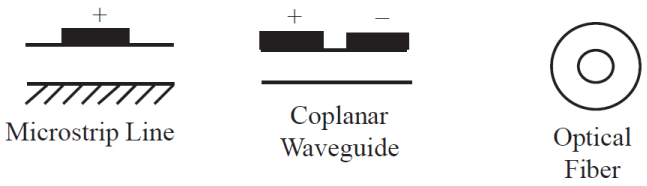


Figure 20.6: Some examples of practical coaxial-like waveguides are microstrip line and coplanar waveguide (left), and the optical fiber (right). The environments of these waveguides are inhomogeneous media, and hence, a pure TEM mode cannot propagate on these waveguides.

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 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 the optical fiber. When a wave is TEM, it is necessary that the wave propagates with the phase velocity of the medium. But when a uniform waveguide has inhomogeneity in between, as shown in Figure 20.6, this is not possible anymore. We can prove this assertion by **reductio ad absurdum**.

Assume only TE wave in a piecewise homogeneous region, then the $\mathbf{E}$ field is

$$\mathbf{E} = \frac{1}{j\omega\varepsilon_i}\nabla \times \nabla \times (\hat{z}\Psi_e) \tag{20.2.1}$$

where ε_i is the permittivity of the region. By doing some algebra, and assume that the field is a waveguide mode such that Ψ_e has $e^{-j\beta_z z}$ dependence, then one can show that E_z is given by

$$E_z = \frac{1}{j\omega\varepsilon_i}(\beta_i^2 - \beta_z^2)\Psi_e \tag{20.2.2}$$

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 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 20.6. 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 where β_i is the wavenumber of the i-th region. Hence, 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 inductances of such lines can be estimated, β_z can still be estimated. As shall be shown later, circuit theory is valid when the frequency is low, or the wavelength is large compared to the size of the structures.

20.2.2 Hybrid Modes–Inhomogeneously-Filled Waveguides

For most inhomogeneously filled waveguides, the modes (eigenmodes or eigenfunctions) inside are not cleanly classed into TE and TM modes, but with some modes that are the hybrid of TE and TM modes. If the inhomogeneity is piecewise constant, some of the equations we have derived before are still valid: In other words, in the homogeneous part (or constant part) of the

waveguide filled with piecewise constant inhomogeneity, the fields can still be decomposed into TE and TM fields. But these fields are coupled to each other by the presence of inhomogeneity, i.e., by the boundary conditions requisite at the interface between the piecewise homogeneous regions. Or both TE and TM waves are coupled together and are present simultaneously, and both $E_z \neq 0$ and $H_z \neq 0$. Some examples of inhomogeneously-filled waveguides where hybrid modes exist are shown in Figure 20.7.

Sometimes, the hybrid modes are called EH or HE modes, as in an optical fiber. Nevertheless, the guidance is 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 metalic waveguides, the reflection is due to the metal walls.

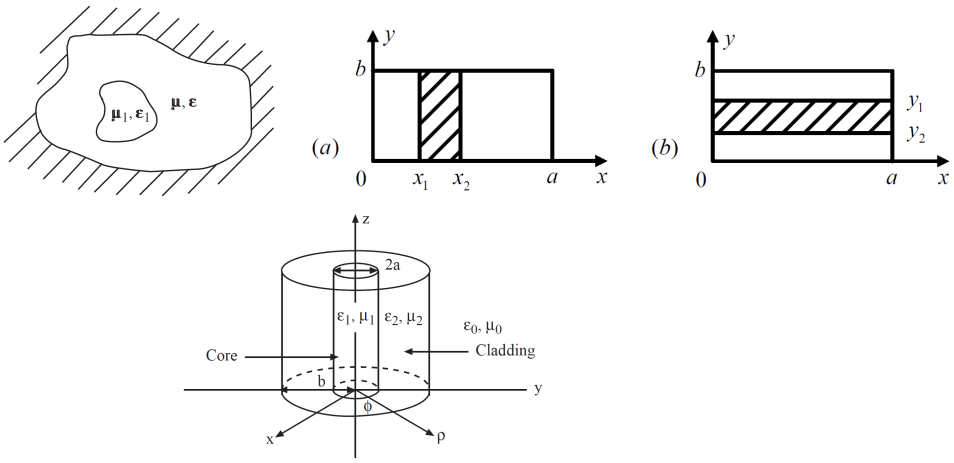


Figure 20.7: Some examples of inhomogeneously filled waveguides where hybrid modes exist: (top-left) A general inhomogeneously filled waveguide, (top-right) slab-loaded rectangular waveguides, and (bottom) an optical fiber with core and cladding.

20.2.3 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 thing 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.

In the case of hollow waveguides, the **E** and **H** fields are coupled through their space and time variations. Namely,

$$\nabla \times \mathbf{E} = -j\omega\mu\mathbf{H}, \qquad \nabla \times \mathbf{H} = j\omega\varepsilon\mathbf{E} \tag{20.2.3}$$

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. Hence, 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.

20.3 Homomorphism of 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 waveguides and transmission lines. But unlike the plane wave in layered medium case, we cannot replace the ∇ operator with $-j\boldsymbol{\beta}$ in a waveguide. Hence, the mathematics is slightly more elaborate. We can show this first for TE modes in a hallow waveguide, and the case for TM modes can be established by invoking duality principle.[4]

20.3.1 TE Case

For this case, $E_z = 0$, and from Maxwell's equations

$$\nabla \times \mathbf{H} = j\omega\varepsilon\mathbf{E} \tag{20.3.1}$$

By letting $\nabla = \nabla_s + \nabla_z$, $\mathbf{H} = \mathbf{H}_s + \mathbf{H}_z$ where $\nabla_z = \hat{z}\frac{\partial}{\partial z}$, and $\mathbf{H}_z = \hat{z}H_z$, and the subscript s implies transverse to z components, then

$$(\nabla_s + \nabla_z) \times (\mathbf{H}_s + \mathbf{H}_z) = \nabla_s \times \mathbf{H}_s + \nabla_z \times \mathbf{H}_s + \nabla_s \times \mathbf{H}_z \tag{20.3.2}$$

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. Therefore, by letting $\mathbf{E} = \mathbf{E}_s + \mathbf{E}_z$, and equating transverse components in (20.3.1), we have[5]

$$\nabla_z \times \mathbf{H}_s + \nabla_s \times \mathbf{H}_z = j\omega\varepsilon\mathbf{E}_s \tag{20.3.3}$$

To simplify the above equation, we shall remove $\mathbf{H}_z$ from above. Next, from Faraday's law, we have

$$\nabla \times \mathbf{E} = -j\omega\mu\mathbf{H} \tag{20.3.4}$$

Again, by letting $\mathbf{E} = \mathbf{E}_s + \mathbf{E}_z$, we can let (20.3.4) be written as

$$\nabla_s \times \mathbf{E}_s + \nabla_z \times \mathbf{E}_s + \nabla_s \times \mathbf{E}_z = -j\omega\mu(\mathbf{H}_s + \mathbf{H}_z) \tag{20.3.5}$$

[4] I have not seen exposition of such mathematical homomorphism elsewhere except in very simple cases [32].

[5] 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.

Equating z components of the above, we have

$$\nabla_s \times \mathbf{E}_s = -j\omega\mu \mathbf{H}_z \tag{20.3.6}$$

The above allows us to express $\mathbf{H}_z$ in terms of $\mathbf{E}_s$. Using (20.3.6), Eq.(20.3.3) can be rewritten as

$$\nabla_z \times \mathbf{H}_s + \nabla_s \times \frac{1}{-j\omega\mu} \nabla_s \times \mathbf{E}_s = +j\omega\varepsilon \mathbf{E}_s \tag{20.3.7}$$

The above can be further simplified by noting that

$$\nabla_s \times \nabla_s \times \mathbf{E}_s = \nabla_s (\nabla_s \cdot \mathbf{E}_s) - \nabla_s \cdot \nabla_s \mathbf{E}_s \tag{20.3.8}$$

But since $\nabla \cdot \mathbf{E} = 0$, and $E_z = 0$ for TE modes, it also implies that $\nabla_s \cdot \mathbf{E}_s = 0$. Also, from Maxwell's equations, we have previously shown that for a homogeneous source-free medium,

$$(\nabla^2 + \beta^2)\mathbf{E} = 0 \tag{20.3.9}$$

with $E_z = 0$ for TE mode, or that

$$(\nabla^2 + \beta^2)\mathbf{E}_s = 0 \tag{20.3.10}$$

Assuming that we have a guided mode, then

$$\mathbf{E}_s \sim e^{\mp j\beta_z z}, \qquad \frac{\partial^2}{\partial z^2}\mathbf{E}_s = -{\beta_z}^2 \mathbf{E}_s \tag{20.3.11}$$

Therefore, (20.3.10) becomes

$$({\nabla_s}^2 + \beta^2 - {\beta_z}^2)\mathbf{E}_s = 0 \tag{20.3.12}$$

or that

$$({\nabla_s}^2 + {\beta_s}^2)\mathbf{E}_s = 0 \tag{20.3.13}$$

where $\beta_s^2 = \beta^2 - \beta_z^2$ is the transverse wave number. Consequently, from (20.3.8)

$$\nabla_s \times \nabla_s \times \mathbf{E}_s = -\nabla_s^2 \mathbf{E}_s = \beta_s^2 \mathbf{E}_s \tag{20.3.14}$$

As such, (20.3.7) becomes

$$\begin{aligned} \nabla_z \times \mathbf{H}_s &= j\omega\varepsilon \mathbf{E}_s + \frac{1}{j\omega\mu} \nabla_s \times \nabla_s \times \mathbf{E}_s \\ &= j\omega\varepsilon \mathbf{E}_s + \frac{1}{j\omega\mu} {\beta_s}^2 \mathbf{E}_s \\ &= j\omega\varepsilon \left(1 - \frac{{\beta_s}^2}{\beta^2}\right) = j\omega\varepsilon \frac{{\beta_z}^2}{\beta^2} \mathbf{E}_s \end{aligned} \tag{20.3.15}$$

Letting $\beta_z = \beta\cos\theta$, then the above can further be rewritten as

$$\nabla_z \times \mathbf{H}_s = j\omega\varepsilon\cos^2\theta\mathbf{E}_s \tag{20.3.16}$$

The above now resembles one of the two telegrapher's equations that we seek. Now looking at (20.3.4) again, assuming $E_z = 0$, equating transverse components, we have

$$\nabla_z \times \mathbf{E}_s = -j\omega\mu\mathbf{H}_s \tag{20.3.17}$$

More explicitly, we can rewrite (20.3.16) and (20.3.17) in the above as

$$\frac{\partial}{\partial z}\hat{z} \times \mathbf{H}_s = j\omega\varepsilon\cos^2\theta\mathbf{E}_s \tag{20.3.18}$$

$$\frac{\partial}{\partial z}\hat{z} \times \mathbf{E}_s = -j\omega\mu\mathbf{H}_s \tag{20.3.19}$$

The above now resembles the telegrapher's equations. We can multiply (20.3.19) by $\hat{z}\times$ to get

$$\frac{\partial}{\partial z}\mathbf{E}_s = j\omega\mu\hat{z} \times \mathbf{H}_s \tag{20.3.20}$$

Now (20.3.18) and (20.3.20) 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 line is then

$$Z_0 = \sqrt{\frac{L}{C}} = \sqrt{\frac{\mu}{\varepsilon\cos^2\theta}} = \sqrt{\frac{\mu}{\varepsilon}}\frac{1}{\cos\theta} = \frac{\omega\mu}{\beta_z} \tag{20.3.21}$$

Therefore, 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

$$\beta_z = \sqrt{\beta^2 - \left(\frac{m\pi}{a}\right)^2 - \left(\frac{n\pi}{b}\right)^2} \tag{20.3.22}$$

Therefore, each TE_{mn} mode will be represented by a different characteristic impedance Z_0, since β_z is different for different TE_{mn} modes.

20.3.2 TM Case

This case can be derived using duality principle. Invoking duality, and after some algebra, then the equivalence of (20.3.18) and (20.3.20) become

$$\frac{\partial}{\partial z}\mathbf{E}_s = j\omega\mu\cos^2\theta\hat{z} \times \mathbf{H}_s \tag{20.3.23}$$

$$\frac{\partial}{\partial z}\hat{z}\times\mathbf{H}_s = j\omega\varepsilon\mathbf{E}_s \tag{20.3.24}$$

To keep the dimensions commensurate, we can let $\mathbf{E}_s \to V$, $\hat{z}\times\mathbf{H}_s \to -I$, $\mu\cos^2\theta \to L$, $\varepsilon \to C$, then the above resembles the telegrapher's equations. We can thus let

$$Z_0 = \sqrt{\frac{L}{C}} = \sqrt{\frac{\mu\cos^2\theta}{\varepsilon}} = \sqrt{\frac{\mu}{\varepsilon}}\cos\theta = \frac{\beta_z}{\omega\varepsilon} \tag{20.3.25}$$

Please note that (20.3.21) and (20.3.25) are very similar to that for the plane wave case, which are the wave impedance for the TE and TM modes, respectively.

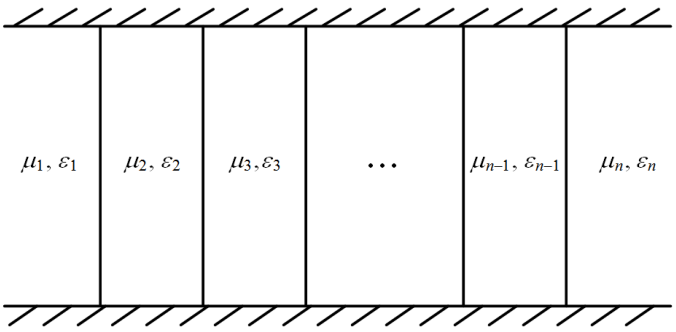


Figure 20.8: A waveguide filled with layered medium is mathematically homomorphic to a multi-section transmission line problem. Hence, transmission-line method can be used to solve this problem, but junction capacitance and inductance are needed to model the junctions correctly.

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, and 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. Hence, the transmission line solution would also imply continuous $\mathbf{E}$ and $\mathbf{H}$ field solutions.

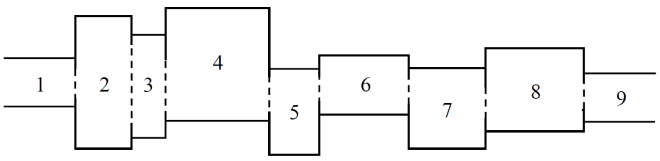


Figure 20.9: A multi-section waveguide is not exactly homormorphic to a multi-section transmission line problem, when the cross section of the waveguides are not equal to each other. Circuit elements are needed at the junctions to capture the physics at the waveguide junctions as shown in the next figure.

20.3.3 Mode Conversion

In the waveguide shown in Figure 20.8, there is no mode conversion at the junction interface. Assuming a rectangular waveguide as an example, what this means is that if we send at 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

$$\mathbf{H}_s = \nabla \times \hat{z}\Psi_{es}(\mathbf{r}_s)e^{\mp j\beta_z z} \tag{20.3.26}$$

$$\mathbf{E}_s = \frac{\mp\beta_z}{\omega\varepsilon}\nabla_s\Psi_{es}(\mathbf{r}_s)e^{\mp j\beta_z z} \tag{20.3.27}$$

In the above, β_s^2 and $\Psi_{es}(\mathbf{r}_s)$ are eigenvalue and eigenfunction, respectively, that depend only on the geometrical shape of the waveguide, but not the materials filling the waveguide. These eigenfunctions are the same throughout different sections of the waveguide. Therefore, boundary conditions can be easily satisfied at the junctions.

However, for a multi-junction waveguide show in Figure 20.9, tangential $\mathbf{E}$ and $\mathbf{H}$ 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 $\mathbf{E}$ and tangential $\mathbf{H}$ 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 [83]. 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.

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 cutoff or evanescent. In this case, the multiple modes at a junction give rise to localized energy storage at a junction. These energies can be either inductive or capacitive. The junction effect may be modeled by a simple circuit model as shown in Figure 20.10. 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.

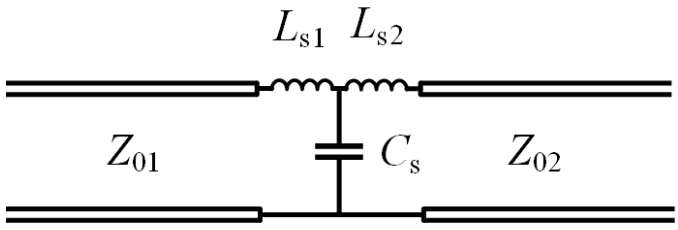


Figure 20.10: Junction circuit elements are used to account for stored electric and magnetic energies at the junction. They also account for that the currents and voltages are not continuous across the junctions anymore as the fields of the dominant modes in each section as shown in Figure 20.9 are not continuous anymore.

Lecture 21

Cavity Resonators

Cavity resonators are important components of microwave and optical systems. They work by constructive and destructive interference of waves in an enclosed region. They can be used as filters, or as devices to enhance certain physical interactions. These can be radiation antennas or electromagnetic sources such as magnetrons or lasers. They can also be used to enhance the sensitivity of sensors. We will study a number of them, and some of them, only heuristically in this lecture.

21.1 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), then no energy is dissipitated as a wave is 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.

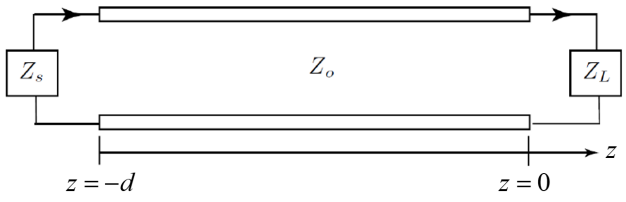


Figure 21.1: A simple resonator can be made by terminating a transmission line with two reactive loads at its two ends, the source end with Z_S and the load end with Z_L.

The transverse resonance condition for 1D problem can be used to derive the resonance condition, namely that

$$1 = \Gamma_S \Gamma_L e^{-2j\beta_z d} \tag{21.1.1}$$

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 a mixed medium. In general,

$$\beta_e = \omega / v_e \tag{21.1.2}$$

where v_e is the effective phase velocity of the wave in the heterogeneous structure.

When the source and load impedances are replaced by short or open circuits, then the reflection coefficients are -1 for a short, and $+1$ for an open circuit. The (21.1.1) above then becomes

$$\pm 1 = e^{-2j\beta_e d} \tag{21.1.3}$$

The $\pm$ sign corresponds to different combinations of open and short circuits at the two ends of the transmission lines. When a "+" sign is chosen, which corresponds to either both ends are short circuit, or are open circuit, the resonance condition is such that

$$\beta_e d = p\pi, \quad p = 0, 1, 2, \ldots, \quad \text{or integer} \tag{21.1.4}$$

For a TEM or a quasi-TEM mode in a transmission line, $p = 0$ is not allowed as the voltage will be uniformly zero on the transmisson line or $V(z) = 0$ for all z implying a trivial solution. The lowest mode then is when $p = 1$ corresponding to a half wavelength on the transmission line.

When the line is open at one end, and shorted at the other end in (21.1.1), the resonance condition corresponds to the "$-$" sign in (21.1.3), which gives rise to

$$\beta_e d = p\pi/2, \quad p \quad \text{odd} \tag{21.1.5}$$

The lowest mode is when $p = 1$ corresponding to a quarter wavelength on the transmission line, which is smaller than that of a transmission line terminated with short or open at both ends. Designing a small 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 made with a coax is shown in Figure 21.2. It is easier to make a short indicated at the left end, 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 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 21.2. This indicates that the current does not abruptly go to zero at the right-hand side due to the presence of fringing field and hence, displacement current.

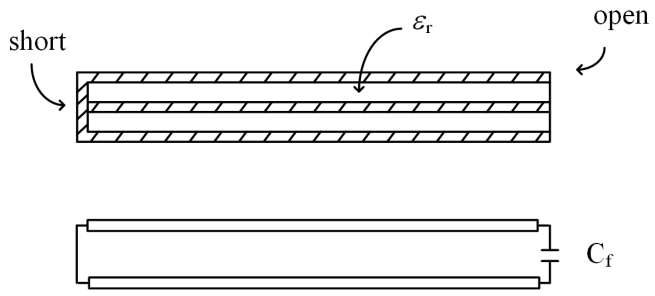


Figure 21.2: A short and open circuited transmission line can be a resonator, but the open end has to be modeled with a fringing field capacitance C_f since there is no exact open circuit.

21.2 Cylindrical Waveguide Resonators

Since a cylindrical waveguide[1] 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 21.3.

[1]Both rectangular and circular waveguides are cylindrical waveguides.

Figure 21.3: 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 also radiates (courtesy of RFcurrent.com).

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 previous lecture,

$$\beta_z = \sqrt{\beta^2 - \left(\frac{m\pi}{a}\right)^2 - \left(\frac{n\pi}{b}\right)^2} \tag{21.2.1}$$

for both TE_{mn} and TM_{mn} modes.[2] If the waveguide is terminated with two shorts (which is easy to make) at its ends, then the resonance condition is that

$$\beta_z = p\pi/d, \quad p \text{ integer} \tag{21.2.2}$$

Together, using (21.2.1), we have the condition that

$$\beta^2 = \frac{\omega^2}{c^2} = \left(\frac{m\pi}{a}\right)^2 + \left(\frac{n\pi}{b}\right)^2 + \left(\frac{p\pi}{d}\right)^2 \tag{21.2.3}$$

The above can only be satisfied by certain select frequencies, and 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.

[2]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.

The entire electromagnetic fields of the cavity can be found from the scalar potentials previously defined, namely that

$$\mathbf{E} = \nabla \times \hat{z}\Psi_h, \quad \mathbf{H} = \nabla \times \mathbf{E}/(-j\omega) \tag{21.2.4}$$

$$\mathbf{H} = \nabla \times \hat{z}\Psi_e, \quad \mathbf{E} = \nabla \times \mathbf{H}/(j\omega\varepsilon) \tag{21.2.5}$$

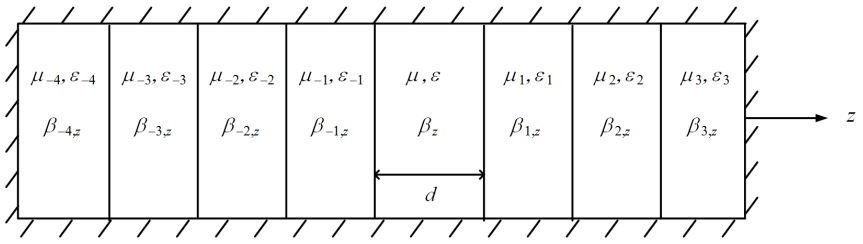


Figure 21.4: A waveguide filled with layered dielectrics can also become a resonator. The transverse resonance condition can be used to find the resonant modes.

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 modes of a waveguide cavity loaded with layered medium as shown in Figure 21.4. This condition is repeated below as:

$$\tilde{R}_- \tilde{R}_+ e^{-2j\beta_z d} = 1 \tag{21.2.6}$$

where d is the length of the waveguide section where the above is applied, and $\tilde{R}_-$ and $\tilde{R}_+$ are the generalized reflection coefficient to the left and right of the center waveguide section. The above is similar to the resonant condition using the transmission line model in (21.1.1), except that now, we have replaced the transmission line reflection coefficient with TE or TM generalized reflection coefficients.

21.2.1 $\beta_z = 0$ Case

In this case, we can still look at the TE and the TM modes in the waveguide. This corresponds to a waveguide mode that bounces off the waveguide wall, but make no progress in the z direction. The modes are independent of z since $\beta_z = 0$. It is quite easy to show that for the TE case, a z-independent $\mathbf{H} = \hat{z}H_0$, and $\mathbf{E} = \mathbf{E}_s$ exist inside the waveguide, and for the TM case, a z-independent $\mathbf{E} = \hat{z}E_0$, and $\mathbf{H} = \mathbf{H}_s$ being the only components in the waveguide.

Consider now a single section waveguide. For the TE mode, if either one of the ends of the waveguide is terminated with a PEC wall, then $\hat{n} \cdot \mathbf{H} = 0$ at the end. This will force the z-independent $\mathbf{H}$ field to be zero in the entire waveguide. Thus for the TE mode, it can only exist if both ends are terminated with open, but this mode is not trapped inside since it easily leaks energy to the outside via the ends of the waveguide.

For the TM mode, since $\mathbf{E} = \hat{z}E_0$, it easily satisfy the boundary condition if both ends are terminated with PEC walls since the boundary condition is that $\hat{n} \times \mathbf{E} = 0$. The wonderful part about this mode is that the length or d of the cavity can be as short as possible.

21.2.2 Lowest Mode of a Rectangular Cavity

The lowest TM mode in a rectanglar waveguide is the TM_{11} mode. At the cutoff of this mode, the $\beta_z = 0$ or $p = 0$, implying no variation of the field in the z direction. When the two ends are terminated with metallic shorts, the tangential magnetic field is not shorted out. But the tangential electric field is shorted to zero in the entire cavity, or that the TE mode cannot exist. Therefore, the longitudinal electric field of the TM mode still exists (see Figures 21.5 and 21.6). As such, for the TM mode, $m = 1$, $n = 1$ and $p = 0$ is possible giving a non-zero field in the cavity. This is the TM_{110} mode of the resonant cavity, which is the lowest mode in the cavity if $a > b > d$. The top and side views of the fields of this mode is shown in Figures 21.5 and 21.6. The corresponding resonant frequency of this mode satisfies the equation

$$\frac{\omega_{110}^2}{c^2} = \left(\frac{\pi}{a}\right)^2 + \left(\frac{\pi}{b}\right)^2 \tag{21.2.7}$$

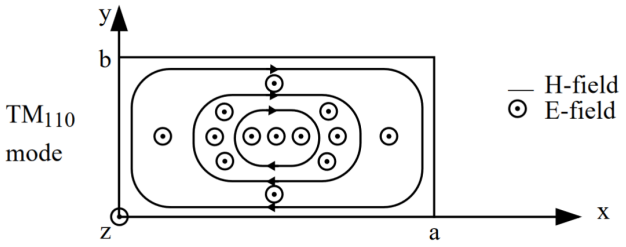


Figure 21.5: The top view of the E and H fields of a rectangular resonant cavity.

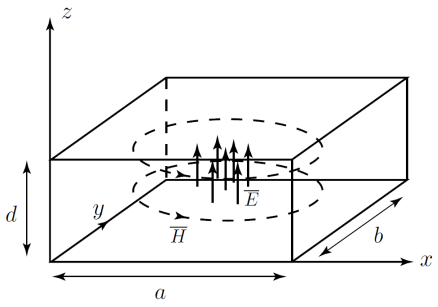


Figure 21.6: The side view of the E and H fields of a rectangular resonant cavity (courtesy of J.A. Kong [32]).

For the TE modes, 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 **E** field. The resonant frequency

of this mode is

$$\frac{\omega_{101}^2}{c^2} = \left(\frac{\pi}{a}\right)^2 + \left(\frac{\pi}{d}\right)^2 \tag{21.2.8}$$

Clearly, this mode has a higher resonant frequency compared to the TM_{110} mode if $d < b$.

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, β_s may have to be determined numerically.

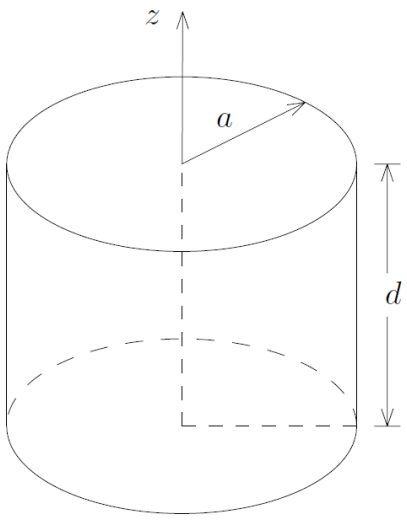


Figure 21.7: A circular resonant cavity made by terminating a circular waveguide (courtesy of Kong [32]).

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 [32].

21.3 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. They can be used in klystrons and magnetrons as microwave sources, 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.

21.3.1 Filters

Microstrip line resonators are often used to make filters. Transmission lines are often used to model microstrip lines in a 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 21.8 [124, 125]. The fringing field gives rise to fringing field capacitance as shown in Figure 21.2. Then the appropriate Γ_S and Γ_L can be used to model the effect of fringing field capacitance. Figure 21.9 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 21.10.

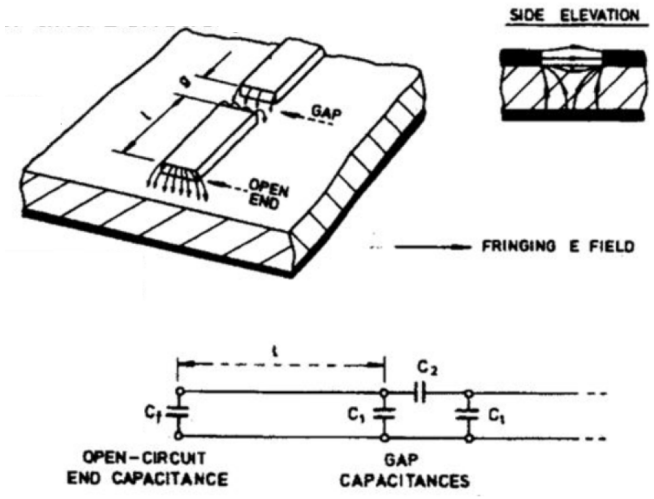


Figure 21.8: End effects and junction effects in a microwave integrated circuit [124, 125] (courtesy of Microwave Journal).

Figure 21.9: 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).

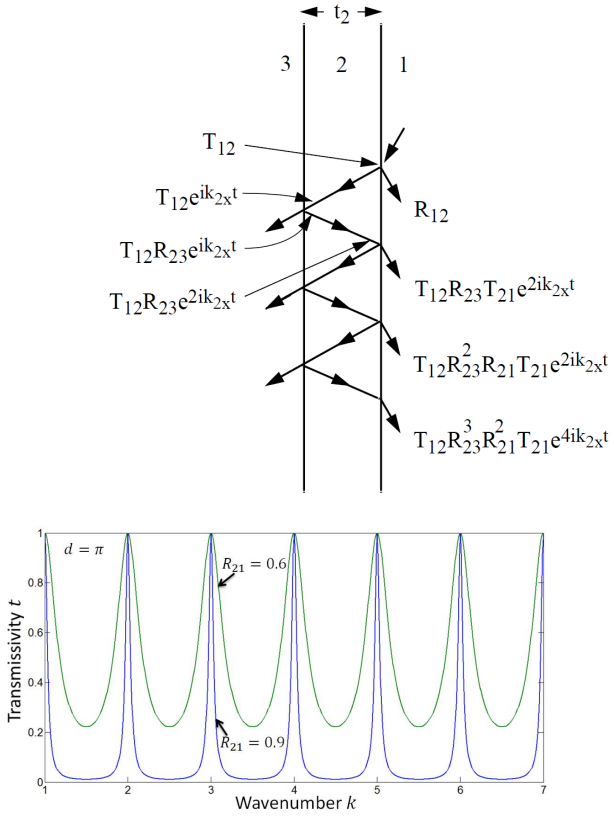


Figure 21.10: Design of a Fabry-Perot resonator [56, 83, 126, 127].

21.3.2 Electromagnetic Sources

Microwave sources are often made by transferring kinetic energy from an electron beam to microwave energy. Klystrons, magnetrons, and traveling wave tubes are such devices.

However, the cavity resonator in a klystron enhances the interaction of the electrons with the microwave field allowing for such energy transfer, causing the field to grow in amplitude as shown in Figure 21.11.

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 absorbing kinetic energy from the hot electrons, giving rise to microwave energy.

Figure 21.13 shows laser cavity resonator to enhance of light wave interaction with material media. By using stimulated emission of electronic transition, light energy can be produced.

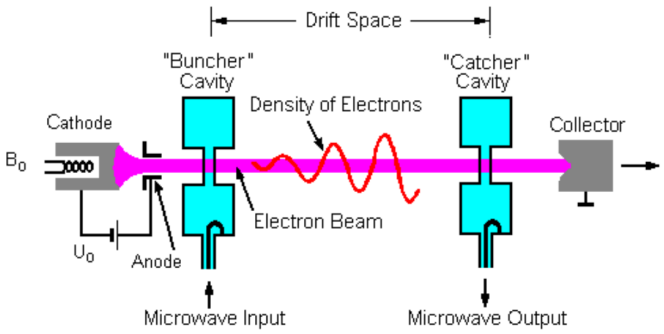


Figure 21.11: A klystron works by converting the kinetic energy of an electron beam into the energy of a traveling microwave next to the beam (courtesy of Wiki [128]).

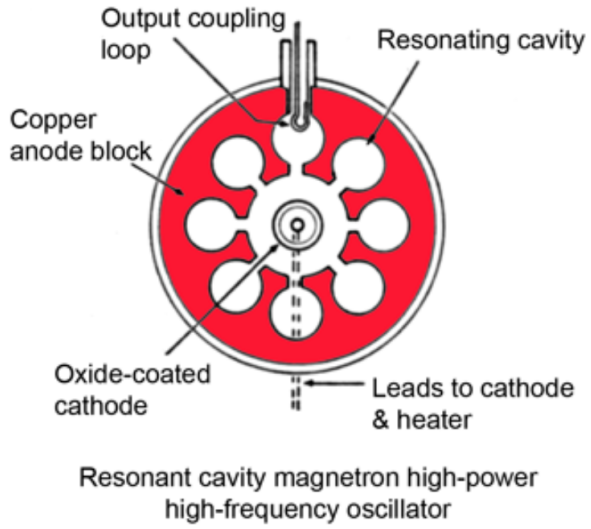


Figure 21.12: A magnetron works by having a high-Q microwave cavity resonator. When the cavity is injected with energetic electrons from the cathode to the anode, the kinetic energy of the electron feeds into the energy of the microwave (courtesy of Wiki [129]).

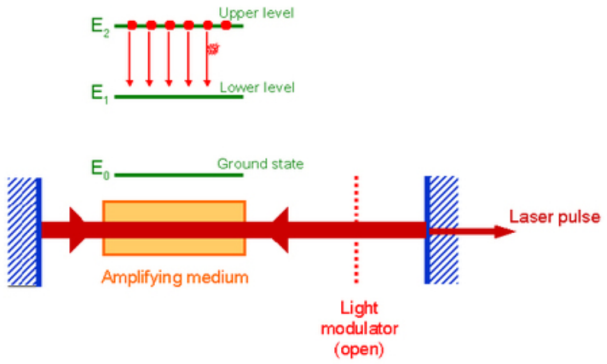


Figure 21.13: A simple view of the physical principle behind the working of the laser (courtesy of www.optique-ingenieur.org).

Energy trapping of a waveguide or a resonator can be used to enhance the efficiency of a semiconductor laser as shown in Figure 21.14. 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.

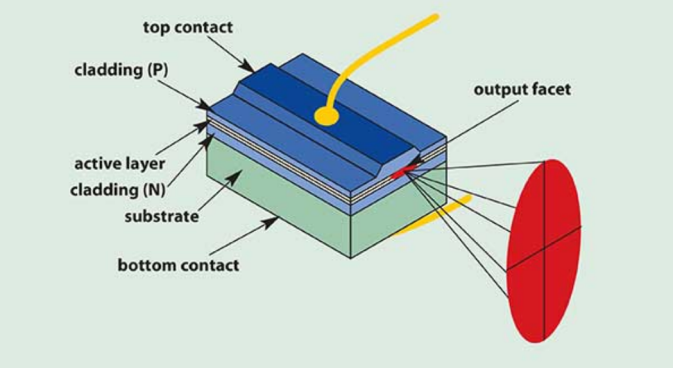


Figure 21.14: A semiconductor laser at work. Room temperature lasing is possible due to both the tight confinement of light and carriers (courtesy of Photonics.com).

21.3.3 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 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 21.15 and 21.16. As seen in the picture, there is an entry microwave port for injecting microwave into the cavity, and another exit port for the microwave to leave the cavity sensor. The passing microwave, when it hits the resonance frequency of the cavity, will create a large field inside it. 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 reveals the frequency of the microwave.

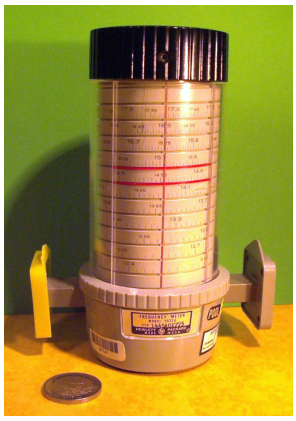


Figure 21.15: An absorption wave meter can be used to measure the frequency of microwave (courtesy of Wiki [130]).

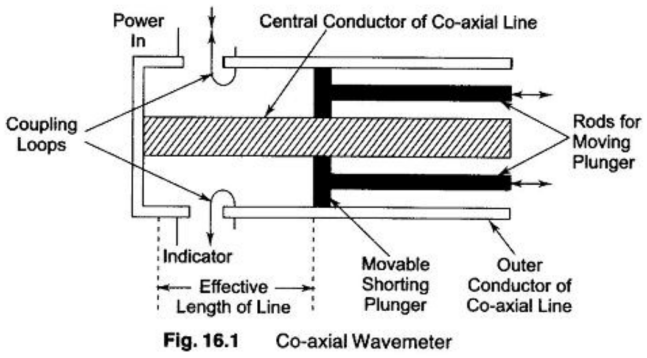


Figure 21.16: The innards of a wavemeter (courtesy of eeeguide.com).

Lecture 22

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 to the cavity. 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 this 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.

22.1 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. 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.

22.1.1 Analogue with an LC Tank Circuit

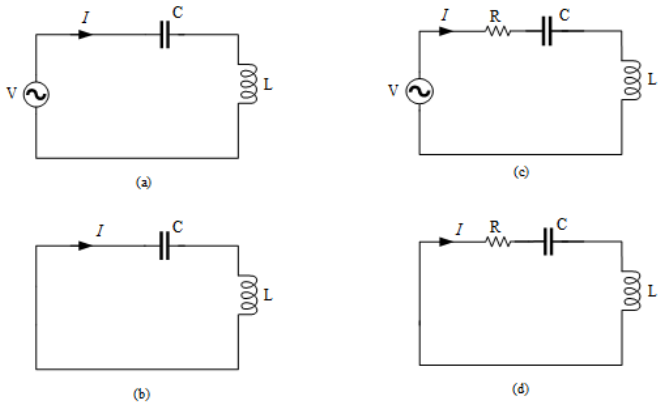


Figure 22.1: For the circuit on the left, it will resonate forever even if the source is turned off. But for the circuit on the right, the current in the circuit will decay with time due to dissipation in the resistor.

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. When there is no loss in an LC tank circuit, it can oscillate forever. Moreover, if we turn off the source, a free oscillation solution exists.[1] One can write the voltage-current relation in the circuit as

$$I = \frac{V}{j\omega L + 1/(j\omega C)} = VY \tag{22.1.1}$$

where

$$Y(\omega) = \frac{1}{j\omega L + 1/(j\omega C)} \tag{22.1.2}$$

The above $Y(\omega)$ can be thought of as the transfer function of the linear system where the in put is $V(\omega)$ and the output is $I(\omega)$. When the voltage is zero or turned off, a non-zero current exist or persist 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 $V = 0$. This resonant frequency, obtained by setting the denominator of Y to zero, is given by $\omega_R = 1/\sqrt{LC}$.

When a small resistor is added in the circuit to give rise to loss, the voltage-current relation becomes

$$I = \frac{V}{j\omega L + R + 1/(j\omega C)} = VY \tag{22.1.3}$$

$$Y = \frac{1}{j\omega L + R + 1/(j\omega C)} \tag{22.1.4}$$

[1] This is analogous to the homogeneous solution of an ordinary differential equation.

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. The homogeneous solution (also called the natural solution, or free oscillation) can only exists at the complex resonant frequencies. With complex resonances, the voltage and the current are decaying sinusoids.

By the same token, because of losses, the free oscillation in a cavity has electromagnetic field with time dependence as follows:

$$\mathbf{E} \propto e^{-\alpha t}\cos(\omega t + \phi_1), \quad \mathbf{H} \propto e^{-\alpha t}\cos(\omega t + \phi_2) \tag{22.1.5}$$

That is, they are decaying sinusoids. The total time-average stored energy, which is proportional to $\frac{1}{4}\epsilon|\mathbf{E}|^2 + \frac{1}{4}\mu|\mathbf{H}|^2$ is of the form

$$\langle W_T\rangle = \langle W_E\rangle + \langle W_H\rangle \doteq W_0 e^{-2\alpha t} \tag{22.1.6}$$

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 22.2). 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 [32]

$$Q \doteq \frac{\omega}{2\alpha} \tag{22.1.7}$$

Q is an approximate concept, and makes sense only if the system has low loss.

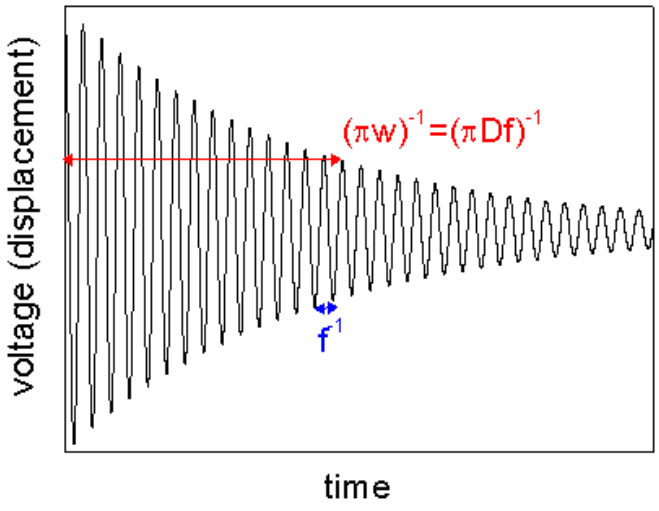


Figure 22.2: A typical time domain response of a high Q system (courtesy of Wikipedia).

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,

$$\langle P_D \rangle = -\frac{d\langle W_T \rangle}{dt} \tag{22.1.8}$$

By further assuming that W_T has to be of the form in (22.1.6), then

$$-\frac{d\langle W_T \rangle}{dt} \doteq 2\alpha W_0 e^{-2\alpha t} = 2\alpha \langle W_t \rangle \tag{22.1.9}$$

From the above, we can estimate the decay constant

$$\alpha \doteq \frac{\langle P_D \rangle}{2\langle W_T \rangle} \tag{22.1.10}$$

Hence, we can write equation (22.1.7) for the Q as

$$Q \doteq \frac{\omega \langle W_T \rangle}{\langle P_D \rangle} \tag{22.1.11}$$

By further letting $\omega = 2\pi/T$, we lent further physical interpretation to express Q as

$$Q \doteq 2\pi \frac{\langle W_T \rangle}{\langle P_D \rangle T} = 2\pi \frac{\text{total energy stored}}{\text{Energy dissipated/cycle}} \tag{22.1.12}$$

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 the Q is a low-loss concept, and hence, the above formulas are only approximately true.

22.1.2 Relation to Bandwidth and Pole Location

The resonance of a system is related to the pole of the transfer function . For instance, in our previous examples of the RLC tank circuit, the admittance Y can be thought of as a transfer function in linear system theory: The input is the voltage, while the output is the current. 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. In other words, in the vicinity of the pole in the frequency domain, the transfer function of the system can be approximated by a single pole which is

$$Y(\omega) \sim \frac{A}{\omega - \omega_R} = \frac{A}{\omega - \omega_0 - j\alpha} \tag{22.1.13}$$

where we have assumed that $\omega_R = \omega_0 + j\alpha$, the resonant frequency is complex. In principle, when $\omega = \omega_R$, the reflection coefficient becomes infinite, but this does not happen in practice because ω_R is complex, and ω, the operating frequency is real. In other words, the pole is displaced slightly off the real axis to account for loss. Using a single pole approximation, 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 $\omega = \omega_0 \pm \alpha$. Therefore, the full-width half maximum (FWHM) bandwidth is $\Delta\omega = 2\alpha$. Thus the Q can be written as in terms of the half-power bandwidth $\Delta\omega$ of the system, viz.,

$$Q \doteq \omega_0/\Delta\omega \tag{22.1.14}$$

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 22.3.

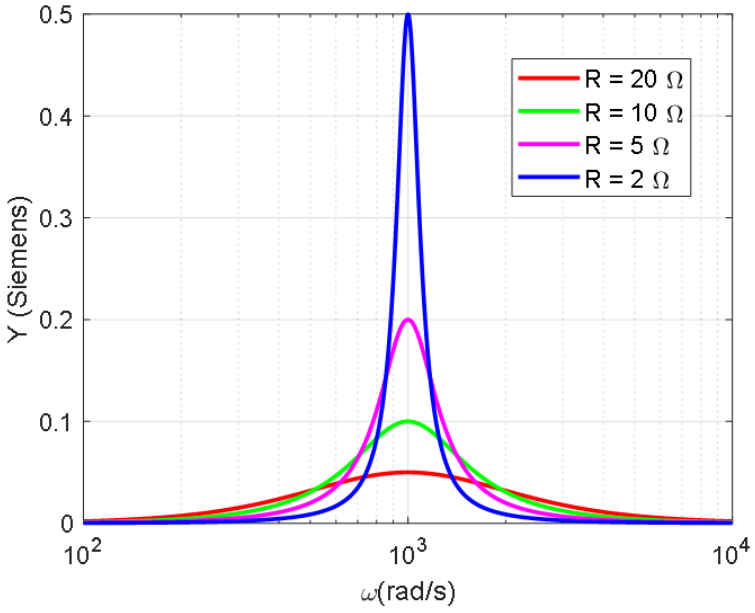


Figure 22.3: A typical system response versus frequency for different Q's using (22.1.4). The Q is altered by changing the resistor R in the circuit.

22.1.3 Wall Loss and Q for a Metallic Cavity

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 power dissipated on the wall is given by [32]

$$\langle P_D \rangle = \frac{1}{2}\Re e \oint_S (\mathbf{E} \times \mathbf{H}^*) \cdot \hat{n}\, dS = \frac{1}{2}\Re e \oint_S (\hat{n} \times \mathbf{E}) \cdot \mathbf{H}^*\, dS \tag{22.1.15}$$

where S is the surface of the cavity wall.[2] Here, $(\hat{n} \times \mathbf{E})$ is the tangential component of the electric field which would have been zero if the cavity wall is made of ideal PEC. Also,

[2]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).

$\hat{n}$ 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 β vector is almost normal to the interface. 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.

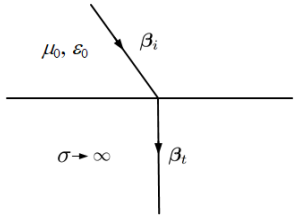


Figure 22.4: Due to high σ in the metal, and large $|\boldsymbol{\beta}_t|$, phase matching condition requires the transmitted wave vector to be almost normal to the surface.

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 $Z_m = \sqrt{\frac{\mu}{\varepsilon_m}} \approx \sqrt{\frac{\mu}{-j\frac{\sigma}{\omega}}} = \sqrt{\frac{\omega\mu}{2\sigma}}(1+j)$,[3] where we have assumed that $\varepsilon_m \approx -\frac{j\sigma}{\omega}$, and $\mathbf{H}_t$ is the tangential magnetic field. This relation between $\mathbf{E}$ and $\mathbf{H}$ will ensure that power is flowing into the metallic surface. Hence,

$$\langle P_D \rangle = \frac{1}{2}\Re e \oint_S \sqrt{\frac{\omega\mu}{2\sigma}}(1+j)\left|\mathbf{H}_t\right|^2 dS = \frac{1}{2}\sqrt{\frac{\omega\mu}{2\sigma}} \oint_S \left|\mathbf{H}_t\right|^2 dS \tag{22.1.16}$$

By further assuming that the stored electric and magnetic energies of a cavity are equal to each other at resonance, the stored energy can be obtained by

$$\langle W_T \rangle = \frac{1}{2}\mu \int_V |\mathbf{H}|^2 dV \tag{22.1.17}$$

Written explicitly, the Q becomes

$$Q = \sqrt{2\omega\mu\sigma}\frac{\int_V |\mathbf{H}|^2 \, dV}{\oint_S |\mathbf{H}_t|^2 \, dS} = \frac{2}{\delta}\frac{\int_V |\mathbf{H}|^2 \, dV}{\oint_S |\mathbf{H}_t|^2 \, dS} \tag{22.1.18}$$

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 (22.1.18), 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

[3]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 22.4, and hence, this formula can be applied.

Q, a dimensionless quantity, is roughly proportional to

$$Q \sim \frac{V}{S\delta} \tag{22.1.19}$$

where V is the volume of the cavity, while S is its surface area. From the above, it is noted that a large cavity compared to its skin depth has a larger Q than an small cavity.

22.1.4 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 $\mathbf{E} = \hat{z}E_z$, where

$$E_z = E_0 \sin\left(\frac{\pi x}{a}\right) \sin\left(\frac{\pi y}{b}\right) \tag{22.1.20}$$

The magnetic field can be derived from the electric field using Maxwell's equation or Faraday's law, and

$$H_x = \frac{j\omega\epsilon}{\omega^2\mu\epsilon}\frac{\partial}{\partial y}E_Z = \frac{j\left(\frac{\pi}{b}\right)}{\omega\mu}E_0 \sin\left(\frac{\pi x}{a}\cos\right)\left(\frac{\pi y}{b}\right) \tag{22.1.21}$$

$$H_y = \frac{-j\omega\epsilon}{\omega^2\mu\epsilon}\frac{\partial}{\partial x}E_Z = -\frac{j\left(\frac{\pi}{a}\right)}{\omega\mu}E_0 \cos\left(\frac{\pi x}{a}\right)\sin\left(\frac{\pi y}{b}\right) \tag{22.1.22}$$

Therefore

$$\begin{aligned}
\int_V |\mathbf{H}|^2\, dV &= \int_{-d}^{0}\int_0^b\int_0^a dx\, dy\, dz \left[|H_x|^2 + |H_y|^2\right] \\
&= \frac{|E_0|^2}{\omega^2\mu^2}\int_{-d}^{0}\int_0^b\int_0^a dx\, dy\, dz \\
&\quad \left[\left(\frac{\pi}{b}\right)^2 \sin^2\left(\frac{\pi x}{a}\right)\cos^2\left(\frac{\pi y}{b}\right) + \left(\frac{\pi}{a}\right)^2 \cos^2\left(\frac{\pi x}{a}\right)\sin^2\left(\frac{\pi y}{b}\right)\right] \\
&= \frac{|E_0|^2}{\omega^2\mu^2}\frac{\pi^2}{4}\left[\frac{a}{b} + \frac{b}{a}\right] d
\end{aligned} \tag{22.1.23}$$

A cavity has six faces, finding the tangential exponent at each face and integrate

$$\begin{aligned}
\oint_S |\mathbf{H}_t|\, dS &= 2\int_0^b\int_0^a dx\, dy \left[|H_x|^2 + |H_y|^2\right] \\
&\quad + 2\int_{-d}^{0}\int_0^a dx\, dz\, |H_x(y=0)|^2 + 2\int_{-d}^{0}\int_0^b dy\, dz\, |H_y(x=0)|^2 \\
&= \frac{2\,|E_0|^2}{\omega^2\mu^2}\frac{\pi^2 ab}{4}\left[\frac{1}{a^2} + \frac{1}{b^2}\right] + \frac{2\left(\frac{\pi}{b}\right)^2}{\omega^2\mu^2}|E_0|^2\,\frac{ad}{2} + \frac{2\left(\frac{\pi}{a}\right)^2}{\omega^2\mu^2}|E_0|^2\,\frac{bd}{2} \\
&= \frac{\pi^2\,|E_0|^2}{\omega^2\mu^2}\left[\frac{b}{2a} + \frac{a}{2b} + \frac{ad}{b^2} + \frac{bd}{a^2}\right]
\end{aligned} \tag{22.1.24}$$

Hence the Q is

$$Q = \frac{2}{\delta} \frac{\left(\frac{ad}{b} + \frac{bd}{a}\right)}{\left(\frac{b}{2a} + \frac{a}{2b} + \frac{ad}{b^2} + \frac{bd}{a^2}\right)} \tag{22.1.25}$$

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.

22.2 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.[4] 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.

22.2.1 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 [77]. Assume that an eigenvalue and an eigenvector exists for the hermitian matrix $\overline{\mathbf{A}}$. Then

$$\overline{\mathbf{A}} \cdot \mathbf{v}_i = \lambda_i \mathbf{v}_i \tag{22.2.1}$$

Dot multiplying the above from the left by $\mathbf{v}_i^\dagger$ where $\dagger$ indicates conjugate transpose, then the above becomes

$$\mathbf{v}_i^\dagger \cdot \overline{\mathbf{A}} \cdot \mathbf{v}_i = \lambda_i \mathbf{v}_i^\dagger \cdot \mathbf{v}_i \tag{22.2.2}$$

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. 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

$$\overline{\mathbf{A}} \cdot \mathbf{v}_i = \lambda_i \mathbf{v}_i \tag{22.2.3}$$

$$\overline{\mathbf{A}} \cdot \mathbf{v}_j = \lambda_j \mathbf{v}_j \tag{22.2.4}$$

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

$$\mathbf{v}_j^\dagger \cdot \overline{\mathbf{A}} \cdot \mathbf{v}_i = \lambda_i \mathbf{v}_j^\dagger \cdot \mathbf{v}_i \tag{22.2.5}$$

$$\mathbf{v}_i^\dagger \cdot \overline{\mathbf{A}} \cdot \mathbf{v}_j = \lambda_j \mathbf{v}_i^\dagger \cdot \mathbf{v}_j \tag{22.2.6}$$

[4] This mathematical homomorphism is not discussed in any other electromagnetic textbooks.

Taking the conjugate transpose of (22.2.5) in the above, and since $\overline{\mathbf{A}}$ is hermitian, their left-hand sides (22.2.5) and (22.2.6) become the same. Subtracting the two equations, we arrive at

$$0 = (\lambda_i - \lambda_j)\mathbf{v}_j^\dagger \cdot \mathbf{v}_i \tag{22.2.7}$$

For distinct eigenvalues, $\lambda_i \neq \lambda_j$, the only way for the above to be satisfied is that

$$\mathbf{v}_j^\dagger \cdot \mathbf{v}_i = C_i \delta_{ij} \tag{22.2.8}$$

Hence, eigenvectors of a hermitian matrix with distinct eigenvalues are orthogonal to each other. The eigenvalues are also real.

22.2.2 Homomorphism with the Waveguide Mode Problem

We shall next show that the problem for finding the waveguide modes or eigenmodes is analogous to the matrix eigenvalue problem. The governing equation for a waveguide mode is a BVP involving the reduced wave equation previously derived, or

$$\nabla_s^2 \psi_i(\mathbf{r}_s) + \beta_{is}^2 \psi_i(\mathbf{r}_s) = 0, \Rightarrow -\nabla_s^2 \psi_i(\mathbf{r}_s) = \beta_{is}^2 \psi_i(\mathbf{r}_s) \tag{22.2.9}$$

with the pertinent homogeneous Dirichlet or Neumann boundary condition, depending on if TE or TM modes are considered. In the above, the differential operator ∇_s^2 is analogous to the matrix operator $\overline{\mathbf{A}}$, the eigenfunction $\psi_i(\mathbf{r}_s)$ is analogous to the eigenvector $\mathbf{v}_i$, and β_{is}^2 is analogous to the eigenvalue λ_i.

In the above, we can think of ∇_s^2 as a linear operator that maps a function to another function, viz.,

$$\nabla_s^2 \psi_i(\mathbf{r}_s) = f_i(\mathbf{r}_s) \tag{22.2.10}$$

Hence it is analogous to a matrix operator $\overline{\mathbf{A}}$, that maps a vector to another vector, viz.,

$$\overline{\mathbf{A}} \cdot \mathbf{x} = \mathbf{f} \tag{22.2.11}$$

We shall elaborate this further next.

Discussion on Functional Space

To think of a function $\psi(\mathbf{r}_s)$ as a vector, one has to think in the discrete world.[5] If one needs to display the function $\psi(\mathbf{r}_s)$, on a computer, one will evaluate the function $\psi(\mathbf{r}_s)$ at discrete N locations $\mathbf{r}_{ls}$, where $l = 1, 2, 3, \ldots, N$. For every $\mathbf{r}_{ls}$ or every l, there is a scalar number $\psi(\mathbf{r}_{ls})$. These scalar numbers can be stored in a column vector in a computer indexed by l. The larger N is, the better is the discrete approximation of $\psi(\mathbf{r}_s)$. In theory, N needs to be infinite to describe this function exactly.[6]

[5] Some of these concepts are discussed in [35, 131].

[6] In mathematical parlance, the index for $\psi(\mathbf{r}_s)$ is uncountably infinite or nondenumerable.

From the above discussion, a function is analogous to a vector and a functional space is analogous to a vector space in linear or matrix algebra. However, a functional space is infinite dimensional. But in order to compute on a computer with finite resource, such functions are approximated with finite length vectors. Or infinite dimensional vector spaces are replaced with finite dimensional ones to make the problem computable. Such an infinite dimensional functional space is also called a Hilbert space.

It is also necessary to define the inner product between two vectors in a functional space just as inner product between two vectors in an matrix vector space. The inner product (or dot product) between two vectors in matrix vector space is

$$\mathbf{v}_i^t \cdot \mathbf{v}_j = \sum_{l=1}^{N} v_{i,l} v_{j,l} \tag{22.2.12}$$

The analogous inner product between two vectors in function space is[7]

$$\langle \psi_i, \psi_j \rangle = \int_S d\mathbf{r}_s \psi_i(\mathbf{r}_s) \psi_j(\mathbf{r}_s) \tag{22.2.13}$$

where S denotes the cross-sectional area of the waveguide over which the integration is performed. The left-hand side is the shorthand notation for inner product in functional space or the infinite dimensional functional Hilbert space.

Another requirement for a vector in a functional Hilbert space is that it contains finite energy or that

$$\mathscr{E}_f = \int_S d\mathbf{r}_s |\psi_i(\mathbf{r}_s)|^2 \tag{22.2.14}$$

is finite. The above is analogous to that for a matrix vector $\mathbf{v}$ as

$$\mathscr{E}_m = \sum_{l=1}^{N} |v_l|^2 \tag{22.2.15}$$

The square root of the above is often used to denote the "length", the "metric", or the "norm" of the vector. Finite energy also implies that the vectors are of finite length.

22.2.3 Proof of Orthogonality of Waveguide Modes[8]

Because of the aforementioned discussion, we see the similarity between a functional Hilbert space, and the matrix vector space. In order to use the result of the matrix EVP, one key step is to prove that the operator ∇_s^2 is hermitian. In matrix algebra, if the elements of a matrix is explicitly available, a matrix $\overline{\mathbf{A}}$ is hermitian if

$$A_{ij} = A_{ji}^* \tag{22.2.16}$$

[7] In many math books, the conjugation of the first function ψ_i is implied, but here, we follow the electromagnetic convention that the conjugation of ψ_i is not implied unless explicitly stated.

[8] This may be skipped on first reading.

But if a matrix element is not explicitly available, a different way to define hermiticity is required: we will use inner products to define it. A matrix operator is hermitian if

$$\mathbf{x}_i^\dagger \cdot \overline{\mathbf{A}} \cdot \mathbf{x}_j = \left(\mathbf{x}_j^\dagger \cdot \overline{\mathbf{A}}^\dagger \cdot \mathbf{x}_i\right)^\dagger = \left(\mathbf{x}_j^\dagger \cdot \overline{\mathbf{A}} \cdot \mathbf{x}_i\right)^\dagger = \left(\mathbf{x}_j^\dagger \cdot \overline{\mathbf{A}} \cdot \mathbf{x}_i\right)^* \tag{22.2.17}$$

The first equality follows from standard matrix algebra,[9] the second equality follows if $\overline{\mathbf{A}} = \overline{\mathbf{A}}^\dagger$, or that $\overline{\mathbf{A}}$ is hermitian. The last equality follows because the quantity in the parenthesis is a scalar, and hence, its conjugate transpose is just its conjugate. Therefore, if a matrix $\overline{\mathbf{A}}$ is hermitian, then

$$\mathbf{x}_i^\dagger \cdot \overline{\mathbf{A}} \cdot \mathbf{x}_j = \left(\mathbf{x}_j^\dagger \cdot \overline{\mathbf{A}} \cdot \mathbf{x}_i\right)^* \tag{22.2.18}$$

The above inner product can be used to define the hermiticity of the matrix operator $\overline{\mathbf{A}}$. It can be easily extended to define the hermiticity of an operator in an infinite dimensional Hilbert space where the matrix elements are not explicitly available.

Hence, using the inner product definition in (22.2.13) for infinite dimensional functional Hilbert space, a functional operator ∇_s^2 is hermitian if

$$\langle \psi_i^*, \nabla_s^2 \psi_j \rangle = (\langle \psi_j^*, \nabla_s^2 \psi_i \rangle)^* \tag{22.2.19}$$

We can rewrite the left-hand side of the above more explicitly as

$$\langle \psi_i^*, \nabla_s^2 \psi_j \rangle = \int_S d\mathbf{r}_s \psi_i^*(\mathbf{r}_s) \nabla_s^2 \psi_j(\mathbf{r}_s) \tag{22.2.20}$$

and then the right-hand side of (22.2.19) can be rewritten more explicitly as

$$(\langle \psi_j^*, \nabla_s^2 \psi_i \rangle)^* = \int_S d\mathbf{r}_s \psi_j(\mathbf{r}_s) \nabla_s^2 \psi_i^*(\mathbf{r}_s) \tag{22.2.21}$$

To prove the above equality in (22.2.19), one uses the identity that

$$\nabla_s \cdot [\psi_i^*(\mathbf{r}_s) \nabla_s \psi_j(\mathbf{r}_s)] = \psi_i^*(\mathbf{r}_s) \nabla_s^2 \psi_j(\mathbf{r}_s) + \nabla_s \psi_i^*(\mathbf{r}_s) \cdot \nabla_s \psi_j(\mathbf{r}_s) \tag{22.2.22}$$

Integrating the above over the cross sectional area S, and invoking Gauss divergence theorem in 2D, one gets that

$$\begin{aligned} \int_C dl \hat{n} \cdot (\psi_i^*(\mathbf{r}_s) \nabla_s \psi_j(\mathbf{r}_s)) = & \int_S d\mathbf{r}_s \left(\psi_i^*(\mathbf{r}_s) \nabla_s^2 \psi_j(\mathbf{r}_s)\right) \\ & + \int_S d\mathbf{r}_s \left(\nabla \psi_i^*(\mathbf{r}_s) \cdot \nabla_s \psi_j(\mathbf{r}_s)\right) \end{aligned} \tag{22.2.23}$$

where C the the contour bounding S or the waveguide wall. By applying the boundary condition that $\psi_i(\mathbf{r}_s) = 0$ or that $\hat{n} \cdot \nabla_s \psi_j(\mathbf{r}_s) = 0$, or a mixture thereof, then the left-hand side of the above is zero. This will be the case be it TE or TM modes.

$$0 = \int_S d\mathbf{r}_s \left(\psi_i^*(\mathbf{r}_s) \nabla_s^2 \psi_j(\mathbf{r}_s)\right) + \int_S d\mathbf{r}_s \left(\nabla \psi_i^*(\mathbf{r}_s) \cdot \nabla_s \psi_j(\mathbf{r}_s)\right) \tag{22.2.24}$$

[9] $(\overline{\mathbf{A}} \cdot \overline{\mathbf{B}} \cdot \overline{\mathbf{C}})^\dagger = \overline{\mathbf{C}}^\dagger \cdot \overline{\mathbf{B}}^\dagger \cdot \overline{\mathbf{A}}^\dagger$ [77].

Applying the same treatment to (22.2.21), we get

$$0 = \int_S d\mathbf{r}_s \left(\psi_j(\mathbf{r}_s)\nabla_s^2\psi_i^*(\mathbf{r}_s)\right) + \int_S d\mathbf{r}_s \left(\nabla\psi_i^*(\mathbf{r}_s)\cdot\nabla_s\psi_j(\mathbf{r}_s)\right) \tag{22.2.25}$$

The above two equations indicate that

$$\langle\psi_i^*, \nabla_s^2\psi_j\rangle = (\langle\psi_j^*, \nabla_s^2\psi_i\rangle)^* \tag{22.2.26}$$

proving that the operator ∇_s^2 is hermitian. One can then use the above property to prove the orthogonality of the eigenmodes when they have distinct eigenvalues, the same way we have proved the orthogonality of eigenvectors. The above proof can be extended to the case of a resonant cavity. The orthogonality of resonant cavity modes is also analogous to the orthogonality of eigenvectors of a hermitian operator.

Lecture 23

Scalar and Vector Potentials

23.1 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 resonator, 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 scalar and vector potential for solving time-harmonic (electrodynamic) problems.

This is important for bridging the gap between static regime where the frequency is zero or low, and dynamic regime where the frequency is not low. For the dynamic regime, it is important to understand the radiation of electromagnetic fields. Electrodynamic regime is important for studying antennas, communications, sensing, wireless power transfer applications, and many more. Hence, it is imperative that we understand how time-varying electromagnetic fields radiate from sources.

It is also important to understand when static or circuit (quasi-static) regimes are important. The circuit regime solves problems that have fueled the microchip and integrated circuit design (ICD) industry, and it is hence imperative 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).

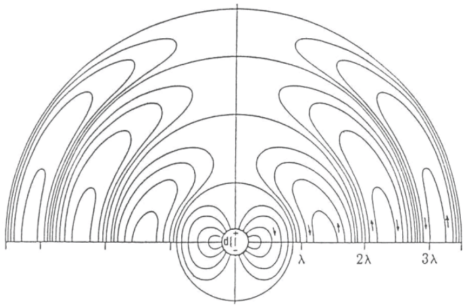


Figure 23.1: Plot of the dynamic electric field around a dipole source, where only the field in the upper half-space is shown. Close to the source, the field ressembles that of a static electric dipole, but far away from the source, the electromagnetic field is detached from the source: the source starts to shed energy to the far region.

23.2 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 medium are then

$$\nabla \times \mathbf{E} = 0 \tag{23.2.1}$$

$$\nabla \times \mathbf{H} = \mathbf{J} \tag{23.2.2}$$

$$\nabla \cdot \varepsilon \mathbf{E} = \varrho \tag{23.2.3}$$

$$\nabla \cdot \mu \mathbf{H} = 0 \tag{23.2.4}$$

Looking at the first equation above, and using the knowledge that $\nabla \times (\nabla \Phi) = 0$, we can construct a solution to (23.2.1) easily. Thus, in order to satisfy the first of Maxwell's equations or Faraday's law above, we let

$$\mathbf{E} = -\nabla \Phi \tag{23.2.5}$$

Using the above in (23.2.3), we get,

$$\nabla \cdot \varepsilon \nabla \Phi = -\varrho \tag{23.2.6}$$

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

$$\nabla^2 \Phi = -\frac{\varrho}{\varepsilon} \tag{23.2.7}$$

which is the Poisson's equation for electrostatics.

Now looking at (23.2.4) where $\nabla \cdot \mu\mathbf{H} = 0$, we let

$$\mu\mathbf{H} = \nabla \times \mathbf{A} \tag{23.2.8}$$

Since $\nabla \cdot (\nabla \times \mathbf{A}) = 0$, the last of Maxwell's equations (23.2.4) is automatically satisfied. Next, using the above in the second of Maxwell's equations above, we get

$$\nabla \times \nabla \times \mathbf{A} = \mu\mathbf{J} \tag{23.2.9}$$

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

$$\nabla^2\mathbf{A} = -\mu\mathbf{J} \tag{23.2.10}$$

which is the vector Poisson's equation. Next, we will repeat the above derivation when $\omega \neq 0$.

23.2.1 Scalar and Vector Potentials for Electrodynamics

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

$$\nabla \times \mathbf{E} = -j\omega\mu\mathbf{H} \tag{23.2.11}$$
$$\nabla \times \mathbf{H} = j\omega\varepsilon\mathbf{E} + \mathbf{J} \tag{23.2.12}$$
$$\nabla \cdot \varepsilon\mathbf{E} = \varrho \tag{23.2.13}$$
$$\nabla \cdot \mu\mathbf{H} = 0 \tag{23.2.14}$$

In order to satisfy the last Maxwell's equation, as before, we let

$$\mu\mathbf{H} = \nabla \times \mathbf{A} \tag{23.2.15}$$

Now, using (23.2.15) in (23.2.11), we have

$$\nabla \times (\mathbf{E} + j\omega\mathbf{A}) = 0 \tag{23.2.16}$$

Since $\nabla \times (\nabla\Phi) = 0$, the above implies that $\mathbf{E} + j\omega\mathbf{A} = -\nabla\Phi$, or

$$\mathbf{E} = -j\omega\mathbf{A} - \nabla\Phi \tag{23.2.17}$$

The above implies 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 second term above, in accordance to Faraday's law, is the contribution to the electric field from the time-varying magnetic field, and hence, is termed the induction term.[1]

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 (23.2.15) and (23.2.17) into (23.2.12) gives

$$\nabla \times \nabla \times \mathbf{A} = -j\omega\mu\varepsilon(-j\omega\mathbf{A} - \nabla\Phi) + \mu\mathbf{J} \tag{23.2.18}$$

[1] Notice that in electrical engineering, most concepts related to magnetic fields are inductive!

Or upon rearrangement, after using that $\nabla \times \nabla \times \mathbf{A} = \nabla\nabla \cdot \mathbf{A} - \nabla \cdot \nabla\mathbf{A}$, we have

$$\nabla^2\mathbf{A} + \omega^2\mu\varepsilon\mathbf{A} = -\mu\mathbf{J} + j\omega\mu\varepsilon\nabla\Phi + \nabla\nabla \cdot \mathbf{A} \tag{23.2.19}$$

Moreover, using (23.2.17) in (23.2.14), we have

$$\nabla \cdot (j\omega\mathbf{A} + \nabla\Phi) = -\frac{\varrho}{\varepsilon} \tag{23.2.20}$$

In the above, (23.2.19) and (23.2.20) 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 $\mathbf{A}$ 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 (23.2.8), 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 [48]. Namely, with

$$\mathbf{A}' = \mathbf{A} + \nabla\Psi \tag{23.2.21}$$

$$\Phi' = \Phi - j\omega\Psi \tag{23.2.22}$$

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, 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 (23.2.15), we need to specify its divergence or $\nabla \cdot \mathbf{A}$ as in the electrostatic case.[2] A clever way to specify the divergence of $\mathbf{A}$ is to choose it to simplify the complicated equations above in (23.2.19). We choose a gauge so that the last two terms in the equation (23.2.19) cancel each other. In other words, we let

$$\nabla \cdot \mathbf{A} = -j\omega\mu\varepsilon\Phi \tag{23.2.23}$$

The above is judiciously chosen so that the pertinent equations (23.2.19) and (23.2.20) will be simplified and decoupled. With the use of (23.2.23) in (23.2.19) and (23.2.20), they now become

$$\nabla^2\mathbf{A} + \omega^2\mu\varepsilon\mathbf{A} = -\mu\mathbf{J} \tag{23.2.24}$$

$$\nabla^2\Phi + \omega^2\mu\varepsilon\Phi = -\frac{\varrho}{\varepsilon} \tag{23.2.25}$$

Equation (23.2.23) is known as the Lorenz gauge[3] and the above equations are Helmholtz equations with source terms. Not only are these equations simplified, they can be solved independently of each other since they are decoupled from each other.

Equations (23.2.24) and (23.2.25) can be solved using the Green's function method we have learnt previously. Equation (23.2.24) actually constitutes three scalar equations for the three x, y, z components, namely that

$$\nabla^2 A_i + \omega^2\mu\varepsilon A_i = -\mu J_i \tag{23.2.26}$$

[2]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 $\mathbf{A}$.

[3]Please note that this Lorenz is not the same as Lorentz.

where i above can be x, y, or z. Therefore, (23.2.24) and (23.2.25) 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

$$\nabla^2 g(\mathbf{r},\mathbf{r}') + \beta^2 g(\mathbf{r},\mathbf{r}') = -\delta(\mathbf{r}-\mathbf{r}') \tag{23.2.27}$$

where $\beta^2 = \omega^2\mu\varepsilon$.

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

$$g(\mathbf{r},\mathbf{r}') = g(|\mathbf{r}-\mathbf{r}'|) = \frac{e^{-j\beta|\mathbf{r}-\mathbf{r}'|}}{4\pi|\mathbf{r}-\mathbf{r}'|} \tag{23.2.28}$$

which can be verified by back substitution.

By using the principle of linear superposition, or convolution, the solutions to (23.2.24) and (23.2.25) are then

$$\mathbf{A}(\mathbf{r}) = \mu \iiint d\mathbf{r}'\mathbf{J}(\mathbf{r}')g(|\mathbf{r}-\mathbf{r}'|) = \mu \iiint d\mathbf{r}'\mathbf{J}(\mathbf{r}')\frac{e^{-j\beta|\mathbf{r}-\mathbf{r}'|}}{4\pi|\mathbf{r}-\mathbf{r}'|} \tag{23.2.29}$$

$$\Phi(\mathbf{r}) = \frac{1}{\varepsilon}\iiint d\mathbf{r}'\varrho(\mathbf{r}')g(|\mathbf{r}-\mathbf{r}'|) = \frac{1}{\varepsilon}\iiint d\mathbf{r}'\varrho(\mathbf{r}')\frac{e^{-j\beta|\mathbf{r}-\mathbf{r}'|}}{4\pi|\mathbf{r}-\mathbf{r}'|} \tag{23.2.30}$$

In the above $d\mathbf{r}'$ is the shorthand notation for $dx\,dy\,dz$ and hence, they are still volume integrals. The above are three-dimensional convolutional integrals in space.

23.2.2 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.[4] Thus the equations then become

$$\nabla \times \mathbf{E} = -j\omega\mu\mathbf{H} - \mathbf{M} \tag{23.2.31}$$

$$\nabla \times \mathbf{H} = j\omega\varepsilon\mathbf{E} + \mathbf{J} \tag{23.2.32}$$

$$\nabla \cdot \mu\mathbf{H} = \varrho_m \tag{23.2.33}$$

$$\nabla \cdot \varepsilon\mathbf{E} = \varrho \tag{23.2.34}$$

The above can be solved in two stages, using the principle of linear superposition. 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. Second, we can set $\mathbf{J} = 0$, $\varrho = 0$ and $\mathbf{M} \neq 0$, $\varrho_m \neq 0$ and solve for the fields next. Then the total general solution, by linearity, is just the linear superposition of these two solutions.

[4] In fact, Maxwell exploited this symmetry [38].

For the second case, we can define an electric vector potential $\mathbf{F}$ such that

$$\mathbf{D} = -\nabla \times \mathbf{F} \tag{23.2.35}$$

and a magnetic scalar potential Φ_m such that

$$\mathbf{H} = -\nabla \Phi_m - j\omega \mathbf{F} \tag{23.2.36}$$

By invoking duality principle, one gather that [47]

$$\mathbf{F}(\mathbf{r}) = \varepsilon \iiint d\mathbf{r}' \mathbf{M}(\mathbf{r}') g(|\mathbf{r} - \mathbf{r}'|) = \varepsilon \iiint d\mathbf{r}' \mathbf{M}(\mathbf{r}') \frac{e^{-j\beta|\mathbf{r}-\mathbf{r}'|}}{4\pi|\mathbf{r} - \mathbf{r}'|} \tag{23.2.37}$$

$$\Phi_m(\mathbf{r}) = \frac{1}{\mu} \iiint d\mathbf{r}' \varrho_m(\mathbf{r}') g(|\mathbf{r} - \mathbf{r}'|) = \frac{1}{\mu} \iiint d\mathbf{r}' \varrho_m(\mathbf{r}') \frac{e^{-j\beta|\mathbf{r}-\mathbf{r}'|}}{4\pi|\mathbf{r} - \mathbf{r}'|} \tag{23.2.38}$$

As mentioned before, even though magnetic sources do not exist, they can be engineered. In many engineering designs, one can use fictitious magnetic sources to enrich the diversity of electromagnetic technologies.

23.3 When is Static Electromagnetic Theory Valid?

We have learnt in the previous section that for electrodynamics,

$$\mathbf{E} = -\nabla \Phi - j\omega \mathbf{A} \tag{23.3.1}$$

where the second term above on the right-hand side is due to induction, or the contribution to the electric field from the time-varying magnetic filed. Hence, much things we learn in potential theory that $\mathbf{E} = -\nabla\Phi$ is not exactly valid. But simple potential theory that $\mathbf{E} = -\nabla\Phi$ is very useful because of its simplicity. We will study when static electromagnetic theory can be used to model electromagnetic systems.

Since the third and the fourth Maxwell's equations are derivable from the first two, let us first study when we can ignore the time derivative terms in the first two of Maxwell's equations, which, in the frequency domain, are

$$\nabla \times \mathbf{E} = -j\omega\mu \mathbf{H} \tag{23.3.2}$$

$$\nabla \times \mathbf{H} = j\omega\varepsilon \mathbf{E} + \mathbf{J} \tag{23.3.3}$$

When the terms multiplied by $j\omega$ above, which are associated with time derivatives, can be ignored, then electrodynamics can be replaced with static electromagnetics, which are much simpler.[5] Quasi-static electromagnetic theory eventually gave rise to circuit theory and telegraphy technology. Circuit theory consists of elements like resistors, capacitors, and inductors. Given that we have now seen electromagnetic theory in its full form, we would like to ponder when we can use simple static electromagnetics to describe electromagnetic phenomena.

[5] That is why Ampere's law, Coulomb's law, and Gauss' law were discovered first.

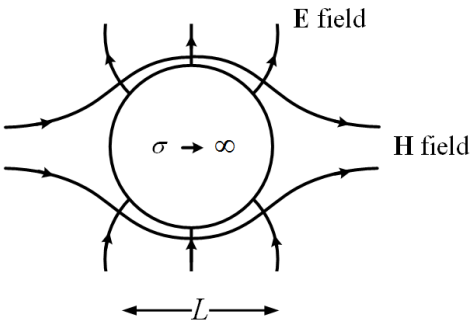


Figure 23.2: The electric and magnetic fields are great contortionists around a perfectly conducting particle. They deform themselves to satisfy the boundary conditions, $\hat{n} \times \mathbf{E} = 0$, and $\hat{n} \cdot \mathbf{H} = 0$ on the PEC surface, even when the particle is very small. In other words, the fields vary on the length-scale of $1/L$. Hence, $\nabla \sim 1/L$ which is large when L is small.

To see this lucidly, it is best to write Maxwell's equations in dimensionless units or the same units. Say if we want to solve Maxwell's equations for the fields close to an object of size L as shown in Figure 23.2. This object can be a small particle like the sphere, or it could be a capacitor, or an inductor, which are small; but how small should it be before we can apply static electromagnetics?

It is clear that these $\mathbf{E}$ and $\mathbf{H}$ fields will have to satisfy boundary conditions, $\hat{n} \times \mathbf{E} = 0$, and $\hat{n} \cdot \mathbf{H} = 0$ on the PEC surface, which is **de rigueur** in the vicinity of the object as shown in Figure 23.2 even when the frequency is low or the wavelength long. The fields become great contortionists in order to do so. Hence, we do not expect a constant field around the object but that the field will vary on the length scale of L. So we renormalize our length scale by this length L by defining a new dimensionless coordinate system such that[6]

$$x' = \frac{x}{L}, \quad y' = \frac{y}{L}, \quad z' = \frac{z}{L} \tag{23.3.4}$$

In other words, by so doing, then $Ldx' = dx$, $Ldy' = dy$, and $Ldz' = dz$, and

$$\frac{\partial}{\partial x} = \frac{1}{L}\frac{\partial}{\partial x'}, \quad \frac{\partial}{\partial y} = \frac{1}{L}\frac{\partial}{\partial y'}, \quad \frac{\partial}{\partial z} = \frac{1}{L}\frac{\partial}{\partial z'} \tag{23.3.5}$$

Take for example a function $f(x) = \sin(\pi x/L)$ which is a periodic function varying on the length scale of L having a period of L. Then it is quite clear that $df(x)/dx = (\pi/L)\cos(\pi x/L)$. When L is small representing a rapidly varying function, $df(x)/dx \sim O(1/L)$.[7] But in the new variable, $f(Lx') = \sin(\pi x')$ and $df/dx' = \cos(\pi x')$ which is $O(1)$.

[6]This dimensional analysis is often used by fluid dynamicists to study fluid flow problems [132].

[7]Reads order of $1/L$ of order $1/L$.

In this manner, $\nabla = \frac{1}{L}\nabla'$ where ∇' is dimensionless; or ∇ will be very large when it operates on fields that vary on the length scale of very small L, where ∇', which is an $O(1)$ operator, will not be large because it is in coordinates normalized with respect to L.

Then, the first two of Maxwell's equations become

$$\frac{1}{L}\nabla' \times \mathbf{E} = -j\omega\mu_0\mathbf{H} \tag{23.3.6}$$

$$\frac{1}{L}\nabla' \times \mathbf{H} = j\omega\varepsilon_0\mathbf{E} + \mathbf{J} \tag{23.3.7}$$

Here, we still have apples and oranges to compare with since $\mathbf{E}$ and $\mathbf{H}$ have different units; we cannot compare quantities if they have different units. For instance, the ratio of $\mathbf{E}$ to the $\mathbf{H}$ field has a dimension of impedance. To bring them to the same unit, we define a new $\mathbf{E}'$ such that

$$\eta_0\mathbf{E}' = \mathbf{E} \tag{23.3.8}$$

where $\eta_0 = \sqrt{\mu_0/\varepsilon_0} \cong 377$ ohms in vacuum. In this manner, the new $\mathbf{E}'$ has the same unit as the $\mathbf{H}$ field. Then, (23.3.6) and (23.3.7) become

$$\frac{\eta_0}{L}\nabla' \times \mathbf{E}' = -j\omega\mu_0\mathbf{H} \tag{23.3.9}$$

$$\frac{1}{L}\nabla' \times \mathbf{H} = j\omega\varepsilon_0\eta_0\mathbf{E}' + \mathbf{J} \tag{23.3.10}$$

With this change, the above can be rearranged to become

$$\nabla' \times \mathbf{E}' = -j\omega\mu_0\frac{L}{\eta_0}\mathbf{H} \tag{23.3.11}$$

$$\nabla' \times \mathbf{H} = j\omega\varepsilon_0\eta_0 L\mathbf{E}' + L\mathbf{J} \tag{23.3.12}$$

By letting $\eta_0 = \sqrt{\mu_0/\varepsilon_0}$, the above can be further simplified to become

$$\nabla' \times \mathbf{E}' = -j\frac{\omega}{c_0}L\mathbf{H} \tag{23.3.13}$$

$$\nabla' \times \mathbf{H} = j\frac{\omega}{c_0}L\mathbf{E}' + L\mathbf{J} \tag{23.3.14}$$

Notice now that in the above, $\mathbf{H}$, $\mathbf{E}'$, and $L\mathbf{J}$ have the same unit, and ∇' is dimensionless and is of order one, and $\omega L/c_0$ is also dimensionless.

Therefore, in the above, one can compare terms, and ignore the frequency dependent $j\omega$ term when

$$\frac{\omega}{c_0}L \ll 1 \tag{23.3.15}$$

The above is the same as $kL \ll 1$ where $k = \omega/c_0 = 2\pi/\lambda_0$. Thus, when

$$2\pi\frac{L}{\lambda_0} \ll 1 \tag{23.3.16}$$

the $j\omega$ terms can be ignored and the first two Maxwell's equations become static equations. Consequently, the above criteria are for the validity of the static approximation when the time-derivative terms in Maxwell's equations can be ignored. When these criteria are satisfied, then Maxwell's equations can be simplified to and approximated by the following equations

$$\nabla' \times \mathbf{E}' = 0 \tag{23.3.17}$$

$$\nabla' \times \mathbf{H} = L\mathbf{J} \tag{23.3.18}$$

which are the static equations, Faraday's law and Ampere's law of electromagnetic theory. They can be solved together with Gauss' laws, or the third or the fourth Maxwell's equations.

In other words, 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. Figure 23.3 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 external field can be written as

$$\mathbf{E}_s = (\hat{r}2\cos\theta + \hat{\theta}\sin\theta)\left(\frac{a}{r}\right)^3 E_s \tag{23.3.19}$$

while the incident field $\mathbf{E}_0$ and the interior field $\mathbf{E}_i$ to the particle can be expressed as

$$\mathbf{E}_0 = \hat{z}E_0 = (\hat{r}\cos\theta - \hat{\theta}\sin\theta)E_0 \tag{23.3.20}$$

$$\mathbf{E}_i = \hat{z}E_i = (\hat{r}\cos\theta - \hat{\theta}\sin\theta)E_i \tag{23.3.21}$$

By matching boundary conditions, as was done in the homework, it can be shown that

$$E_s = \frac{\varepsilon_s - \varepsilon}{\varepsilon_s + 2\varepsilon}E_0 \tag{23.3.22}$$

$$E_i = \frac{3\varepsilon}{\varepsilon_s + 2\varepsilon}E_0 \tag{23.3.23}$$

For a plasmonic nano-particle, the particle medium behaves like a plasma (see Lecture 8), and ε_s in the above can be negative, making the denominators of the above expression very close to zero. 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!

Figure 23.4 shows a nano-particle induced in plasmonic oscillation by a light wave. Figure 23.5 shows that different color fluids can be obtained by immersing nano-particles in fluids with different background permittivity (ε in (23.3.22) and (23.3.23)) causing the plasmonic particles to resonate at different frequencies.

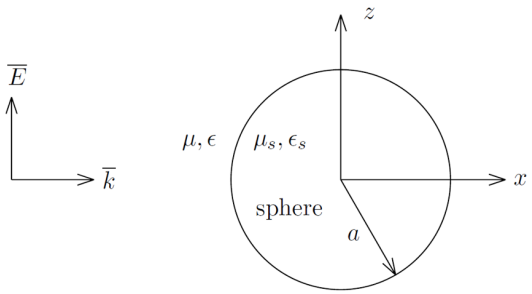


Figure 23.3: A plane electromagnetic wave incident on a particle. When the particle size is small compared to wavelength, static analysis can be used to solve this problem (courtesy of Kong [32]).

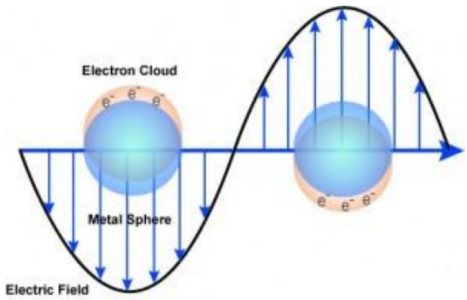


Figure 23.4: A nano-particle undergoing electromagnetic oscillation when an electromagnetic wave impinges on it (courtesy of sigmaaldric.com). The oscillation is inordinately large when the incident wave's frequency coincides with the resonance frequency of the plasmonic particle.

Figure 23.5: Different color fluids containing nano-particles can be obtained by changing the permittivity of the background fluids (courtesy of nanocomposix.com).

In (23.3.16), this criterion has been expressed in terms of the dimension of the object L compared to the wavelength λ_0. Alternatively, we can express this criterion in terms of transit time. The transit time for an electromagnetic wave to traverse an object of size L is $\tau = L/c_0$ and $\omega = 2\pi/T$ where T is the period of one time-harmonic oscillation. Therefore, (23.3.15) can be re-expressed as

$$\frac{2\pi\tau}{T} \ll 1 \tag{23.3.24}$$

The above implies that if the transit time τ needed for the electromagnetic field to traverse the object of length L is much small than the period of oscillation of the electromagnetic field, then static theory can be used.

The finite speed of light gives rise to delay or retardation of electromagnetic signal when it propagates through space. When this retardation effect can be ignored, then static electromagnetic theory can be used. In other words, if the speed of light had been infinite, then there would be no retardation effect, and static electromagnetic theory could always be used. Alternatively, the infinite speed of light will give rise to infinite wavelength, and criterion (23.3.16) will always be satisfied, and static theory prevails always.

23.3.1 Quasi-Static Electromagnetic Theory

In closing, we would like to make one more remark. The right-hand side of (23.3.11), which is 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 inductor in this world. To understand this dilemma, let us rewrite (23.3.11) in integral form, namely,

$$\oint_C \mathbf{E}' \cdot d\mathbf{l} = -j\omega\mu_0 \frac{L}{\eta_0} \iint_S d\mathbf{S} \cdot \mathbf{H} \tag{23.3.25}$$

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 circuit theory is not diminished unless $\omega = 0$.

By the same token, displacement current in (23.3.11) 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

$$\oint_C \mathbf{H} \cdot d\mathbf{l} = j\omega\varepsilon\eta_0 L \iint_S d\mathbf{S} \cdot \mathbf{E}' \tag{23.3.26}$$

The displacement 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 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. That explains the success of circuit theory in electromagnetic engineering!

Lecture 24

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 physics of highly complex circuits and structures, which is very important in the computer and micro-chip industry (or the IC design industry). 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 [31, 32, 54, 65].

The two most important laws in circuit theory are Kirchoff current law (KCL) and Kirchhoff voltage law (KVL) [14, 52]. These two laws are derivable from the current continuity equation and from Faraday's law.

24.1 Kirchhoff Current Law

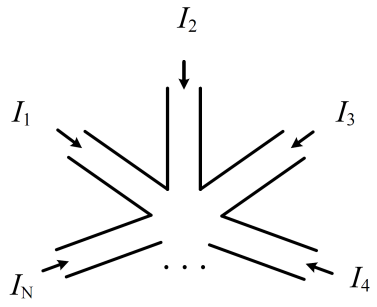


Figure 24.1: Schematic showing the derivation of Kirchhoff current law. All currents flowing into a node must add up to zero.

Kirchhoff current law (KCL) is a consequence of the current continuity equation, or that

$$\nabla \cdot \mathbf{J} = -j\omega\varrho \tag{24.1.1}$$

It is a consequence of charge conservation. But it is also derivable from generalized Ampere's law and Gauss' law for charge.[1]

First, we assume that all currents are flowing into a node as shown in Figure 24.1, and that the node is non-charge accumulating with $\omega \to 0$. Then the charge continuity equation becomes[2]

$$\nabla \cdot \mathbf{J} = 0 \tag{24.1.2}$$

By integrating the above current continuity equation over a volume containing the node, it is easy to show that

$$\sum_{i}^{N} I_i = 0 \tag{24.1.3}$$

which is the statement of KCL. This is shown for the schematic of Figure 24.1.

24.2 Kirchhoff Voltage Law

Kirchhoff voltage law is the consequence of Faraday's law. For the truly static case when $\omega = 0$, it is

$$\nabla \times \mathbf{E} = 0 \tag{24.2.1}$$

The above implies that $\mathbf{E} = -\nabla\Phi$, from which we can deduce that

$$-\oint_C \mathbf{E} \cdot \mathbf{dl} = 0 \tag{24.2.2}$$

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

$$V_{ba} = -\int_a^b \mathbf{E} \cdot \mathbf{dl} = \int_a^b \nabla\Phi \cdot \mathbf{dl} = \Phi(\mathbf{r}_b) - \Phi(\mathbf{r}_a) = V_b - V_a \tag{24.2.3}$$

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

$$\int_a^b \frac{d}{dx}\Phi \cdot \mathbf{dx} = \Phi(\mathbf{r}_b) - \Phi(\mathbf{r}_a) \tag{24.2.4}$$

[1] Some authors will say that charge conservation is more fundamental, and that Gauss' law and Ampere's law are consistent with charge conservation and the current continuity equation.

[2] One can also say that this is a consequenc of static Ampere's law, $\nabla \times \mathbf{H} = \mathbf{J}$. By taking the divergence of this equation yields (24.1.2) directly.

A curved line can be thought of as a concatenation of many small straight line segments.

As has been shown before, to be exact, $\mathbf{E} = -\nabla\Phi - \partial/\partial t\mathbf{A}$, but we have ignored the induction effect. Therefore, this concept is only valid in the low frequency or long wavelength limit, or that the dimension over which the above is applied is very small so that retardation effect can be ignored.

A good way to remember the above formula is that if $V_b > V_a$. 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 24.2 gives Kirchhoff voltage law (KVL), or that

$$\sum_i^N V_j = 0 \tag{24.2.5}$$

Notice that the voltage drop across a resistor is always positive, since the voltages to the left of the resistors in Figure 24.2 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.

If one of the voltage drops is due to a voltage source, it can be modeled by a negative resistor as shown in Figure 24.3. 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.

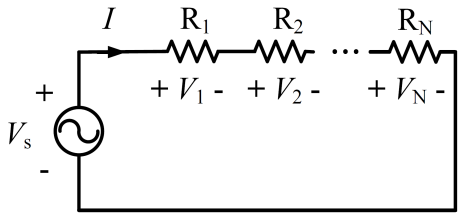


Figure 24.2: Kichhoff voltage law where the sum of all voltages around a loop is zero, which is the consequence of static Faraday's law.

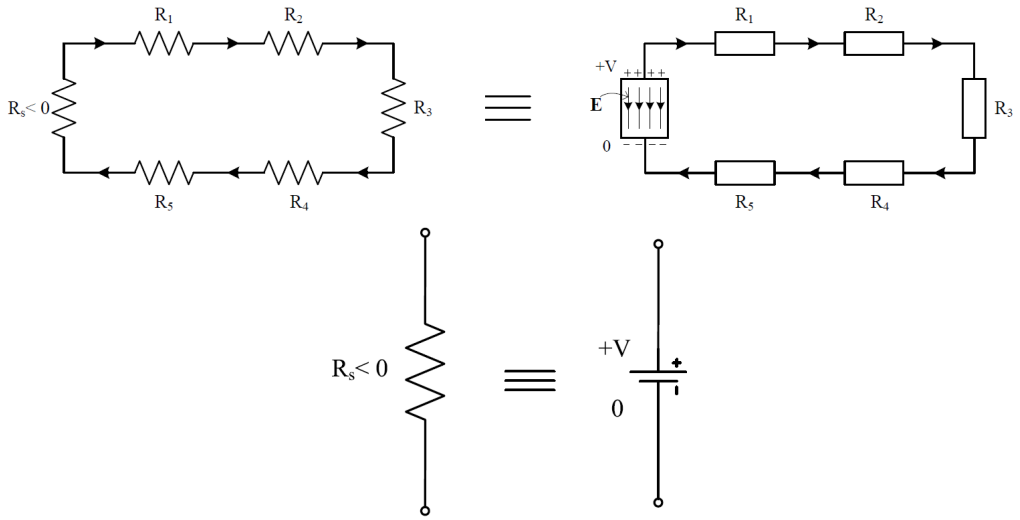


Figure 24.3: A voltage source can also be modeled by a negative resistor.

Faraday's law for the time-varying **B** flux case is

$$\nabla \times \mathbf{E} = -\frac{\partial \mathbf{B}}{\partial t} \tag{24.2.6}$$

Writing the above in integral form, one gets

$$-\oint_C \mathbf{E} \cdot \mathbf{dl} = \frac{d}{dt} \int_s \mathbf{B} \cdot \mathbf{dS} \tag{24.2.7}$$

We can apply the above to a loop shown in Figure 24.4, or a loop C that goes from a to b to c to d to a. We can further assume that this loop is very small compared to wavelength so that potential theory that $\mathbf{E} = -\nabla\Phi$ can be applied. Furthermore, we assume that this loop C does not have any magnetic flux through it so that the right-hand side of the above can be set to zero, or

$$-\oint_C \mathbf{E} \cdot \mathbf{dl} = 0 \tag{24.2.8}$$

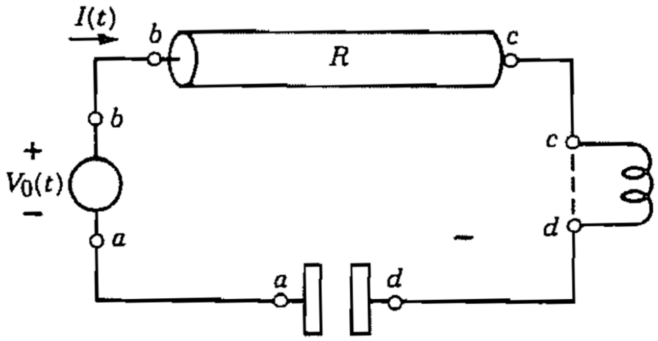


Figure 24.4: The Kirchhoff voltage law for a circuit loop consisting of resistor, inductor, and capacitor can also be derived from Faraday's law at low frequency (courtesy of Ramo et al).

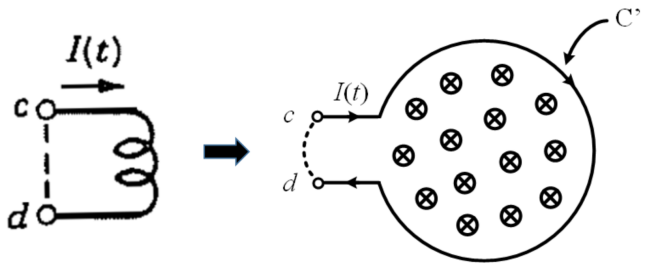


Figure 24.5: The voltage-current relation of an inductor can be obtained by unwrapping an inductor coil, and then calculating its flux linkage.

Notice that this loop does not go through the inductor, but goes directly from c to d. Then there is no flux linkage in this loop and thus

$$-\int_a^b \mathbf{E}\cdot\mathbf{dl} - \int_b^c \mathbf{E}\cdot\mathbf{dl} - \int_c^d \mathbf{E}\cdot\mathbf{dl} - \int_d^a \mathbf{E}\cdot\mathbf{dl} = 0 \tag{24.2.9}$$

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 (24.2.3), we have

$$V_0(t) + V_{cb} + V_{dc} + V_{ad} = 0 \tag{24.2.10}$$

Notice that in the above, in accordance to (24.2.3), $V_b > V_c$, $V_c > V_d$, and $V_a > V_a$. 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 carefully next.

24.3 Inductor

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 24.5 where 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, and hence, $\mathbf{E} = -\nabla\Phi$. 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)[3] 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 to the inductor. But this new loop will contain the flux $\mathbf{B}$ generated by the inductor current. Thus

$$\oint_{C'} \mathbf{E}\cdot\mathbf{dl} = \int_d^c \mathbf{E}\cdot\mathbf{dl} = -V_{dc} = -\frac{d}{dt}\int_{S'} \mathbf{B}\cdot\mathbf{dS} \tag{24.3.1}$$

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. The inductance L is defined as the flux linkage per unit current, or

$$L = \left[\int_{S'} \mathbf{B}\cdot\mathbf{dS}\right]/I \tag{24.3.2}$$

So the voltage in (24.3.1) is then

$$V_{dc} = \frac{d}{dt}(LI) = L\frac{dI}{dt} \tag{24.3.3}$$

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

$$\oint \mathbf{E}\cdot\mathbf{dl} = R_L I - V_{dc} = -\frac{d}{dt}\int_S \mathbf{B}\cdot\mathbf{dS} \tag{24.3.4}$$

Consequently,

$$V_{dc} = R_L I + L\frac{dI}{dt} \tag{24.3.5}$$

Thus, to account for the loss of the coil, we add a resistor in the equation. The above becomes simpler in the frequency domain, namely

$$V_{dc} = R_L I + j\omega L I \tag{24.3.6}$$

[3] Lenz's law can also be explained from Faraday's law (1831).

24.4 Capacitance

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

$$C = \frac{Q}{V} \tag{24.4.1}$$

From the current continuity equation, one can easily show that in Figure 24.6,

$$I = \frac{dQ}{dt} = \frac{d}{dt}(CV_{da}) = C\frac{dV_{da}}{dt} \tag{24.4.2}$$

where C is time independent. Integrating the above equation, one gets

$$V_{da}(t) = \frac{1}{C}\int_{-\infty}^{t} I dt' \tag{24.4.3}$$

The above looks quite cumbersome in the time domain, but in the frequency domain, it becomes

$$I = j\omega C V_{da} \tag{24.4.4}$$

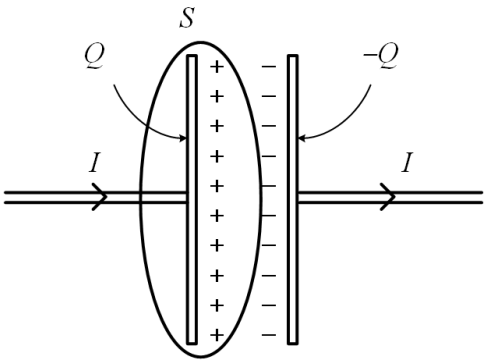


Figure 24.6: Schematic showing the calculation of the capacitance of a capacitor.

24.5 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 dominates. As is well known,

$$\mathbf{J} = \sigma \mathbf{E} \tag{24.5.1}$$

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

$$V_{cb} = -\int_b^c \mathbf{E} \cdot \mathbf{dl} = -\int_b^c \frac{\mathbf{J}}{\sigma} \cdot \mathbf{dl} \tag{24.5.2}$$

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, $\mathbf{J} \cdot \mathbf{d}l = Idl/A$, implying that

$$V_{cb} = -\int_b^c \frac{Idl}{\sigma A} = -I\int_b^c \frac{dl}{\sigma A} = -IR \tag{24.5.3}$$

where[4]

$$R = \int_b^c \frac{dl}{\sigma A} = \int_b^c \frac{\rho dl}{A} \tag{24.5.4}$$

Again, for simplicity, we assume long wavelength or low frequency in the above derivation.

24.6 Some Remarks

In this course, we have learnt that given the sources ϱ and $\mathbf{J}$ of an electromagnetic system, one can find Φ and $\mathbf{A}$, from which we can find $\mathbf{E}$ and $\mathbf{H}$. This is even true at DC or statics. We have also looked at the definition of inductor L and capacitor C. 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 lost 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 amplified by capacitor and inductor, 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. Since the wave is slow, it has a smaller wavelength, and 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 generally have higher losses than air or vacuum.

24.7 Energy Storage Method for Inductor and Capacitor

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

[4]The resisitivity $\rho = 1/\sigma$ where ρ has the unit of ohm-m, while σ has the unit of siemen/m.

easily. This manner of thinking allows for an alternative way of calculating inductances and capacitances as well [31].

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

$$W_m = \frac{1}{2}LI^2 \tag{24.7.1}$$

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, a piece of wire in fact behaves like a small inductor, and it 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

$$W_e = \frac{1}{2}CV^2 \tag{24.7.2}$$

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.

24.8 Finding Closed-Form Formulas for Inductance and Capacitance

Finding closed form solutions for inductors and capacitors is a difficult endeavor. As in solving Maxwell's equations or the waveguide proble, only certain 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

$$\mathbf{H}(\mathbf{r}) = \int \frac{I(\mathbf{r}')\mathbf{dl}' \times \hat{R}}{4\pi R^2} \tag{24.8.1}$$

But the above cannot be evaluated in closed form save in terms of complicate elliptic integrals [117, 133]. Thus it is simpler to just measure the inductance.

However, if we have a solenoid as shown in Figure 24.7, 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 [29, 31]. Figure 24.8 shows the schematic used to find the approximate inductance of this inductor.

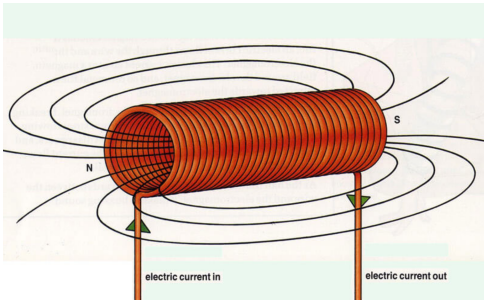


Figure 24.7: The flux-linkage method is used to estimate the inductor of a solenoid (courtesy of SolenoidSupplier.Com).

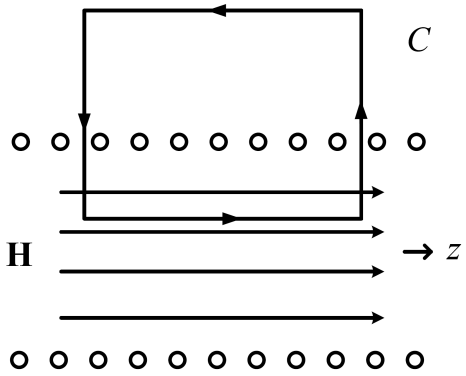


Figure 24.8: Finding the inductor flux linkage approximately by assuming the magnetic field is uniform inside a long solenoid.

The capacitance of a parallel plate capacitor can be found by solving a boundary value problem (BVP) for electrostatics as shown in Section 3.3.4. The electrostatic BVP for capacitor involves Poisson's equation and Laplace equation which are scalar equations [48]. Alternatively, variational expressions can be used to find the lower and upper bounds of capacitors using, for example, Thomson's theorem [48].

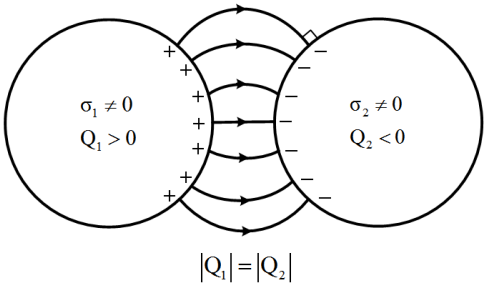


Figure 24.9: The capacitance between two charged conductors can be found by solving a boundary value problem (BVP) involving Laplace equation as discussed in 3.3.4.

Assume a geometry of two conductors charged to $+V$ and $-V$ volts as shown in Figure 24.9. 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

$$\Phi(\mathbf{r}) = \frac{1}{\varepsilon}\int_S dS' \frac{\sigma(\mathbf{r}')}{4\pi|\mathbf{r}-\mathbf{r}'|} \tag{24.8.2}$$

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

$$\Phi(\mathbf{r}) = \frac{1}{\varepsilon}\int_S dS' \frac{\sigma(\mathbf{r}')}{4\pi|\mathbf{r}-\mathbf{r}'|} = \begin{cases} +V, & \mathbf{r} \in S_1 \\ -V, & \mathbf{r} \in S_2 \end{cases} \tag{24.8.3}$$

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. The integral equation can be solved by numerical methods.

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 $C = Q/(2V)$.

24.9 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 using circuit theory.

Figure 24.10 shows the schematic and the cross section of a computer chip at different levels: with the transistor level at the bottom-most. 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. Since these nanometer-size structures are much smaller than the wavelength, 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. This is important at the package level where the dimensions of the components are larger.

A process of parameter extraction where computer software or field solvers (software that solve Maxwell's equations numerically) 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 commercial software called SPICE (Simulation Program with Integrated-Circuit Emphasis) [134], which is a computer-aided software, solves for the voltages and currents in this network.

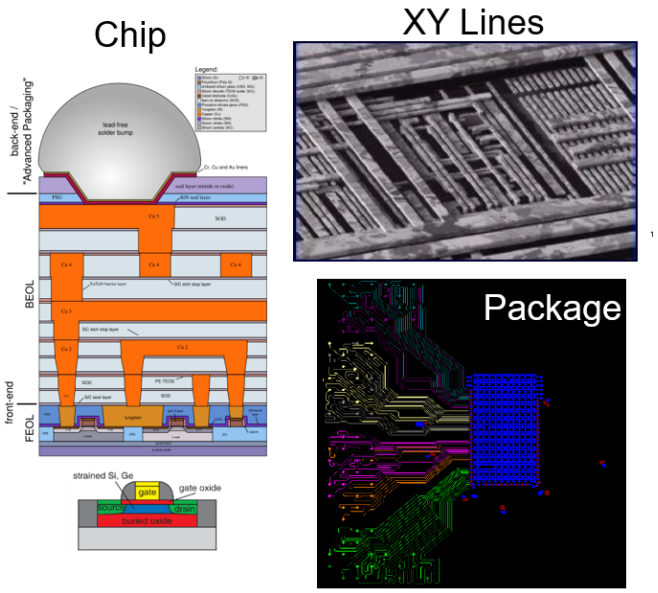


Figure 24.10: Cross section of a chip (top left) and the XY lines in the chip (top right), and the interconnects in the package needed to take the signal out of the chip (bottom right) (courtesy of Wikipedia and Intel).

Initially, SPICE software was written primarily to solve circuit problems. But the SPICE software now has many capabilities, including modeling of transmission lines for microwave engineering, which are important for modeling retardation effects. Figure 24.11 shows an interface of an RF-SPICE that allows the modeling of transmission line with a Smith chart interface.

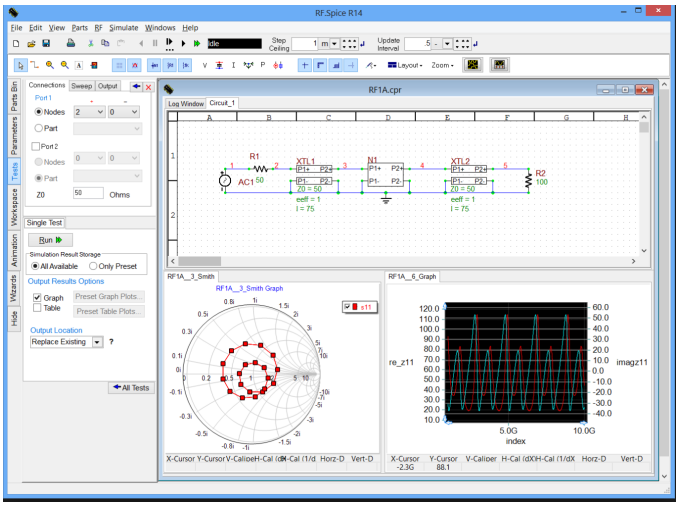


Figure 24.11: SPICE is also used to solve RF problems. A transmission line is used in combination with circuit theory to account for retardation effects in a computer circuit (courtesy of EMAG Technologies Inc.).

24.10 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 of the chip as they can remove high-frequency noise and voltage fluctuation from a circuit as shown in Figure 24.12. Figure 24.13 shows a 3D IC computer chip where decoupling capacitors are integrated into its design.

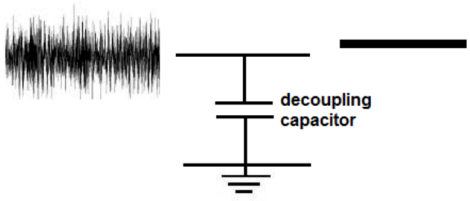


Figure 24.12: A decoupling capacitor is essentially a low-pass filter allowing low-frequency signal to pass through, while high-frequency signal is short-circuited (courtesy learningabout-electronics.com).

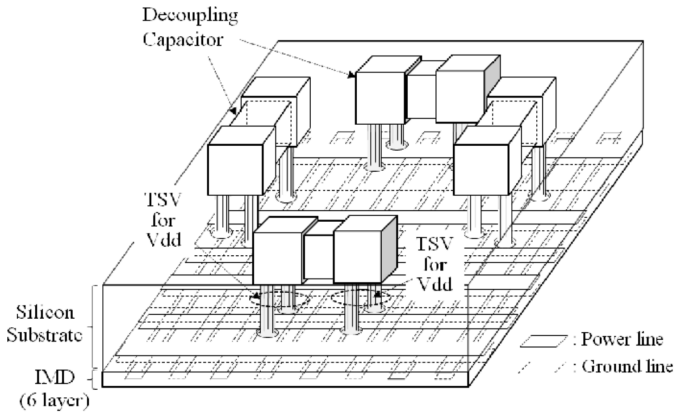


Figure 24.13: Modern computer chip design is 3D and is like a jungle. There are different levels in the chip and they are connected by through silicon vias (TSV). IMD stands for inter-metal dielectrics. One can see different XY lines serving as power and ground lines (courtesy of Semantic Scholars).

Inductors are also indispensable in IC design, as they can be used as a high frequency choke. However, designing compact inductor 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 [135].

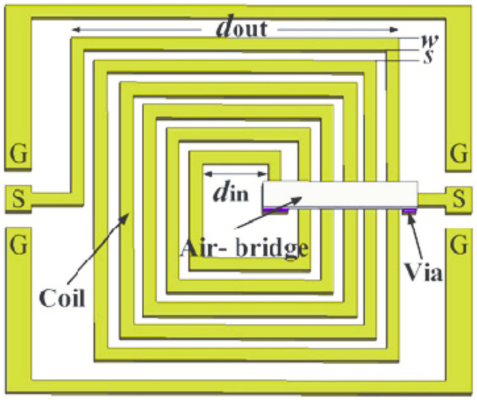


Figure 24.14: Spiral inductors are difficult to build on a chip, but by using laminal structure, it can be integrated into the IC fabrication process (courtesy of Quan Yuan, Research Gate).

Lecture 25

Radiation by a Hertzian Dipole

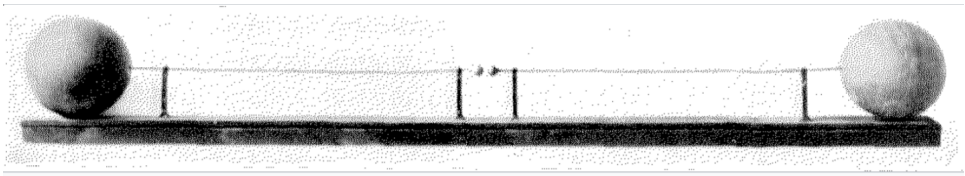
Radiation of electromagnetic field is of ultimate importance for wireless communication systems. The first demonstration of the wave nature of electromagnetic field was by Heinrich Hertz in 1888 [18], some 23 years after Maxwell's equations were fully established. Guglielmo Marconi, after much perserverence with a series of experiments, successfully transmitted wireless radio signal from Cornwall, England to Newfoundland, Canada in 1901 [136]. 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 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.

25.1 History

The original historic Hertzian dipole experiment is shown in Figure 25.1. It was done in 1887 by Heinrich Hertz [18]. The schematic for the original experiment is also shown in Figure 25.2.

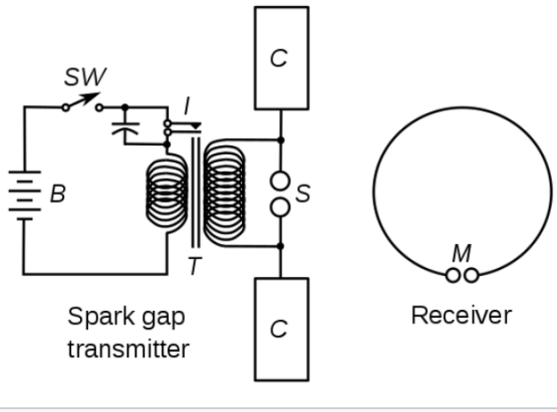
A metallic sphere has a capacitance in closed form with respect to infinity or a ground plane.[1] Hertz could use those knowledge to estimate the capacitance of the sphere, and also, he could estimate the inductance of the leads that are attached to the dipole, and hence, the resonance frequency of his antenna. 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.

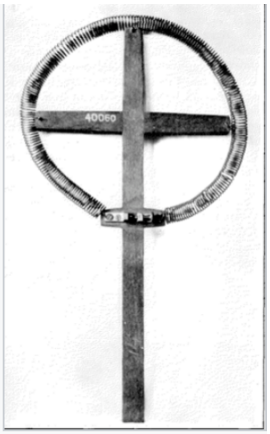
[1] We shall learn later that this problem can be solved in closed form using image theorem.



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 25.1: Hertz's original experiment on a small dipole (courtesy of Wikipedia [18]).





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 25.2: More on Hertz's original experiment on a small dipole (courtesy of Wikipedia [18]). The antenna was powered by a transformer. The radiated electromagnetic field was picked up by a loop receiver antenna that generates a spark at its gap M.

25.2 Approximation by a Point Source

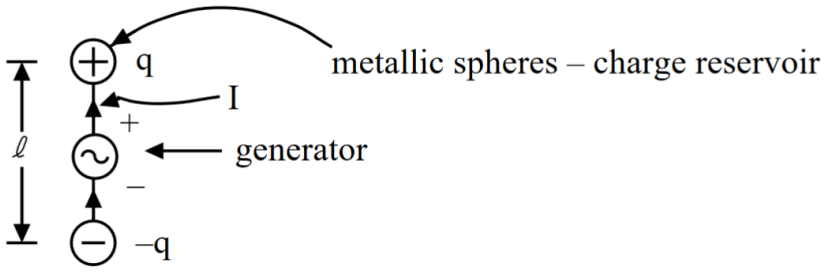


Figure 25.3: Schematic of a small Hertzian dipole which is a close approximation of that first proposed by Hertz.

Figure 25.3 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 in time harmonically because it is driven by the generator. Since

$$\frac{dq}{dt} = I,$$

we have the current moment

$$Il = \frac{dq}{dt} l = j\omega ql = j\omega p \tag{25.2.1}$$

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 [32, 44]

$$\mathbf{J}(\mathbf{r}) = \hat{z} Il\delta(x)\delta(y)\delta(z) = \hat{z} Il\delta(\mathbf{r}) \tag{25.2.2}$$

The dipole is as shown in Figure 25.3 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.

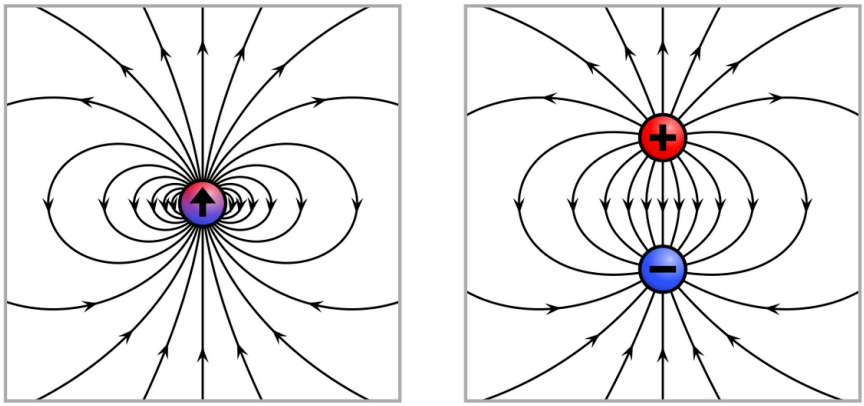


Figure 25.4: The field of a point dipole field versus that of a dipole field. When one is far away from the dipole sources, their fields are similar to each other (courtesy of Wikepedia).

We have learnt previously that the vector potential is related to the current as follows:

$$\mathbf{A}(\mathbf{r}) = \mu \iiint d\mathbf{r}' \mathbf{J}(\mathbf{r}') \frac{e^{-j\beta|\mathbf{r}-\mathbf{r}'|}}{4\pi|\mathbf{r}-\mathbf{r}'|} \tag{25.2.3}$$

Since the current is a 3D delta function in space, using the sifting property of a delta function, the corresponding vector potential is given by

$$\mathbf{A}(\mathbf{r}) = \hat{z} \frac{\mu I l}{4\pi r} e^{-j\beta r} \tag{25.2.4}$$

Since the vector potential $\mathbf{A}(\mathbf{r})$ is cylindrically symmetric, the corresponding magnetic field is obtained, using cylindrical coordinates, as

$$\mathbf{H} = \frac{1}{\mu} \nabla \times \mathbf{A} = \frac{1}{\mu} \left(\hat{\rho} \frac{1}{\rho} \frac{\partial}{\partial \phi} A_z - \hat{\phi} \frac{\partial}{\partial \rho} A_z \right) \tag{25.2.5}$$

where $\frac{\partial}{\partial \phi} = 0$, $r = \sqrt{\rho^2 + z^2}$. In the above, we have used the chain rule that

$$\frac{\partial}{\partial \rho} = \frac{\partial r}{\partial \rho} \frac{\partial}{\partial r} = \frac{\rho}{\sqrt{\rho^2 + z^2}} \frac{\partial}{\partial r} = \frac{\rho}{r} \frac{\partial}{\partial r}.$$

As a result,

$$\mathbf{H} = -\hat{\phi} \frac{\rho}{r} \frac{Il}{4\pi} \left(-\frac{1}{r^2} - j\beta \frac{1}{r} \right) e^{-j\beta r} \tag{25.2.6}$$

In spherical coordinates, $\frac{\rho}{r} = \sin\theta$, and (25.2.6) becomes [32]

$$\mathbf{H} = \hat{\phi} \frac{Il}{4\pi r^2} (1 + j\beta r) e^{-j\beta r} \sin\theta \tag{25.2.7}$$

The electric field can be derived using Maxwell's equations.

$$\mathbf{E} = \frac{1}{j\omega\epsilon}\nabla \times \mathbf{H} = \frac{1}{j\omega\epsilon}\left(\hat{r}\frac{1}{r\sin\theta}\frac{\partial}{\partial\theta}\sin\theta H_\phi - \hat{\theta}\frac{1}{r}\frac{\partial}{\partial r}rH_\phi\right)$$
$$= \frac{Ile^{-j\beta r}}{j\omega\epsilon 4\pi r^3}\left[\hat{r}2\cos\theta(1+j\beta r) + \hat{\theta}\sin\theta(1+j\beta r-\beta^2r^2)\right] \tag{25.2.8}$$

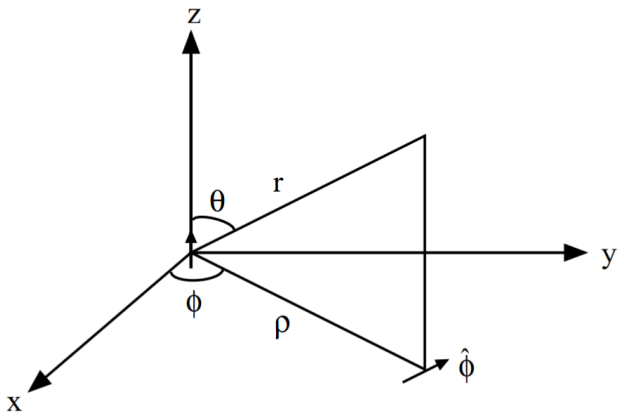


Figure 25.5: Spherical coordinates are used to calculate the fields of a Hertzian dipole.

25.2.1 Case I. Near Field, $\beta r \ll 1$

Since $\beta r \ll 1$, 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 (25.2.8), with $Il = j\omega p$

$$\mathbf{E} \cong \frac{p}{4\pi\epsilon r^3}(\hat{r}2\cos\theta + \hat{\theta}\sin\theta), \qquad \beta r \ll 1 \tag{25.2.9}$$

For the $\mathbf{H}$ field, from (25.2.7), with $\beta r \ll 1$, then

$$\mathbf{H} = \hat{\phi}\frac{j\omega p}{4\pi r^2}\sin\theta \tag{25.2.10}$$

or

$$\eta_0\mathbf{H} = \hat{\phi}\frac{j\beta r p}{4\pi\varepsilon r^3}\sin\theta \tag{25.2.11}$$

Thus, it is seen that

$$\eta_0\mathbf{H} \ll \mathbf{E}, \qquad \text{when } \beta r \ll 1 \tag{25.2.12}$$

where $p = ql$ is the dipole moment.[2] The above implies that in the near field, the electric field dominates over the magnetic field.

In the above, βr could be made very small by making $\frac{r}{\lambda}$ small or by making $\omega \to 0$. The above is 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.[3]

For instance,

$$\frac{\partial}{\partial x} \gg \frac{\partial}{c\partial t}$$

Alternatively, we can say that the above is equivalent to

$$\frac{\partial}{\partial x} \gg \frac{\omega}{c}$$

or that

$$\nabla^2 - \frac{1}{c^2}\frac{\partial^2}{\partial t^2} \approx \nabla^2$$

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 lengthscale.

25.2.2 Case II. Far Field (Radiation Field), $\beta\mathbf{r} \gg 1$

This is also known as the far zone. In this case, retardation effect is important. In other words, phase delay cannot be ignored.

$$\mathbf{E} \cong \hat{\theta} j\omega\mu \frac{Il}{4\pi r} e^{-j\beta r} \sin\theta \tag{25.2.13}$$

and

$$\mathbf{H} \cong \hat{\phi} j\beta \frac{Il}{4\pi r} e^{-j\beta r} \sin\theta \tag{25.2.14}$$

Note that $\frac{E_\theta}{H_\phi} = \frac{\omega\mu}{\beta} = \sqrt{\frac{\mu}{\epsilon}} = \eta_0$. 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.

25.3 Radiation, Power, and Directive Gain Patterns

The time average power flow in the far field is given by

$$\langle \mathbf{S} \rangle = \frac{1}{2}\Re e[\mathbf{E} \times \mathbf{H}^*] = \hat{r}\frac{1}{2}\eta_0 |H_\phi|^2 = \hat{r}\frac{\eta_0}{2}\left(\frac{\beta Il}{4\pi r}\right)^2 \sin^2\theta \tag{25.3.1}$$

[2]Here, $\eta_0 = \sqrt{\mu/\varepsilon}$. We multiply $\mathbf{H}$ by η_0 so that the quantities we are comparing have the same unit.

[3]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 over a small lengthscale compared to wavelength.

The **radiation field pattern** of a Hertzian dipole is the plot of $|\mathbf{E}|$ as a function of θ at a constant **r**. Hence, it is proportional to $\sin\theta$, and it can be proved that it is a circle. The **radiation power pattern** is the plot of $\langle S_r \rangle$ at a constant r.

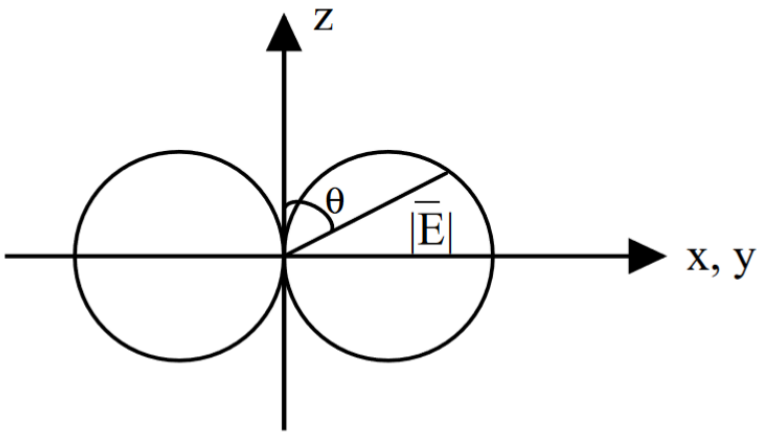


Figure 25.6: Radiation field pattern of a Hertzian dipole. It can be shown that the pattern is a circle.

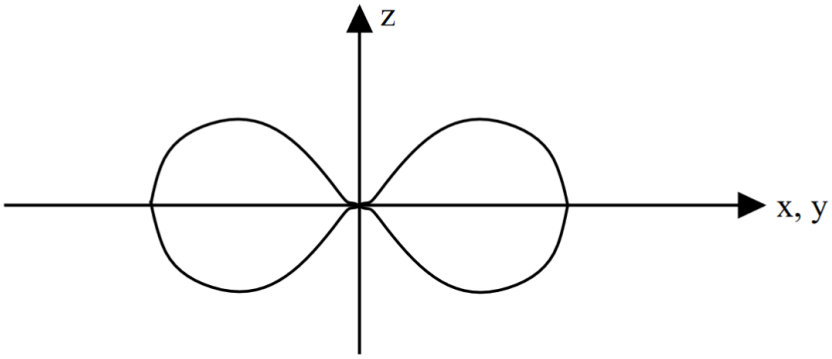


Figure 25.7: Radiation power pattern of a Hertzian dipole which is also the same as the directive gain pattern.

The total power radiated by a Hertzian dipole is thus given by

$$P = \int_0^{2\pi} d\phi \int_0^{\pi} d\theta r^2 \sin\theta \langle S_r \rangle = 2\pi \int_0^{\pi} d\theta \frac{\eta_0}{2} \left(\frac{\beta I l}{4\pi}\right)^2 \sin^3\theta \tag{25.3.2}$$

Since

$$\int_0^{\pi} d\theta \sin^3\theta = -\int_1^{-1} (d\cos\theta)[1-\cos^2\theta] = \int_{-1}^{1} dx(1-x^2) = \frac{4}{3} \tag{25.3.3}$$

then

$$P = \frac{4}{3}\pi\eta_0 \left(\frac{\beta I l}{4\pi}\right)^2 = \frac{\eta_0 (\beta I l)^2}{12\pi} \tag{25.3.4}$$

The **directive gain** of an antenna, $G(\theta, \phi)$, is defined as [32]

$$G(\theta,\phi) = \frac{\langle S_r \rangle}{\langle S_{av} \rangle} = \frac{\langle S_r \rangle}{\frac{P}{4\pi r^2}} \tag{25.3.5}$$

where

$$\langle S_{av} \rangle = \frac{P}{4\pi r^2} \tag{25.3.6}$$

is the power density if the power P were uniformly distributed over a sphere of radius r. Notice that $\langle S_{av} \rangle$ is independent of angle. Hence, the angular dependence of the directive gain $G(\theta, \phi)$ is coming from $\langle S_r \rangle$.

Substituting (25.3.1) and (25.3.4) into the above, we have

$$G(\theta,\phi) = \frac{\frac{\eta_0}{2}\left(\frac{\beta I l}{4\pi r}\right)^2 \sin^2\theta}{\frac{1}{4\pi r^2}\frac{4}{3}\eta_0\pi\left(\frac{\beta I l}{4\pi}\right)^2} = \frac{3}{2}\sin^2\theta \tag{25.3.7}$$

The peak of $G(\theta, \phi)$ is known as the **directivity** of an antenna. It is 1.5 in the case of a Hertzian dipole. If an antenna is radiating isotropically, its directivity is 1, which is the lowest possible value, whereas it can be over 100 for some antennas like reflector antennas (see Figure 25.8). A **directive gain pattern** is a plot of the above function $G(\theta, \phi)$ and it resembles the radiation power pattern.

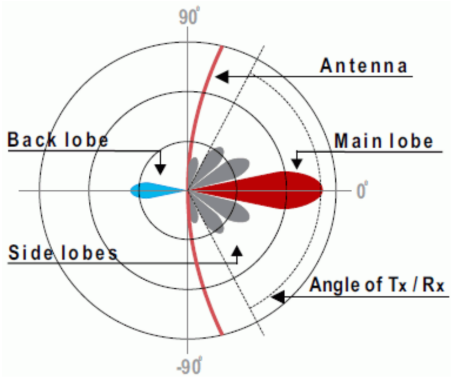


Figure 25.8: The gain of a reflector antenna can be increased by deflecting the power radiated in the desired direction by the use of a reflector (courtesy of racom.eu).

If the total power fed into the antenna instead of the total radiated power is used in the denominator of (25.3.5), the ratio is known as the **power gain** or just **gain** and the pattern is the **power gain pattern**. The total power fed into the antenna is not equal to the total radiated power because there could be some loss in the antenna system like metallic loss.

25.3.1 Radiation Resistance

Engineers love to replace complex systems with simpler systems. Simplicity rules again! This will make interface with electronic driving circuits for the antenna easier to derive. 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.

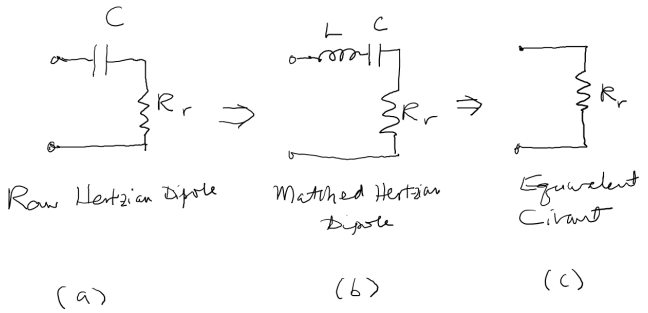


Figure 25.9: (a) Equivalent circuit of a raw Hertzian dipole without matching. (b) Equivalent circuit of a matched Hertzian dipole (using maximum power transfer theorem). (c) Equivalent circuit of a matched dipole at the resonance frequency of the LC tank circuit.

Hence, the way to drive the Hertzian dipole effectively is to use matching network for maximum power transfer. 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. 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 $P = \frac{1}{2}I^2R_r$. Thus, the **radiation resistance** R_r is the effective resistance that will dissipate the same power as the radiation power P when a current I flows through the resistor. Hence, it is defined by [32]

$$R_r = \frac{2P}{I^2} = \eta_0 \frac{(\beta l)^2}{6\pi} \approx 20(\beta l)^2, \qquad \text{where } \eta_0 = 377 \approx 120\pi \ \Omega \tag{25.3.8}$$

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. 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 25.10. 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.[4] Such a formula can be used to estimate the radiation resistance of a dipole.

For example, a half-wave dipole does not have a triangular current distribution a sinusoidal one as shown in Figure 25.11. Nevertheless, we approximate the current distribution of a half-wave dipole with a triangular distribution, and apply the above formula. We pick $a = \frac{\lambda}{2}$, and let $l_{\text{eff}} = \frac{\lambda}{4}$ in (25.3.8), we have

$$R_r \approx 50\Omega \tag{25.3.9}$$

[4]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.

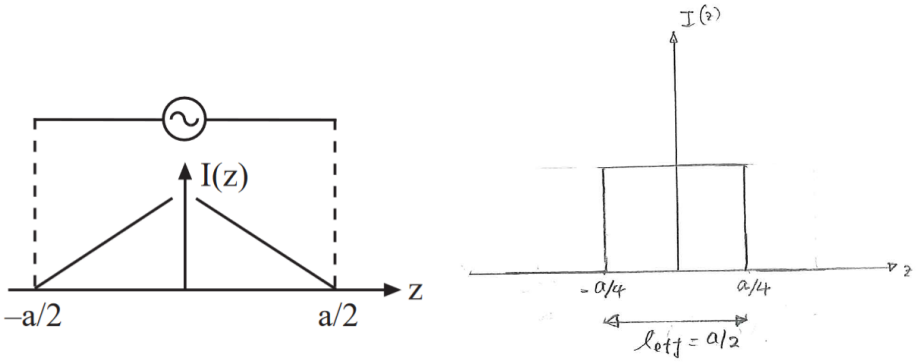


Figure 25.10: The current pattern on a short dipole can be approximated by a triangle since the current has to vanish at the end points of the short dipole. Furthermore, this dipole can be approximated by an effective Hertzian dipole half its length with uniform current.

The true current distribution on a half-wave dipole resembles that shown in Figure 25.11. 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. A more precise calculation shows that $R_r = 73\Omega$ for a half-wave dipole [54]. This also implies that a half-wave dipole with sinusoidal current distribution is a better radiator than a dipole with triangular current distribution.

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 25.12 [137]. 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 gets 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 [138].

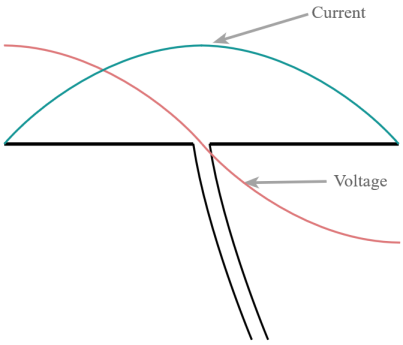
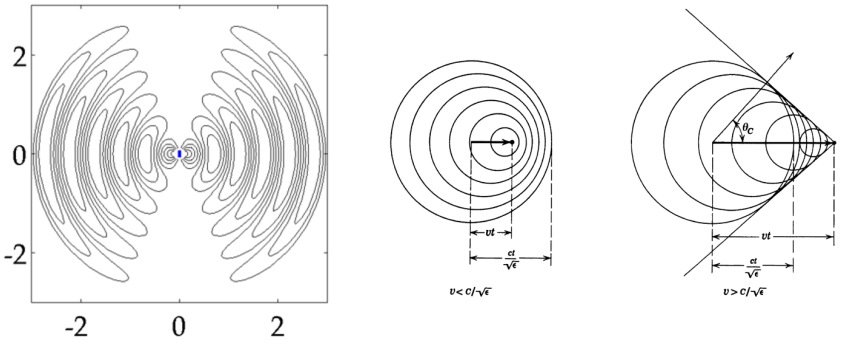


Figure 25.11: Approximate current distribution on a half-wave dipole (courtesy of electronics-notes.co). The currents are zero at the two end tips due to the current continuity equation, or KCL.

Figure 25.12: A bicone antenna can be thought of as a transmission line with gradually changing characteristic impedance. This enhances impedance matching and the radiation of the antenna (courtesy of antennasproduct.com).

Lecture 26

Radiation Fields



(a) (b)

Figure 26.1: (a) Electric field around a time-oscillating dipole (courtesy of physics stack exchange). (b) Equi-potential lines aroud a moving charge that gives rise Cherenkov radiation (courtesy of J.D. Jackson [48]). We will not study Cherenkov (Cerenkov) radiation in this course, but it is written up in [48] and [32]. Its was a Nobel Prize winning discovery.

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 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 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 26.1.

We have shown how to connect the vector and scalar potentials to the sources $\mathbf{J}$ and ϱ of an electromagnetic system. This is a very important connection: it implies that once we know the sources, we know how to find the fields. 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 by assuming that the point where the field is observed is very far from the source location in terms of wavelegth. 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.

26.1 Radiation Fields or Far-Field Approximation

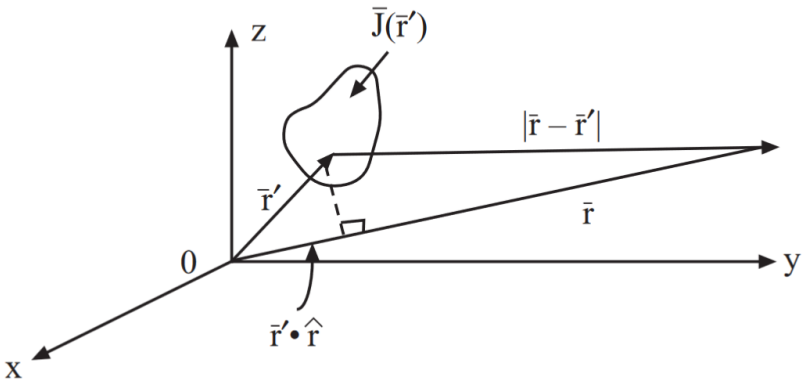


Figure 26.2: The relation of the observation point located a $\mathbf{r}$ to the source location at $\mathbf{r}'$. The distance of the observation point $\mathbf{r}$ to the source location $\mathbf{r}'$ is $|\mathbf{r}-\mathbf{r}'|$.

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 (23.2.29) and (23.2.30)[1] They are given by

$$\mathbf{A}(\mathbf{r}) = \mu \iiint_V d\mathbf{r}' \mathbf{J}(\mathbf{r}') \frac{e^{-j\beta|\mathbf{r}-\mathbf{r}'|}}{4\pi|\mathbf{r}-\mathbf{r}'|} \tag{26.1.1}$$

$$\Phi(\mathbf{r}) = \frac{1}{\varepsilon} \iiint_V d\mathbf{r}' \varrho(\mathbf{r}') \frac{e^{-j\beta|\mathbf{r}-\mathbf{r}'|}}{4\pi|\mathbf{r}-\mathbf{r}'|} \tag{26.1.2}$$

[1]This topic is found in many standard textbooks in electromagnetics [32, 47, 54]. They are also found in lecture notes [44, 139].

where $\beta = \omega\sqrt{\mu\varepsilon} = \omega/c$ is the wavenumber. The integrals in (26.1.1) and (26.1.2) 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.

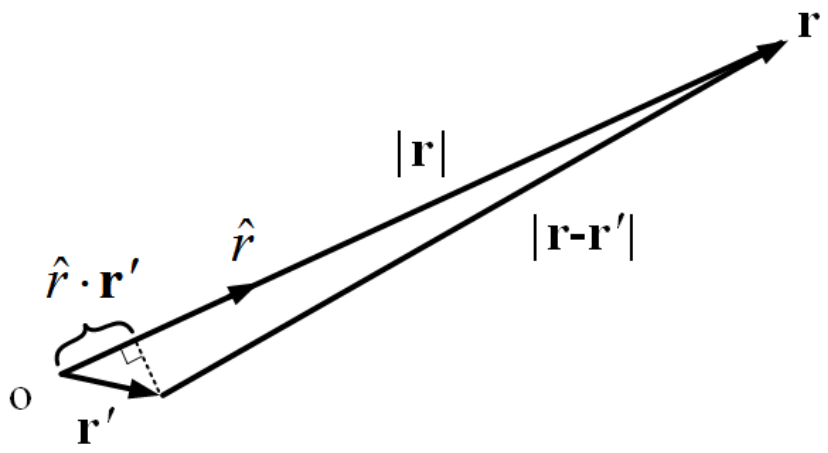


Figure 26.3: The relation between $|\mathbf{r}|$ and $|\mathbf{r}-\mathbf{r}'|$ using the parallax method, or that $|\mathbf{r}-\mathbf{r}'| \approx |\mathbf{r}| - \mathbf{r}' \cdot \hat{r}$. It is assumed that $\mathbf{r}$ is almost parallel to $\mathbf{r} - \mathbf{r}'$.

26.1.1 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 or by geometrical argument as shown in Figure 26.3. Thus (26.1.1) above becomes

$$\mathbf{A}(\mathbf{r}) \approx \frac{\mu}{4\pi} \iiint_V d\mathbf{r}' \frac{\mathbf{J}(\mathbf{r}')}{r - \mathbf{r}' \cdot \hat{r}} e^{-j\beta r + j\beta \mathbf{r}' \cdot \hat{r}} \approx \frac{\mu e^{-j\beta r}}{4\pi r} \iiint_V d\mathbf{r}' \mathbf{J}(\mathbf{r}') e^{j\beta \mathbf{r}' \cdot \hat{r}} \tag{26.1.3}$$

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 (26.1.3) 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 $\mathbf{r}' = \hat{x}x' + \hat{y}y' + \hat{z}z'$, then

$$e^{j\beta \mathbf{r}' \cdot \hat{r}} = e^{j\boldsymbol{\beta} \cdot \mathbf{r}'} = e^{j\beta_x x' + j\beta_y y' + j\beta_z z'} \tag{26.1.4}$$

Therefore (26.1.3) resembles a 3D Fourier transform integral,[2] namely, the above integral

[2]Except that the vector $\boldsymbol{\beta}$ is of fixed length.

becomes

$$\mathbf{A}(\mathbf{r}) \approx \frac{\mu e^{-j\beta r}}{4\pi r} \iiint_V d\mathbf{r}' \mathbf{J}(\mathbf{r}') e^{j\boldsymbol{\beta}\cdot\mathbf{r}'} \tag{26.1.5}$$

and (26.1.5) can be rewritten as

$$\mathbf{A}(\mathbf{r}) \cong \frac{\mu e^{-j\beta r}}{4\pi r} \mathbf{F}(\boldsymbol{\beta}) \tag{26.1.6}$$

where

$$\mathbf{F}(\boldsymbol{\beta}) = \iiint_V d\mathbf{r}' \mathbf{J}(\mathbf{r}') e^{j\boldsymbol{\beta}\cdot\mathbf{r}'} \tag{26.1.7}$$

is the 3D Fourier transform of $\mathbf{J}(\mathbf{r}')$ with the Fourier transform variable $\boldsymbol{\beta} = \hat{r}\beta$.

It is to be noted that this is not a normal 3D Fourier transform because $|\boldsymbol{\beta}|^2 = {\beta_x}^2 + {\beta_y}^2 + {\beta_z}^2 = \beta^2$ 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 normal 3D Fourier transform, β_x, β_y, and β_z are independent variables, each with values in the range $[-\infty, \infty]$. Or the value of ${\beta_x}^2 + {\beta_y}^2 + {\beta_z}^2$ ranges from zero to infinity.

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.

26.1.2 Locally Plane Wave Approximation

We can write $\hat{r}$ or $\boldsymbol{\beta}$ in terms of direction cosines in spherical coordinates or that

$$\hat{r} = \hat{x}\cos\phi\sin\theta + \hat{y}\sin\phi\sin\theta + \hat{z}\cos\theta \tag{26.1.8}$$

Hence,

$$\mathbf{F}(\boldsymbol{\beta}) = \mathbf{F}(\beta\hat{r}) = \mathbf{F}(\beta, \theta, \phi) \tag{26.1.9}$$

It is not truly a 3D function, since β, the length of the vector $\boldsymbol{\beta}$, is fixed. It is a 3D Fourier transform with data restricted on a spherical surface.

Also in (26.1.6), 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 ϕ, the observation angles. In other words, the prefactor in (26.1.6), $\exp(-j\beta r)/r$, can be thought of as resembling a spherical wave. Hence, if one follows a ray of this spherical wave and moves in the r direction, the predominant variation of the field is due to $e^{-j\beta r}$, whereas the direction of the vector $\boldsymbol{\beta}$ changes little, and hence, $\mathbf{F}(\boldsymbol{\beta})$ changes little. Furthermore, $\mathbf{r}'$ in (26.1.7) are restricted to small or finite number, making $\mathbf{F}(\boldsymbol{\beta})$ a weak function of $\boldsymbol{\beta}$ (see Figure 26.4).

The above shows that 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 $2\pi r/\lambda \gg 1$. Or r is many wavelengths away from

the source. Hence, we can 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 26.4 and Figure 26.5

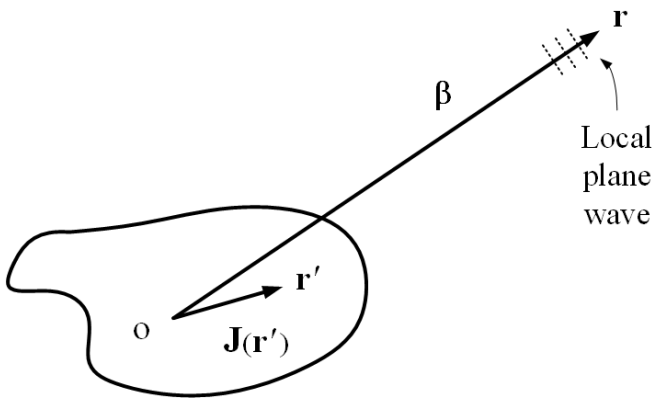


Figure 26.4: A source radiates a field that resembles a spherical wave. In the vicinity of the observation point $\mathbf{r}$, when β is large, the field is strongly dependent on r via $\exp(-j\beta r)$ but weakly dependent on $\boldsymbol{\beta}$ (***b**eta* hardly changes direction in the vicinity of the observation point). Hence, the field becomes locally a plane wave in the far field.

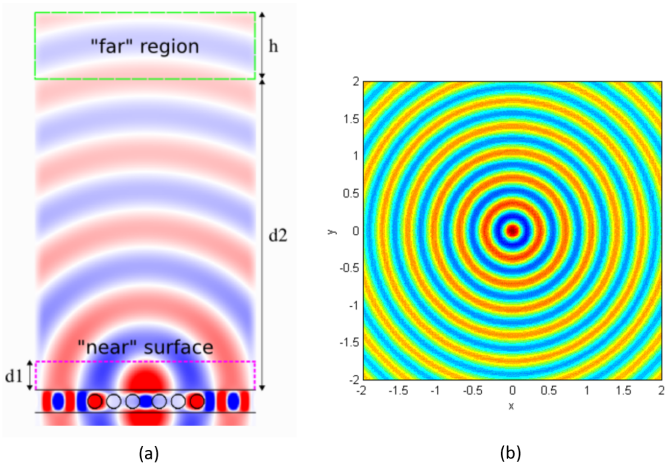


Figure 26.5: (a) A leaky hole in a waveguide leaks a spherical (courtesy of MEEP, MIT). (b) A point source radiates a spherical wave (courtesy of ME513, Purdue Engineering). Most of these simulations are done with FDTD (finite-difference time-domain) method that we will learn later in the course. When the wavelength is short, or the frequency high, a spherical wave front looks locally plane. This is similar to the notion that as humans, who are short, 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, $\nabla \to -j\boldsymbol{\beta} = -j\beta\hat{r}$, we have

$$\mathbf{H} = \frac{1}{\mu}\nabla \times \mathbf{A} \approx -j\frac{\beta}{\mu}\hat{r} \times (\hat{\theta}A_\theta + \hat{\phi}A_\phi) = j\frac{\beta}{\mu}(\hat{\theta}A_\phi - \hat{\phi}A_\theta) \tag{26.1.10}$$

Similarly [44, 139],

$$\mathbf{E} = \frac{1}{j\omega\varepsilon}\nabla \times \mathbf{H} \cong -j\frac{\beta}{\omega\epsilon}\hat{r} \times \mathbf{H} \cong -j\omega(\hat{\theta}A_\theta + \hat{\phi}A_\phi) \tag{26.1.11}$$

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, 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 (26.1.10) for the magnetic field, the electric field can also be written as

$$\mathbf{E} = \frac{1}{j\omega\mu\varepsilon}\nabla \times \nabla \times \mathbf{A} \tag{26.1.12}$$

Using the formula for the double-curl operator, the above can be rewritten as

$$\mathbf{E} = \frac{1}{j\omega\mu\varepsilon}\left(\nabla\nabla \cdot \mathbf{A} - \nabla^2\mathbf{A}\right) \cong \frac{1}{j\omega\mu\varepsilon}\left(-\boldsymbol{\beta}\boldsymbol{\beta} + \beta^2\bar{\mathbf{I}}\right) \cdot \mathbf{A} \tag{26.1.13}$$

where we have used that $\nabla \cong -j\boldsymbol{\beta}$ and $\nabla^2\mathbf{A} = -\beta^2\mathbf{A}$.[3] Alternatively, we can factor $\beta^2 = \omega^2\mu\epsilon$ out of the parenthesis, and rewrite the above as

$$\mathbf{E} \cong -j\omega\left(-\hat{\beta}\hat{\beta} + \bar{\mathbf{I}}\right)\cdot\mathbf{A} = -j\omega\left(-\hat{r}\hat{r} + \bar{\mathbf{I}}\right)\cdot\mathbf{A} \tag{26.1.14}$$

Since $\bar{\mathbf{I}} = \hat{r}\hat{r} + \hat{\theta}\hat{\theta} + \hat{\phi}\hat{\phi}$, then the above becomes

$$\mathbf{E} \cong -j\omega\left(\hat{\theta}\hat{\theta} + \hat{\phi}\hat{\phi}\right)\cdot\mathbf{A} = -j\omega(\hat{\theta}A_\theta + \hat{\phi}A_\phi) \tag{26.1.15}$$

which is the same as previously derived. It also shows that the electric field is transverse to the $\boldsymbol{\beta}$ vector.[4]

Furthermore, it can be shown that in the far field, using the local plane-wave approximation,

$$|\mathbf{E}|/|\mathbf{H}| \approx \eta \tag{26.1.16}$$

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

$$\langle\mathbf{S}\rangle = \frac{1}{2}\Re e\left(\mathbf{E}\times\mathbf{H}^*\right) \approx \frac{1}{2\eta}|\mathbf{E}|^2\hat{r} = \langle S_r\rangle\hat{r} \tag{26.1.17}$$

which resembles also the property of a plane wave.[5] Since the radiated field is a spherical wave, the Poynting's vector is radial. Therefore,

$$\langle\mathbf{S}\rangle = \hat{r}\langle S_r(\theta,\phi)\rangle, \quad \text{where} \quad \langle S_r(\theta,\phi)\rangle = \frac{1}{2\eta}|\mathbf{E}|^2 \tag{26.1.18}$$

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.

26.1.3 Directive Gain Pattern Revisited

We have defined the directive gain pattern for a Hertzian dipole before in Section 25.3. But this concept can be applied to a general radiating source or antenna. 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 can be found by integrating over all angles, viz.,

$$P_T = \int_0^\pi \int_0^{2\pi} r^2 \sin\theta d\theta d\phi \langle S_r(\theta,\phi)\rangle \tag{26.1.19}$$

[3]Note that $\nabla\cdot\mathbf{A} \neq 0$ here.

[4]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!

[5]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).

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

$$\langle S_{\text{av}} \rangle = \frac{P_T}{4\pi r^2} \tag{26.1.20}$$

A dimensionless directive gain pattern can be defined as before in Section 25.3 such that [32, 139]

$$G(\theta, \phi) = \frac{\langle S_r(\theta, \phi) \rangle}{\langle S_{\text{av}} \rangle} = \frac{4\pi r^2 \langle S_r(\theta, \phi) \rangle}{P_T} \tag{26.1.21}$$

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. As in the Hertzian dipole case, the directivity of an antenna $D = \max(G(\theta, \phi))$, is the maximum value of the directive gain. It is to be noted that by its mere definition,

$$\int d\Omega G(\theta, \phi) = 4\pi \tag{26.1.22}$$

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.

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

$$P_{\text{received}} = \langle S_{\text{inc}} \rangle A_e \tag{26.1.23}$$

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. If the receive antenna is made of PEC, induced current will form on the receive antenna so as to generated a field that will cancel the incident field on the PEC surface. This induced current generates a voltage at the receiver load, and hence power received by the antenna.

A wonderful relationship exists between the directive gain pattern $G(\theta, \phi)$ and the effective aperture, namely that[6]

$$A_e = \frac{\lambda^2}{4\pi} G(\theta, \phi) \tag{26.1.24}$$

Therefore, the effective aperture of an antenna is also direction dependent. The above implies that the radiation property of an antenna is related to its receiving property. This is a

[6]The proof of this formula is beyond the scope of this lecture, but we will elaborate on it when we discuss reciprocity theorem.

beautiful consequence of reciprocity theorem that we will study later! 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 [32], or using blackbody radiation law [139, 140].

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 26.6. 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 26.7 shows the gain pattern of a sophisticated antenna array design for 5G applications.

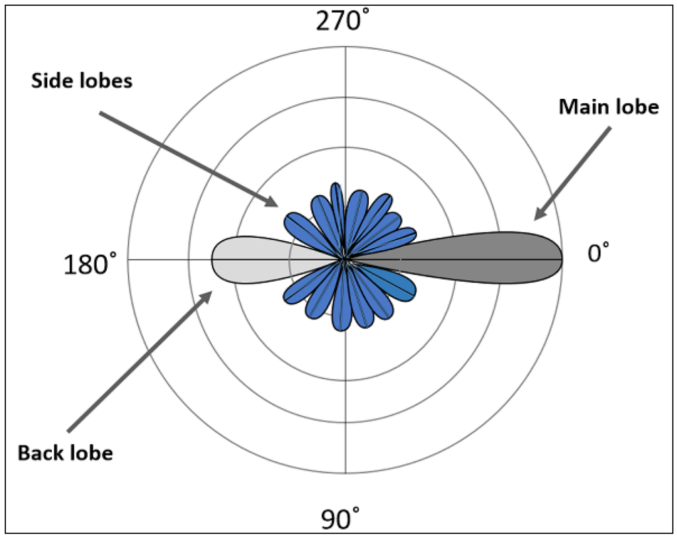


Figure 26.6: The directive gain pattern of an array antenna. The directivity is increased by constructive interference (courtesy of Wikepedia).

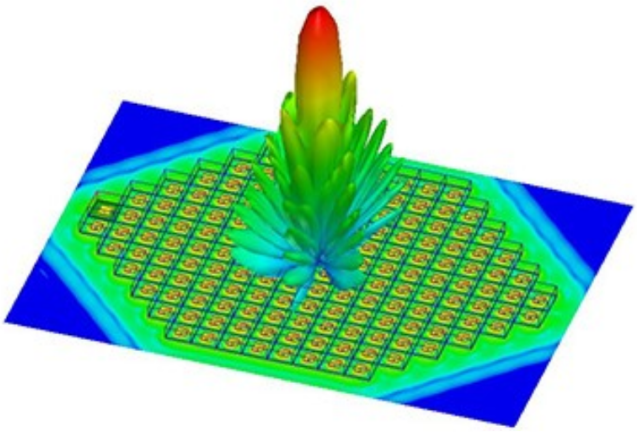


Figure 26.7: The directive gain pattern of a sophisticated array antenna for 5G applications (courtesy of Ozeninc.com).

Lecture 27

Array Antennas, Fresnel Zone, Rayleigh Distance

Our world is beset with the dizzying impact of wireless communication. It has greatly impact our lives.[1] Wireless communication is impossible without using antennas. Hence it is important to design these communication systems with utmost efficiency and sensitivity. We have seen that a simple Hertzian dipole has low directivity in Section 25.3. The radiation pattern looks like that of a donut, and the directivity of the 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. They can be made to constructively interfere in the desired direction, and destructively interfere in other directions to enhance their directivity. 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, clever engineering can be done 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 [141, 142].

27.1 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. The relative phases of the array elements can be changed in time 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, and they are the input to Maxwell's equations, and the fields are

[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.

the output of the system, then we are dealing with a linear 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 can be obtained by inverse Fourier transform. This is provided that the current sources are impressed and they are not affected by the fields that they radiate.

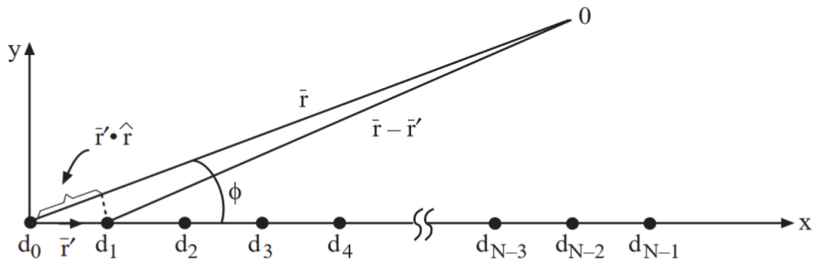


Figure 27.1: Schematics of a dipole array. To simplify the math, the far-field approximation can be used to find its far field or radiation field.

A simple linear dipole array is shown in Figure 27.1. 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:

$$\begin{aligned}\mathbf{J}(\mathbf{r}') = \hat{z}Il[&A_0\delta(x') + A_1\delta(x'-d_1) + A_2\delta(x'-d_2) + \cdots \\ &+ A_{N-1}\delta(x'-d_{N-1})]\delta(y')\delta(z')\end{aligned} \tag{27.1.1}$$

The far field can be found using the approximate formula derived in the previous lecture, viz., (26.1.3) reproduced below:

$$\mathbf{A}(\mathbf{r}) \approx \frac{\mu e^{-j\beta r}}{4\pi r} \iiint_V d\mathbf{r}'\mathbf{J}(\mathbf{r}')e^{j\boldsymbol{\beta}\cdot\mathbf{r}'} \tag{27.1.2}$$

To reiterate, the above implies that the far field is related to the Fourier transform of the current source $\mathbf{J}(\mathbf{r}')$.

27.1.1 Far-Field Approximation

The vector potential on the xy-plane in the far field, using the sifting property of delta function, yield the following equation for $\mathbf{A}(\mathbf{r})$ using (27.1.2),

$$\mathbf{A}(\mathbf{r}) \cong \hat{z}\frac{\mu Il}{4\pi r}e^{-j\beta r} \iiint d\mathbf{r}'[A_0\delta(x') + A_1\delta(x'-d_1) + \cdots]\delta(y')\delta(z')e^{j\beta\mathbf{r}'\cdot\hat{r}} \tag{27.1.3}$$

In the above, for simplicity, we will assume that the observation point is on the xy plane, or that $\mathbf{r} = \boldsymbol{\rho} = \hat{x}x + \hat{y}y$. 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, we have

$$\mathbf{A}(\mathbf{r}) \cong \hat{z} \frac{\mu I l}{4\pi r} e^{-j\beta r} [A_0 + A_1 e^{j\beta d_1 \cos\phi} + A_2 e^{j\beta d_2 \cos\phi} + \cdots + A_{N-1} e^{j\beta d_{N-1} \cos\phi}] \tag{27.1.4}$$

Next, for further simplification, we let $d_n = nd$, implying an equally space array with distance d between adjacent elements. Then we let $A_n = e^{jn\psi}$, which assumes a progressively increasing phase shift between different elements. Such an antenna array is called a **linear phase array**. Thus, (27.1.3) in the above becomes

$$\begin{aligned} \mathbf{A}(\mathbf{r}) \cong \hat{z} \frac{\mu I l}{4\pi r} e^{-j\beta r} [1 + e^{j(\beta d \cos\phi + \psi)} + e^{j2(\beta d \cos\phi + \psi)} + \cdots \\ + e^{j(N-1)(\beta d \cos\phi + \psi)}] \end{aligned} \tag{27.1.5}$$

With the simplifying assumptions, the above series can be summed in closed form.

27.1.2 Radiation Pattern of an Array

The above(27.1.5) can be summed in closed form using the formula

$$\sum_{n=0}^{N-1} x^n = \frac{1 - x^N}{1 - x} \tag{27.1.6}$$

Then in the far field,

$$\mathbf{A}(\mathbf{r}) \cong \hat{z} \frac{\mu I l}{4\pi r} e^{-j\beta r} \frac{1 - e^{jN(\beta d \cos\phi + \psi)}}{1 - e^{j(\beta d \cos\phi + \psi)}} \tag{27.1.7}$$

Ordinarily, as shown previously in (26.1.11), $\mathbf{E} \approx -j\omega(\hat{\theta} A_\theta + \hat{\phi} A_\phi)$. But since $\mathbf{A}$ is $\hat{z}$ directed, $A_\phi = 0$. Furthermore, on the xy plane, $E_\theta \approx -j\omega A_\theta = j\omega A_z$. As a consequence,

$$\begin{aligned} |E_\theta| &= |E_0| \left| \frac{1 - e^{jN(\beta d \cos\phi + \psi)}}{1 - e^{j(\beta d \cos\phi + \psi)}} \right|, \quad \mathbf{r} \to \infty \\ &= |E_0| \left| \frac{\sin\left(\frac{N}{2}(\beta d \cos\phi + \psi)\right)}{\sin\left(\frac{1}{2}(\beta d \cos\phi + \psi)\right)} \right|, \quad \mathbf{r} \to \infty \end{aligned} \tag{27.1.8}$$

The factor multiplying $|E_0|$ above is also called the array factor. The above can be used to plot the far-field pattern of an antenna array.

Equation (27.1.8) has an array factor that is of the form

$$\frac{|\sin(Nx)|}{|\sin x|}$$

This function appears in digital signal processing frequently, and is known as the digital sinc function [143]. 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 in digital signal processing, is the digital sinc.

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 27.2. The function peaks when both the numerator and the denominator of the digital sinc vanish. The value can be found by Taylor series expansion. This happens when $x = n\pi$ for integer n. Such an apparent singularity is called a removable singularity.

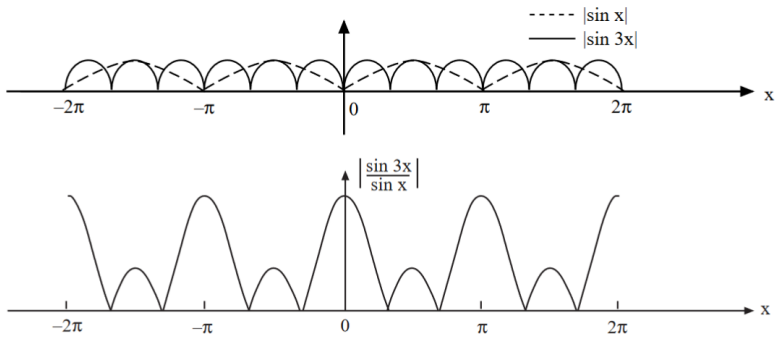


Figure 27.2: Plot of the digital sinc, $\frac{|\sin 3x|}{|\sin x|}$.

In equation (27.1.8), $x = \frac{1}{2}(\beta d \cos\phi + \psi)$. We notice that the **maximum** in (27.1.8) would occur if $x = n\pi$, or if

$$\beta d \cos\phi + \psi = 2n\pi, \qquad n = 0, \pm 1, \pm 2, \pm 3, \cdots \tag{27.1.9}$$

The **zeros** or **nulls** will occur at $Nx = n\pi$, or

$$\beta d \cos\phi + \psi = \frac{2n\pi}{N}, \qquad n = \pm 1, \pm 2, \pm 3, \cdots, \quad n \neq mN \tag{27.1.10}$$

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**. 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 the broadside direction. Hence, this is called a **broadside array**. The radiation pattern of such an array is shown in Figure 27.3.

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 27.4 shows the radiation pattern of such an array.

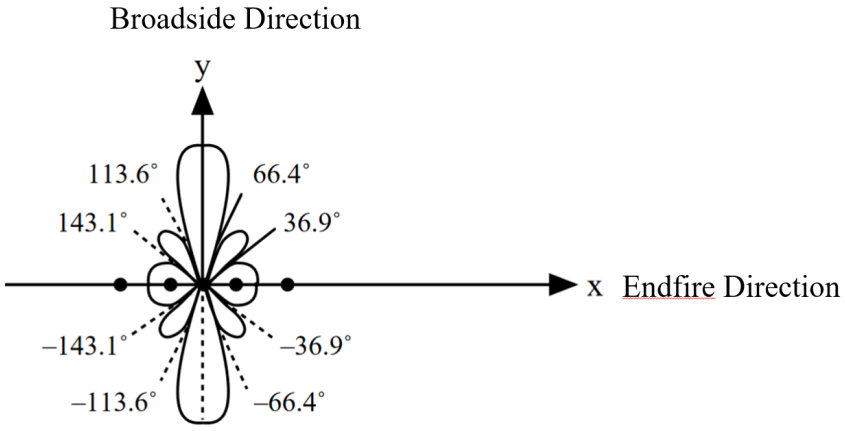


Figure 27.3: The radiation pattern of a five-element broadside array. The broadside and endfire directions of the array are also labeled.

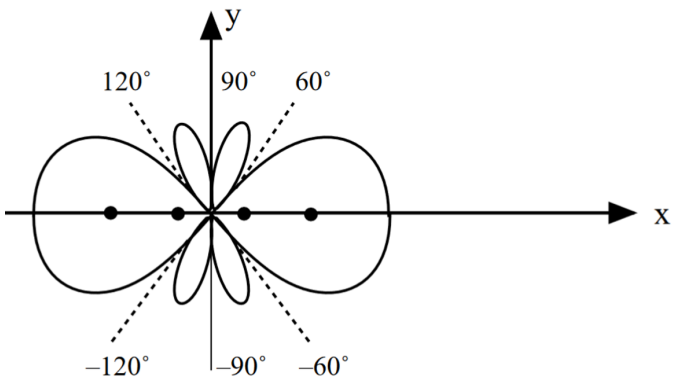


Figure 27.4: By changing the phase of the linear array, the radiation pattern of the antenna array can be changed to become an endfire array.

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 linear 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 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 or lobes 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.

27.2 When is Far-Field Approximation Valid?

In making the far-field approximation in (27.1.3), it will be interesting to ponder when the far-field approximation is valid? That is, when we can approximate

$$e^{-j\beta|\mathbf{r}-\mathbf{r}'|} \approx e^{-j\beta r+j\beta\mathbf{r}'\cdot\hat{r}} \tag{27.2.1}$$

to arrive at (27.1.3). 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 become a spherical wave.

To answer this question, we need to study the approximation in (27.2.1) more carefully. First, we have

$$|\mathbf{r}-\mathbf{r}'|^2 = (\mathbf{r}-\mathbf{r}')\cdot(\mathbf{r}-\mathbf{r}') = r^2 - 2\mathbf{r}\cdot\mathbf{r}' + r'^2 \tag{27.2.2}$$

We can take the square root of the above to get

$$|\mathbf{r}-\mathbf{r}'| = r\left(1 - \frac{2\mathbf{r}\cdot\mathbf{r}'}{r^2} + \frac{r'^2}{r^2}\right)^{1/2} \tag{27.2.3}$$

Next, we use the Taylor series expansion to get, for small x, that

$$(1+x)^n \approx 1 + nx + \frac{n(n-1)}{2!}x^2 + \cdots \tag{27.2.4}$$

or that

$$(1+x)^{1/2} \approx 1 + \frac{1}{2}x - \frac{1}{8}x^2 + \cdots \tag{27.2.5}$$

We can apply this approximation by letting

$$x \cong -\frac{2\mathbf{r}\cdot\mathbf{r}'}{r^2} + \frac{r'^2}{r^2}$$

To this end, we arrive at[2]

$$|\mathbf{r}-\mathbf{r}'| \approx r\left[1 - \frac{\mathbf{r}\cdot\mathbf{r}'}{r^2} + \frac{1}{2}\frac{r'^2}{r^2} - \frac{1}{2}\left(\frac{\mathbf{r}\cdot\mathbf{r}'}{r^2}\right)^2 + \cdots\right] \tag{27.2.6}$$

[2]The art of making such approximation is called perturbation expansion [45].

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 (27.2.6) can be neglected.

We can multiply out the right-hand side of the above to further arrive at

$$\begin{aligned} |\mathbf{r} - \mathbf{r}'| &\approx r - \frac{\mathbf{r} \cdot \mathbf{r}'}{r} + \frac{1}{2}\frac{r'^2}{r} - \frac{1}{2}\frac{(\mathbf{r} \cdot \mathbf{r}')^2}{r^3} + \cdots \\ &= r - \hat{r} \cdot \mathbf{r}' + \frac{1}{2}\frac{r'^2}{r} - \frac{1}{2r}(\hat{r} \cdot \mathbf{r}')^2 + \cdots \end{aligned} \tag{27.2.7}$$

The last two terms in the last line of (27.2.7) are of the same order.[3] Moreover, their sum is bounded by $r'^2/(2r)$ since $\hat{r} \cdot \mathbf{r}'$ is always less than r'. Hence, the far field approximation is valid if

$$\beta \frac{r'^2}{2r} \ll 1 \tag{27.2.8}$$

In the above, β is involved because the approximation has to be valid in the exponent, namely $\exp(-j\beta|\mathbf{r} - \mathbf{r}'|)$ where β multiplies $|\mathbf{r} - \mathbf{r}'|$ or its approximation. If (27.2.8) is valid, then

$$e^{j\beta \frac{r'^2}{2r}} \approx 1$$

and thus, the first two terms on the right-hand side of (27.2.7) 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.

27.2.1 Rayleigh Distance

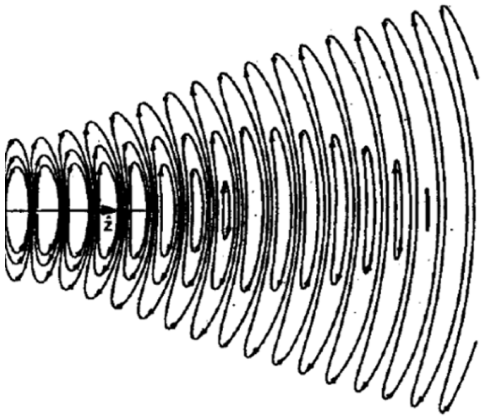


Figure 27.5: The right half of a Gaussian beam [82] displays the physics of the near field, the Fresnel zone, and the far zone. In the far zone, the field behaves like a spherical wave.

[3]The math parlance for saying that these two terms are approximately of the same magnitude as each other.

If we have an infinite time-harmonic current sheet, it can be shown that by matching boundary conditions, it will launch plane waves on both sides of the current sheet [32][p. 652]. 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. Thus, when a wave field leaves an aperture antenna, it can be approximately described by a Gaussian beam [82] (see Figure 27.5). 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.

Consequently, after using that $\beta = 2\pi/\lambda$, for the far-field approximation to be valid, we need (27.2.8) to be valid, or that

$$r \gg \frac{\pi}{\lambda} r'^2 \tag{27.2.9}$$

If the aperture of the antenna is of radius W, then $r' < r_{\max} \cong W$ and the far field approximation is valid if

$$r \gg \frac{\pi}{\lambda} W^2 = r_R \tag{27.2.10}$$

If r is larger than this distance, then an antenna beam behaves like a spherical wave and starts to diverge. This distance r_R is also known as the Rayleigh distance. After this distance, the wave from a finite size source resembles a spherical wave which is diverging in all directions (see Figure 27.5). Also, notice that the shorter the wavelength λ, the larger is this distance. 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 [32], 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 (27.2.8), one gets

$$\beta \frac{r'^2}{2r} \approx \frac{\pi}{8} \tag{27.2.11}$$

Using the approximation, the Rayleigh distance is defined to be

$$r_R = \frac{2D^2}{\lambda} \tag{27.2.12}$$

where $D = 2W$ is the diameter of the antenna aperture. This concept is important in both optics and microwave.

27.2.2 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. 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. As a result, the field in the near zone decays rapidly, but the field in the far zone decays as $1/r$ for energy conservation. In other words, the far field convects energy to infinity.

Lecture 28

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 transmission lines, waveguides, resonators, and dipoles solutions. Examples of approximate solutions are 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 for new designs.

Fortunately for us, Maxwell's equations are accurate 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 have been possible for Maxwell's equations of complex structures. This field of solving Maxwell's equations numerically is known as computational electromagnetics. It shall be discussed later in this course, and many commercial software are available to solve Maxwell's equations to high fidelity. Therefore, design engineers these days do not require higher 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. 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. In the parlance of the field, it is known as virtual proto-typing. It is mainly driven by heuristics and cut-and-try engineering. Therefore, we will discuss the functions of different antennas heuristically in this lecture.

28.1 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 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 can 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 28.1 [137]. These antennas are using resonance tunneling to increase the currents on them 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. It can be gradually morphed from a quarter-wavelength transmission line as shown in Figure 28.1. 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. The dipole antenna, though a simple device, has been extensively studied by King [144].[1]

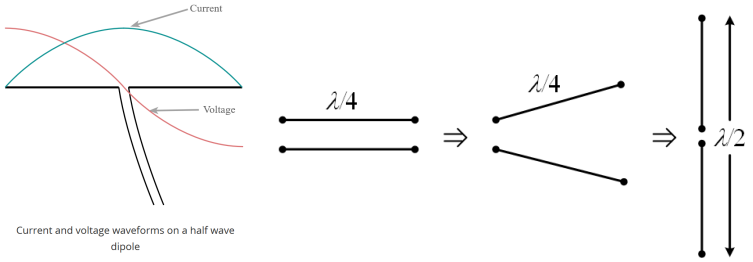


Figure 28.1: A half-wave dipole can be thought of as a resonator with radiation loss. It can be thought of as a quarter-wavelength transmission line that is gradually opened up (courtesy of electronics-notes.com).

[1] He has reputed to have produced over 100 PhD students studying the dipole antenna.

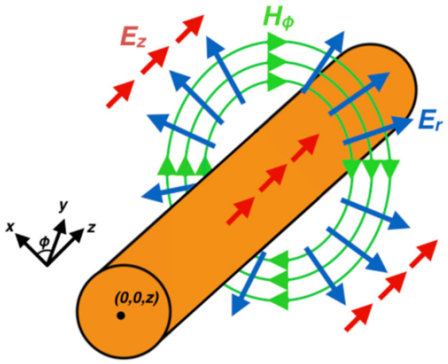


Figure 28.2: The electromagnetic field around a Goubau line (courtesy of [145]). The field resembles that of a coaxial cable with the outer conductor located at infinity.

One can also think of a piece of a wire as a waveguide. It is called a **Goubau line** as shown in Figure 28.2, which can be thought of the limiting case of a coaxial cable where the outer conductor is infinitely far away [145]. The wave is weakly guided since it 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.

A folded dipole is often used to alter the input impedance of a dipole antenna [146]. Even though it can have a resonant frequency lower than that of a normal dipole, the lowest resonant mode does not radiate well. The mode that radiates well has the same resonant length as an unfolded dipole. It has a radiation resistance 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 [147]. Figure 28.3 shows a Yagi-Uda antenna driven by a folded dipole.

Figure 28.3: A Yagi-Uda antenna was invented by heuristics in 1926. The principal element of the antenna is the folded dipole. When a wire dipole antenna is less than half a wavelength, it acts as a waveguide, or a director. When the wire antenna is slightly more than half a wavelength, it acts as a reflector [148]. Therefore, the antenna radiates predominantly in one direction (courtesy of Wikipedia [149]).

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 [148]. Physical intuition was a tool of engineers of yesteryears while modern engineers tend to use sophisticated computer-aided design (CAD) software. The principal driver element of the antenna is the folded dipole. Surprisingly, the elements of dipoles in front of the driver element are acting like a waveguide in space, while the element at the back 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 [150].

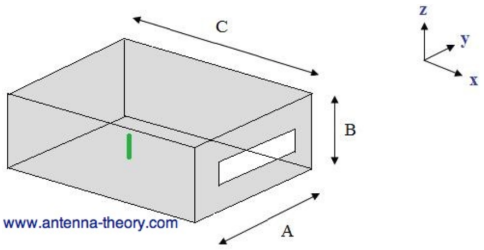


Figure 28.4: A cavity-backed slot antenna radiates well because when the small dipole radiates close to the resonant frequency of the cavity, the field strength is strongly enhanced inside the cavity, and hence around the slot. This makes the slot into a good radiator (courtesy of antenna-theory.com).

Slot antenna is a simple antenna to make [151]. 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 28.4. 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.

Another antenna that resembles a cavity backed slot antenna is the microstrip patch antenna, or patch antenna. This is shown in in Figure 28.5. This antenna also radiates efficiently by resonant tunneling. Roughly, when L (see left of Figure 28.5) is half a wavelength, the patch antenna resonates. This 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 second design (right of Figure 28.5) has an inset feed. This allows the antenna to resonate at a lower frequency because the current has a longer path to slosh through when it is at resonance.

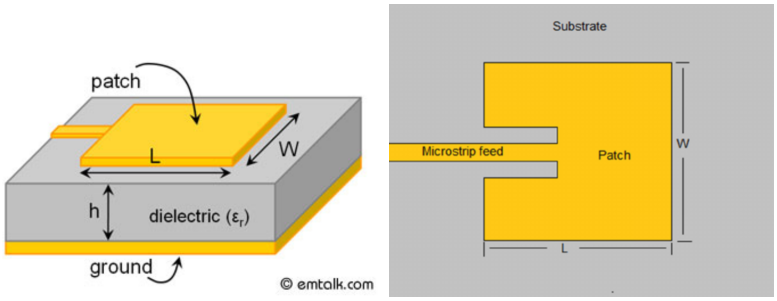


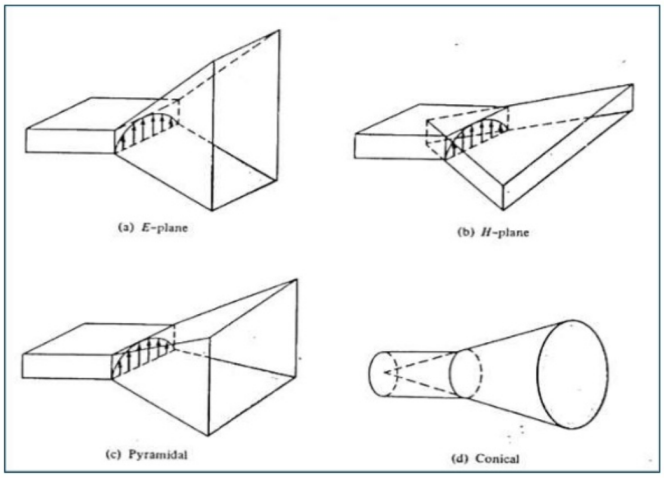
Figure 28.5: A microstrip patch antenna also radiates well when it resonates. The patch antenna resembles a cavity resonator with magnetic wall (courtesy of emtalk.com).

28.2 Horn Antennas

The impedance of free space is 377 ohms while that of most transmission line is 50 ohms. This impedance mismatch can be mitigated by using a flared horn (see Figure 28.6) [152].

One can think that the characteristic impedance of a transmission line made of two pieces of metal as $Z_0 = \sqrt{L/C}$. As the horn flares, C becomes smaller, increasing its characteristic impedance to get close to that of free space. This allows for better impedance matching from the source to free space. This 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 $Z_T = \sqrt{Z_0 Z_L}$

A corrugated horn, as we have discussed previously in a circular waveguide in Section 20.1.1, discourages current flows in the non-axial symmetric mode. It encourages the propagation of the TE_{01} mode in the circular waveguide and hence, the circular horn antenna. This mode is axially symmetric, and thus, this antenna can radiate fields that are axially symmetric [153, 154].



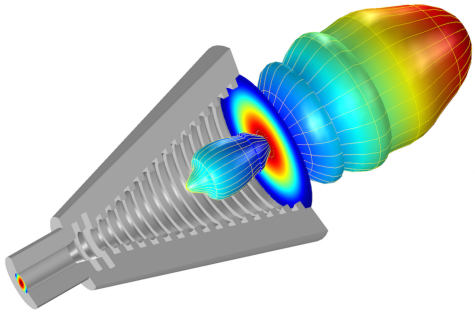


Figure 28.6: A horn antenna works with the same principle as the biconical antenna. Its flared horn changes the waveguide impedance so as to match the impedance of a waveguide to the impedance of free space. The lower figure is that of a corrugated circular horn antenna. The corrugation enhances the 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 [155]), is shown in Figure 28.7.[2] It is also called a notch antenna. It works by the same principle to gradually match the impedance of the source to that of free space. But such a gradually 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 notch. Thus, this antenna can radiate effectively over a broad range of frequencies, and thus this gives the antenna a broad bandwidth performance. It is good for transmitting a pulsed signal which has a broad frequency spectrum.

[2]He must have loved the musician Vivaldi so much:)

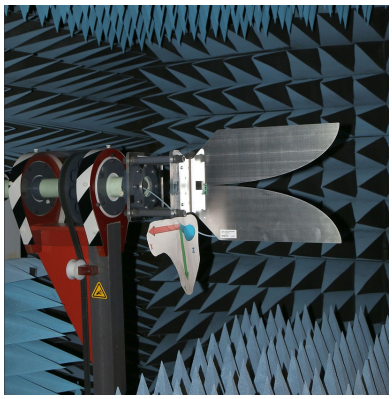


Figure 28.7: A Vivaldi antenna, also called a notch antenna, works like a horn antenna, but uses very little metal. Hence, it is cheap to build, and its flared notch makes it broadband (courtesy of Wikipedia [156]).

28.3 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 28.8. The reflector antenna in this case is a Cassegrain design [157][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 [159] are prevalent in radio astronomy and space exploration due to their high directivity and sensitivity.

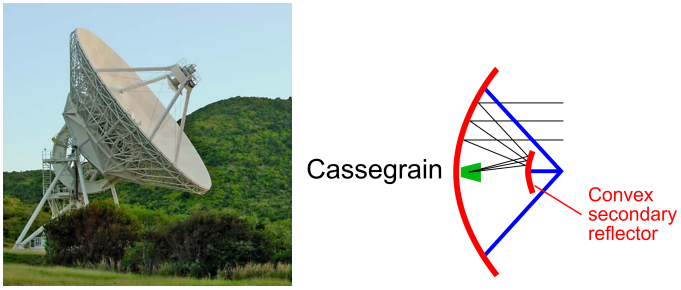


Figure 28.8: The left picture of an NRAO radio telescope antenna of Cassegrain design in Virginia, USA (courtesy of Britannica.com). The bottom is the detail of the Cassegrain design (courtesy of rev.com).

[3]The name came from an optical telescope of similar design [158]

Another recent invention is the reflectarray antenna [160,161] which is very popular. One of them is as shown in Figure 28.9. 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 meta-surface [162, 163]. It can greatly economize on the space usage compared to a reflector antenna.

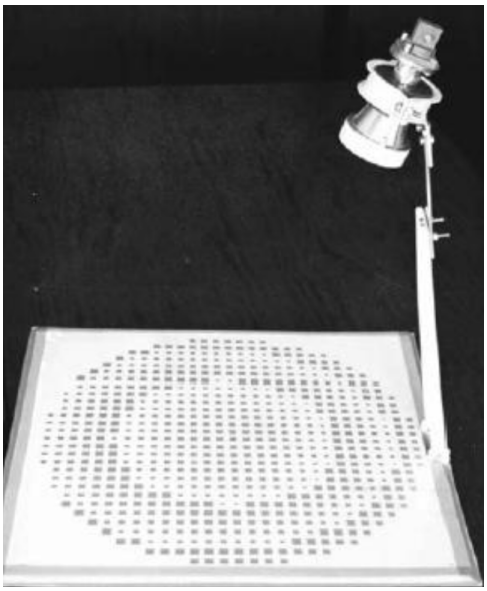


Figure 28.9: A reflectarray where the reflector is a flat surface. Patches are unequally spaced to give the array the focussing effect (courtesy of antenna-theory.com).

Another quasi-optical antenna is the lens antenna as shown in Figure 28.10 [164]. 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 curve surface is similar to that of a flat surface. This is called the tangent-plane approximation of a curve surface, and is valid at high frequencies.

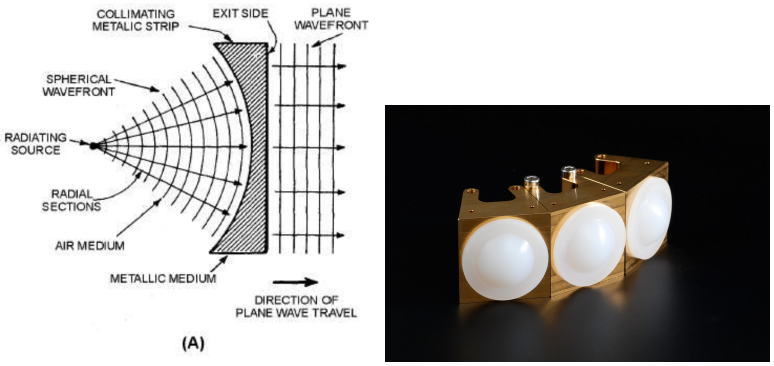


Figure 28.10: The left figure shows a lens antenna where the lens is made of artificial dielectrics made from metallic strips (courtesy of electriciantutoring.tpub.com). The right figure shows some dielectric lens at the aperture of an open waveguide to focus the microwave exiting from the waveguide opening (courtesy of micro-radar.de).

28.4 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) shown in Figure 28.11 [165]. 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. The low Q gives this antenna a broader bandwidth.

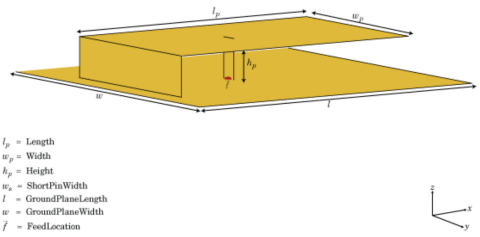


Figure 28.11: A PIFA (planar inverted F antenna) is compact, broadband, and easy to fabricate. It is good for cell phone antennas due to its small size (courtesy of Mathworks).

An interesting small antenna is the U-slot antenna shown in Figure 28.12 [166, 167]. 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 [168]. The L probe has an inductive part as well as a capacitive part. Their reactance cancel each other, allowing the electromagnetic energy to tunnel through the antenna, making it a better radiator.

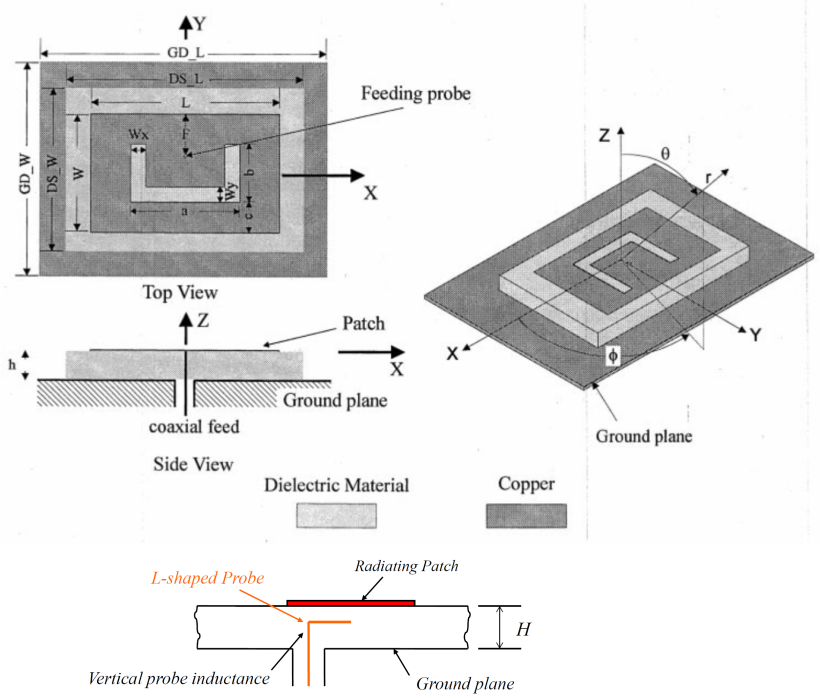


Figure 28.12: The top figure shows a U slot patch antenna design. The bottom figure shows a patch antenna fed by an L probe with significant increase in bandwidth (courtesy of K.M. Luk) [168].

Another area where small antennas are needed is in RFID (radio frequency identification) tag [169]. 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. The communication is done through an RFID reader. The RFID reader can talk to a small com-

[4]Remember that larger inductance implies more store magnetic field energy, and hence, the higher Q of the system.

puter 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 28.13. 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 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.

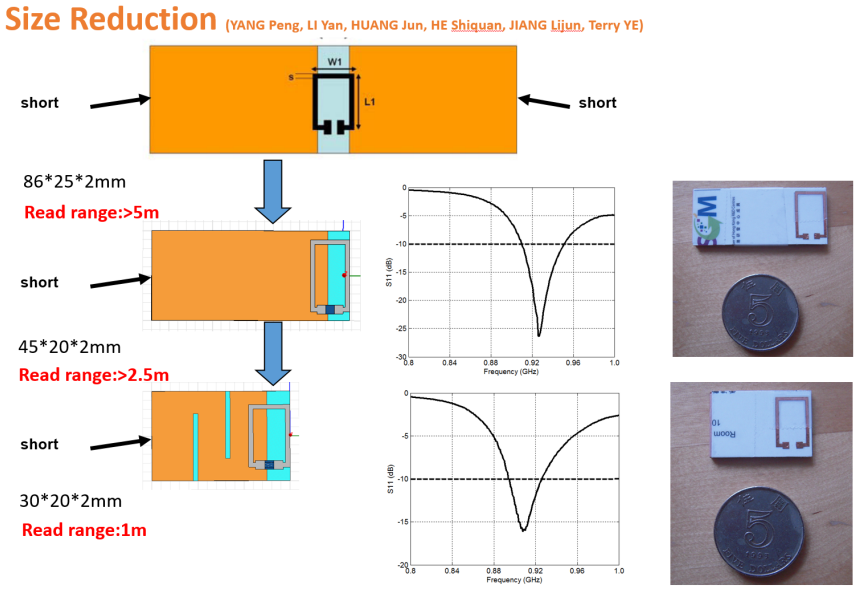


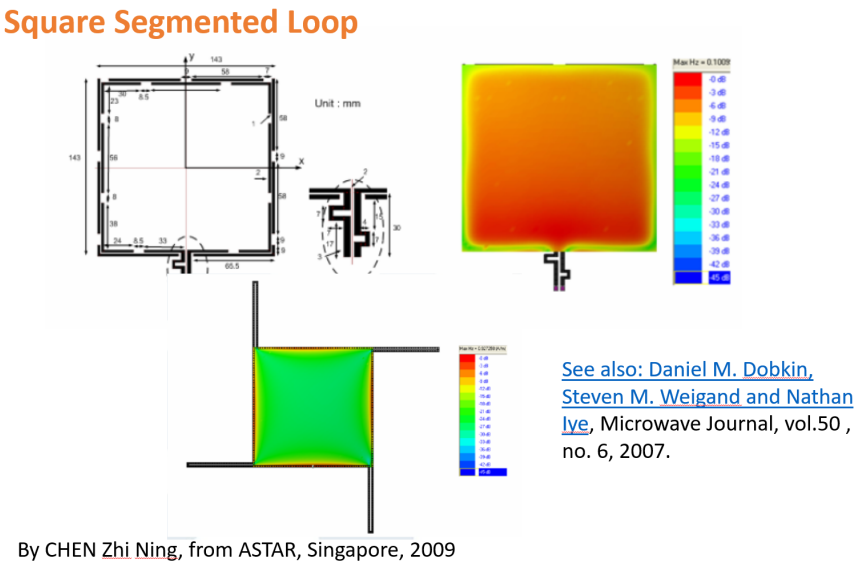
Figure 28.13: Some RFID antennas designed at The University of Hong Kong (courtesy of P. Yang, Y. Li, J. Huang, L.J. Jiang, S.Q. He, T. Ye, and W.C. Chew).

An RFID reader can be designed to read the information from a batch of vials or test tubes containing different chemicals. 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 line [145], will have resonant frequencies. When a loop antenna resonates, the current is non-uniform on it. This happens at higher frequencies. (Fundamentally, this comes from the retardation effect of electromagnetic field.) This will result in a non-uniform field inside the loop defeating the design of the RFID reader.

One way to view how the non-uniform current come about is that a piece of wire becomes a tiny inductor. Across an inductor, $V = j\omega LI$, implying a 90° phase shift between the voltage and the current. In other words, the voltage drop is always nonzero, and therefore, the voltage cannot be constant around the loop. Since the voltage and current are locally

related by the local inductance, the current cannot be constant also.

To solve the problem of the current and voltage being non-constant around the loop, a local inductor is connected in series with a capacitor [170]. This causes the local LC tank circuit to resonate. At resonance, the current-voltage relationship across the LC tank circuit is such that there is no voltage drop across the tank circuit since it becomes a short circuit. In this way, the voltage is equilized between two points and becomes 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 28.14.



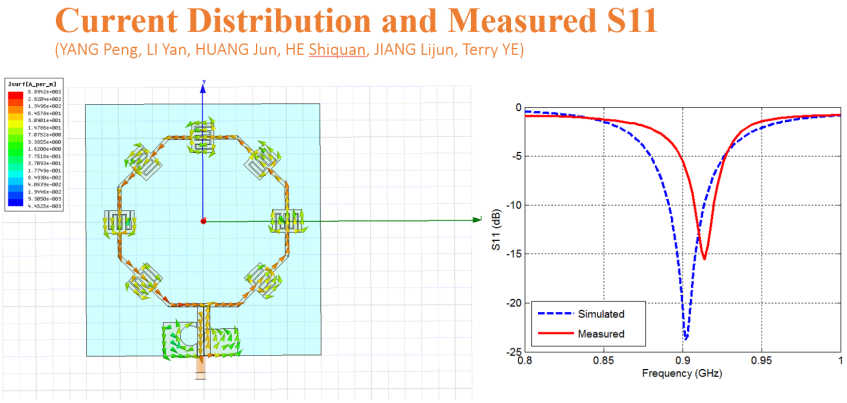


Figure 28.14: The top figure shows a RFID reader designed by [171]. The bottom figure shows simulation and measurement done at The University of Hong Kong (courtesy of Z.N. Chen [171] and P. Yang, Y. Li, J. Huang, L.J. Jiang, S.Q. He, T. Ye, and W.C. Chew).

Lecture 29

Uniqueness Theorem

The uniqueness of a solution to a linear system of equations is an important concept in mathematics. Under certain conditions, ordinary differential equation partial differential equation and matrix equations will have unique solutions. But uniqueness is not always guaranteed as we shall see. This issue is discussed in many math books and linear algebra books [77,89]. The proof of uniqueness for Laplace and Poisson equations are given in [31,54] which is slightly different from electrodynamic problems.

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]

29.1 The Difference Solutions to Source-Free Maxwell's Equations

In this section, we will prove uniqueness theorem for electrodynamic problems [32, 35, 50, 65, 83]. First, let us assume that there exist two solutions in the presence of one set of common impressed sources $\mathbf{J}_i$ and $\mathbf{M}_i$.[2] Namely, these two 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. Are $\mathbf{E}^a = \mathbf{E}^b$, $\mathbf{H}^a = \mathbf{H}^b$?

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, lets assume two possible solutions with one given set of sources $\mathbf{J}_i$ and

[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 it was used in [172] in 1936. These are immutable sources that cannot be changed by the environment in which they are immersed.

$\mathbf{M}_i$, it follows that

$$\nabla \times \mathbf{E}^a = -j\omega\overline{\boldsymbol{\mu}} \cdot \mathbf{H}^a - \mathbf{M}_i \tag{29.1.1}$$
$$\nabla \times \mathbf{E}^b = -j\omega\overline{\boldsymbol{\mu}} \cdot \mathbf{H}^b - \mathbf{M}_i \tag{29.1.2}$$
$$\nabla \times \mathbf{H}^a = j\omega\overline{\boldsymbol{\varepsilon}} \cdot \mathbf{E}^a + \mathbf{J}_i \tag{29.1.3}$$
$$\nabla \times \mathbf{H}^b = j\omega\overline{\boldsymbol{\varepsilon}} \cdot \mathbf{E}^b + \mathbf{J}_i \tag{29.1.4}$$

By taking the difference of these two solutions, we have

$$\nabla \times (\mathbf{E}^a - \mathbf{E}^b) = -j\omega\overline{\boldsymbol{\mu}} \cdot (\mathbf{H}^a - \mathbf{H}^b) \tag{29.1.5}$$
$$\nabla \times (\mathbf{H}^a - \mathbf{H}^b) = j\omega\overline{\boldsymbol{\varepsilon}} \cdot (\mathbf{E}^a - \mathbf{E}^b) \tag{29.1.6}$$

Or alternatively, defining $\delta\mathbf{E} = \mathbf{E}^a - \mathbf{E}^b$ and $\delta\mathbf{H} = \mathbf{H}^a - \mathbf{H}^b$, we have

$$\nabla \times \delta\mathbf{E} = -j\omega\overline{\boldsymbol{\mu}} \cdot \delta\mathbf{H} \tag{29.1.7}$$
$$\nabla \times \delta\mathbf{H} = j\omega\overline{\boldsymbol{\varepsilon}} \cdot \delta\mathbf{E} \tag{29.1.8}$$

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 (29.1.7) and (29.1.8).

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, it can be shown that

$$\nabla \cdot (\delta\mathbf{E} \times \delta\mathbf{H}^*) = \delta\mathbf{H}^* \cdot \nabla \times \delta\mathbf{E} - \delta\mathbf{E} \cdot \nabla \times \delta\mathbf{H}^* \tag{29.1.9}$$

Then by taking the left dot product of $\delta\mathbf{H}^*$ with (29.1.7), and then the left dot product of $\delta\mathbf{E}$ with the complex conjugation of (29.1.8), we obtain

$$\begin{aligned} \delta\mathbf{H}^* \cdot \nabla \times \delta\mathbf{E} &= -j\omega\delta\mathbf{H}^* \cdot \overline{\boldsymbol{\mu}} \cdot \delta\mathbf{H} \\ \delta\mathbf{E} \cdot \nabla \times \delta\mathbf{H}^* &= -j\omega\delta\mathbf{E} \cdot \overline{\boldsymbol{\varepsilon}}^* \cdot \delta\mathbf{E}^* \end{aligned} \tag{29.1.10}$$

Now, taking the difference of the above, we get

$$\begin{aligned} \delta\mathbf{H}^* \cdot \nabla \times \delta\mathbf{E} - \delta\mathbf{E} \cdot \nabla \times \delta\mathbf{H}^* &= \nabla \cdot (\delta\mathbf{E} \times \delta\mathbf{H}^*) \\ &= -j\omega\delta\mathbf{H}^* \cdot \overline{\boldsymbol{\mu}} \cdot \delta\mathbf{H} + j\omega\delta\mathbf{E} \cdot \overline{\boldsymbol{\varepsilon}}^* \cdot \delta\mathbf{E}^* \end{aligned} \tag{29.1.11}$$

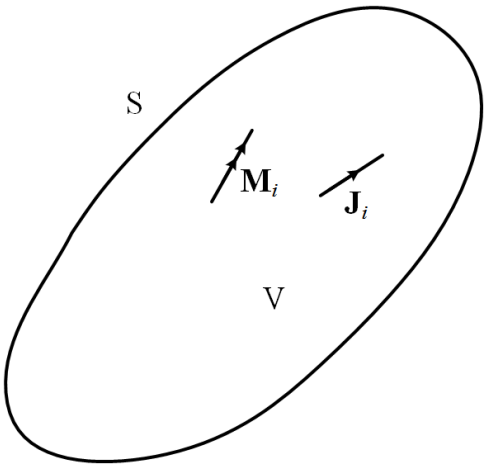


Figure 29.1: Geometry for proving the uniqueness theorem. We like to know the requisite boundary conditions on S plus the type of media inside V in order to guarantee the uniqueness of the solution in V.

Next, we integrate the above equation (29.1.11) over a volume V bounded by a surface S as shown in Figure 29.1. After making use of Gauss' divergence theorem, we arrive at

$$\begin{aligned}\iiint_V \nabla\cdot(\delta\mathbf{E}\times\delta\mathbf{H}^*)dV &= \oiint_S(\delta\mathbf{E}\times\delta\mathbf{H}^*)\cdot d\mathbf{S}\\ &= \iiint_V[-j\omega\delta\mathbf{H}^*\cdot\overline{\boldsymbol{\mu}}\cdot\delta\mathbf{H}+j\omega\delta\mathbf{E}\cdot\overline{\boldsymbol{\varepsilon}}^*\cdot\delta\mathbf{E}^*]dV \qquad (29.1.12)\end{aligned}$$

And next, we would like to know the kind of boundary conditions that would make the left-hand side equal to zero. 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]
$\oiint_S(\delta\mathbf{E}\times\delta\mathbf{H}^*)\cdot\hat{n}dS=\oiint_S(\hat{n}\times\delta\mathbf{E})\cdot\delta\mathbf{H}^*dS=0.$

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

[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 [83].

[4]We use the vector identity 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, $d\mathbf{S}=\hat{n}dS$.

[5]In math parlance, $S_1\cup S_2=S$.

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 (29.1.12) is zero, and then the right-hand side of (29.1.12) becomes

$$\iiint_V [-j\omega\delta\mathbf{H}^* \cdot \overline{\boldsymbol{\mu}} \cdot \delta\mathbf{H} + j\omega\delta\mathbf{E} \cdot \overline{\boldsymbol{\varepsilon}}^* \cdot \delta\mathbf{E}^*] dV = 0 \tag{29.1.13}$$

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

$$\iiint_V [-\delta\mathbf{H}^* \cdot \overline{\boldsymbol{\mu}} \cdot \delta\mathbf{H} + \delta\mathbf{E} \cdot \overline{\boldsymbol{\varepsilon}}^* \cdot \delta\mathbf{E}^*] dV = 0 \tag{29.1.14}$$

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.

For resonant solutions, the stored electric energy can balance the stored magnetic energy. The above resonant solutions are those of the difference solutions satisfying PEC or PMC boundary condition or mixture thereof. Also, they are the resonant solutions of the source-free Maxwell's equations (29.1.7). Therefore, $\delta\mathbf{H}$ and $\delta\mathbf{E}$ need not be zero, even though (29.1.14) is zero. This happens when we encounter solutions that are the resonant modes of the volume V bounded by the surface S.

29.2 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.[7] First we begin with the isotropic case.

29.2.1 Isotropic Case

It is easier to see this for lossy isotropic media. Then (29.1.13) simplifies to

$$\iiint_V [-j\omega\mu|\delta\mathbf{H}|^2 + j\omega\varepsilon^*|\delta\mathbf{E}|^2] dV = 0 \tag{29.2.1}$$

[6] Tensors are a special kind of matrices.

[7] See Chew, Theory of Microwave and Optical Waveguides.

For isotropic lossy media, $\mu = \mu' - j\mu''$ and $\varepsilon = \varepsilon' - j\varepsilon''$. Taking the real part of the above, we have from (29.2.1) that

$$\iiint_V [-\omega\mu''|\delta\mathbf{H}|^2 - \omega\varepsilon''|\delta\mathbf{E}|^2]dV = 0 \tag{29.2.2}$$

Since the integrand in the above is always negative definite, the integral can be zero only if

$$\delta\mathbf{E} = 0, \quad \delta\mathbf{H} = 0 \tag{29.2.3}$$

everywhere in V, implying that $\mathbf{E}_a = \mathbf{E}_b$, and that $\mathbf{H}_a = \mathbf{H}_b$. Hence, 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) which is pure time-harmonic solution cannot exist due to loss. The modes, which are the source-free solutions of Maxwell's equations, are decaying sinusoids. But when we express equations (29.1.1) to (29.1.4) in the frequency domain, we are seeking solutions for which ω is real. Thus decaying sinusoids are not among the possible solutions, and hence, they are not in the solution space.

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. In this case, no time-harmonic solution exists, and the resonant solution is a growing sinusoid.

29.2.2 General Anisotropic Case

The proof for general anisotropic media is more complicated. For the lossless anisotropic media, we see that (29.1.13) is purely imaginary. However, when the medium is lossy, this same equation will have a real part. Hence, we need to find the real part of (29.1.13) 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[8] of (29.1.13), which is scalar, and add it to itself to get its real part. To this end, we will find the conjugate of its integrand which is a scalar number.

First, the complex conjugate of the first scalar term in the integrand of (29.1.13) is[9]

$$(-j\omega\delta\mathbf{H}^* \cdot \overline{\boldsymbol{\mu}} \cdot \delta\mathbf{H})^* = j\omega\delta\mathbf{H} \cdot \overline{\boldsymbol{\mu}}^* \cdot \delta\mathbf{H}^* = j\omega\delta\mathbf{H}^* \cdot \overline{\boldsymbol{\mu}}^\dagger \cdot \delta\mathbf{H} \tag{29.2.4}$$

Similarly, the complex conjugate of the second scalar term in the same integrand is

$$(j\omega\delta\mathbf{E} \cdot \overline{\boldsymbol{\varepsilon}}^* \cdot \delta\mathbf{E}^*)^* = -j\omega\delta\mathbf{E}^* \cdot \overline{\boldsymbol{\varepsilon}}^\dagger \cdot \delta\mathbf{E} \tag{29.2.5}$$

[8] Also called hermitian conjugate.

[9] 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.

But

$$j\omega\delta\mathbf{E}\cdot\overline{\boldsymbol{\varepsilon}}^{*}\cdot\delta\mathbf{E}^{*}=j\omega\delta\mathbf{E}^{*}\cdot\overline{\boldsymbol{\varepsilon}}^{\dagger}\cdot\delta\mathbf{E} \tag{29.2.6}$$

The above gives us the complex conjugate of the scalar quantity (29.1.13) and adding it to itself, we have

$$\iiint_{V}[-j\omega\delta\mathbf{H}^{*}\cdot(\overline{\boldsymbol{\mu}}-\overline{\boldsymbol{\mu}}^{\dagger})\cdot\delta\mathbf{H}-j\omega\delta\mathbf{E}^{*}\cdot(\overline{\boldsymbol{\varepsilon}}-\overline{\boldsymbol{\varepsilon}}^{\dagger})\cdot\delta\mathbf{E}]dV=0 \tag{29.2.7}$$

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. Thus, uniqueness is guaranteed in a lossy anisotropic medium.

Similar statement can be made for the isotropic 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.

29.3 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 [77]. 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

$$\overline{\mathbf{A}}\cdot\mathbf{x}=\mathbf{b} \tag{29.3.1}$$

If a solution to a matrix equation exists without excitation, namely, when $\mathbf{b}=0$, then the solution is the null space solution [77], namely, $\mathbf{x}=\mathbf{x}_N$. In other words,

$$\overline{\mathbf{A}}\cdot\mathbf{x}_N=0 \tag{29.3.2}$$

These null space solutions exist without a "driving term" $\mathbf{b}$ on the right-hand side. For Maxwell's equations, $\mathbf{b}$ corresponds to the source terms. The solution in (29.3.2) is like the homogeneous solution of an ordinary differential equation or a partial differential equation [89]. 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 (29.1.7) and (29.1.8) are Maxwell's equations without the source terms. In a closed region V bounded by a surface S, only resonant solutions for $\delta\mathbf{E}$ and $\delta\mathbf{H}$ with the relevant boundary conditions can exist when there are no source terms.

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 (see Section 22.1.1 for an analogue with LC tank circuit). Since we are looking for solutions in the frequency domain, or time harmonic solutions, the solutions considered are on the real ω axis on the complex ω plane. Thus the non-sinusoidal 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. Therefore,

complex resonant solutions which are off the real axis, and are homogeneous solutions, are not found on the real axis.

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 to time *ad infinitum*. 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. One reason for non-uniqueness is because we are seeking the solutions in the frequency domain.

29.4 Connection to Poles of a Linear System

The output to input of a linear system can be represented by a transfer function $H(\omega)$ [52,173]. 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 gives rise to 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 29.2.

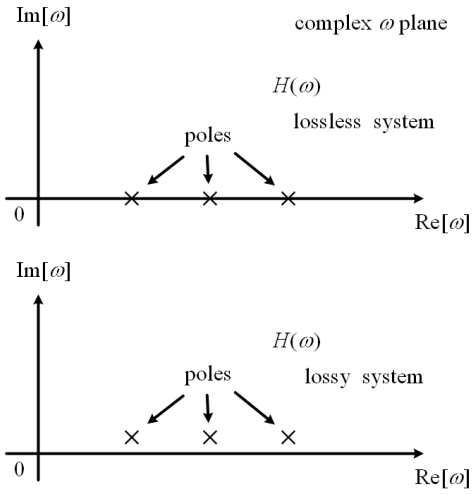


Figure 29.2: The non-uniqueness problem is intimately related to the locations of the poles of a transfer function being on the real axis, when one solves a linear system using Fourier transform technique. For a lossless system, the poles are located on the real axis. When performing a Fourier inverse 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.,

$$y(t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} d\omega e^{j\omega t} H(\omega) F(\omega) \tag{29.4.1}$$

where the Fourier inversion integral path is on the real axis on the complex ω plane. The Fourier inversion integral 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 29.2. Then the transfer function is uniquely determined for all frequencies on the real axis. In this way, the Fourier inversion integral in (29.4.1) 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 [52]. The Laplace transform technique allows the specification of initial values, which is similar to solving the problem as an initial value problem (IVP).

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 $Y(\omega)$ of (22.1.4). The transfer function has two poles. If there is no loss, then the poles are located on the real axis, rendering the Fourier inversion contour undefined in (29.4.1). Hence, the solution is non-unique. However, if infinitesimal loss is introduced by setting $R \neq 0$, then the poles will migrate off the real axis making (29.4.1) well defined!

29.5 Radiation from Antenna Sources and Radiation Condition

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. To see how this can come about, we first study the radiation of sources into a region V bounded by a large surface S_{inf} as shown in Figure 29.4 [35].

Even when $\hat{n} \times \mathbf{E}$ or $\hat{n} \times \mathbf{H}$ are specified on the surface at S_{inf}, the solution is nonunique because the volume V bounded by S_{inf}, can have many resonant solutions. In fact, the region will be replete with resonant solutions as one makes S_{inf} become very large.

To have more insight, we look at the resonant condition of a large rectangular cavity given by (21.2.3) reproduced here as

$$\beta^2 = \frac{\omega^2}{c^2} = \left(\frac{m\pi}{a}\right)^2 + \left(\frac{n\pi}{b}\right)^2 + \left(\frac{p\pi}{d}\right)^2 \tag{29.5.1}$$

The above is an equation of an Ewald sphere in a 3D mode space, but the values of $\beta_x = \frac{m\pi}{a}$, $\beta_y = \frac{n\pi}{b}$, and $\beta_z = \frac{p\pi}{d}$ are discrete. 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 that the above equation can be

satisfied or approximately satisfied becomes very large. This is illustrated Figure 29.3 in 2D. Hence, the chance of the operating frequency ω coinciding with a resonant mode of the cavity is very high giving rise to nonuniqueness. This is even more so when the cavity becomes very large. Hence, the chance of operating inside the cavity with unique solution is increasingly difficult. This above argument applies to cavities of other shapes as well.

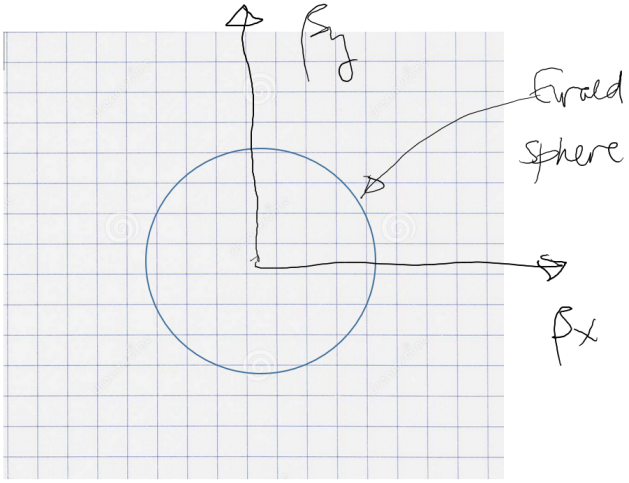


Figure 29.3: For very large cavity, the grid spacing in the mode space (or Fourier space) becomes very small. Then the chance that the sphere surface encounters a resonant mode is very high. When this happens, the solution to the cavity problem is non-unique.

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_{inf} to infinity, and the solution will always be 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** [174]. . 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.

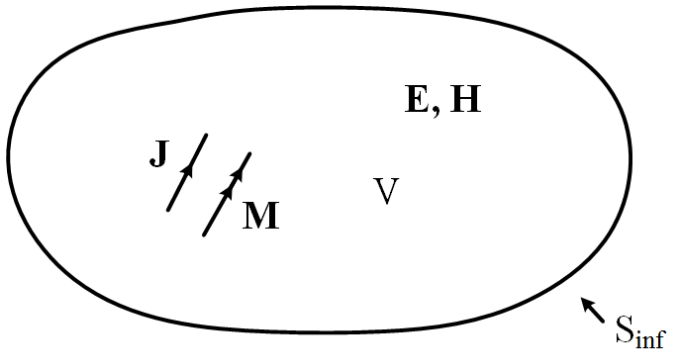


Figure 29.4: The solution for antenna radiation is unique because we impose the Sommerfeld radiation condition when seeking the solution. That is we assume that the radiation wave travels to infinity but never to return. This is equivalent to assuming an infinitesimal loss when seeking the solution in V.

Lecture 30

Reciprocity Theorem

Reciprocity theorem is one of the most important theorems in electromagnetics. With it we can develop physical intuition to ascertain if a certain design or experiment is right or wrong. It also tells us what is possible or impossible in the design of many systems. 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.

子貢問曰：“有一言而可以終身行之者乎？”子曰：“其恕乎！己所不欲、勿施於人。”

Zi Gong [a disciple] asked: "Is there any one word that could guide a person throughout life?"

The Master replied: "How about 'reciprocity'! Never impose on others what you would not choose for yourself."

Analects XV.24, tr. David Hinton

Figure 30.1: (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 translated by D. Hinton [175]. (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 [176]. The subsequent portraits of these two sages are more humanly urbane.

30.1 Mathematical Derivation

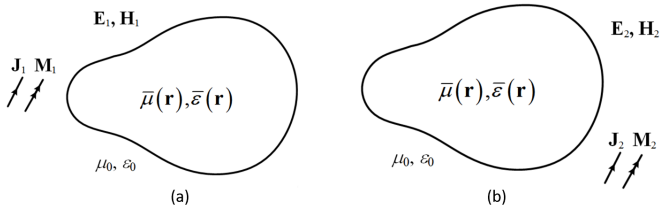


Figure 30.2: The geometry for proving reciprocity theorem. We perform two experiments on the same object or scatterer: (a) With sources $\mathbf{J}_1$ and $\mathbf{M}_1$ turned on, generating fields $\mathbf{E}_1$ and $\mathbf{H}_1$, and (b) With sources $\mathbf{J}_2$ and $\mathbf{M}_2$ turned on, generating fields $\mathbf{E}_2$ and $\mathbf{H}_2$. Magnetic currents, by convention, are denoted by double arrows.

Consider a general 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 30.2. This representation of the medium is quite general, and it can include dispersive and conductive media as well. It can represent complex terrain, or complicated electronic circuit structures in circuit boards or microchips, as well as complicated antenna structures.

We can do a Gedanken experiment[1] where a scatterer or an object is illuminated by fields from two sets of sources which are turned on and off consecutively. This is illustrated in Figure 30.2: When only $\mathbf{J}_1$ and $\mathbf{M}_1$ are turned on, they generate fields $\mathbf{E}_1$ and $\mathbf{H}_1$ in this medium. On the other hand, when only $\mathbf{J}_2$ and $\mathbf{M}_2$ are turned on, 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]

$$\nabla \times \mathbf{E}_1 = -j\omega\overline{\boldsymbol{\mu}} \cdot \mathbf{H}_1 - \mathbf{M}_1 \tag{30.1.1}$$

$$\nabla \times \mathbf{H}_1 = j\omega\overline{\boldsymbol{\varepsilon}} \cdot \mathbf{E}_1 + \mathbf{J}_1 \tag{30.1.2}$$

$$\nabla \times \mathbf{E}_2 = -j\omega\overline{\boldsymbol{\mu}} \cdot \mathbf{H}_2 - \mathbf{M}_2 \tag{30.1.3}$$

$$\nabla \times \mathbf{H}_2 = j\omega\overline{\boldsymbol{\varepsilon}} \cdot \mathbf{E}_2 + \mathbf{J}_2 \tag{30.1.4}$$

We would like to find a simplifying expression for the divergence of the following quantity,

$$\nabla \cdot (\mathbf{E}_1 \times \mathbf{H}_2) = \mathbf{H}_2 \cdot \nabla \times \mathbf{E}_1 - \mathbf{E}_1 \cdot \nabla \cdot \mathbf{H}_2 \tag{30.1.5}$$

so that the divergence theorem can be invoked. To this end, and from the above, we can show

[1] Thought experiment in German.

[2] The current sources are impressed currents so that they are immutable, and not changed by the environment they are immersed in [50, 172].

that (after left dot-multiply (30.1.1) with $\mathbf{H}_2$ and (30.1.4) with $\mathbf{E}_1$),

$$\mathbf{H}_2 \cdot \nabla \times \mathbf{E}_1 = -j\omega \mathbf{H}_2 \cdot \overline{\boldsymbol{\mu}} \cdot \mathbf{H}_1 - \mathbf{H}_2 \cdot \mathbf{M}_1 \tag{30.1.6}$$

$$\mathbf{E}_1 \cdot \nabla \times \mathbf{H}_2 = j\omega \mathbf{E}_1 \cdot \overline{\boldsymbol{\varepsilon}} \cdot \mathbf{E}_2 + \mathbf{E}_1 \cdot \mathbf{J}_2 \tag{30.1.7}$$

Then, using the above, and the following identity, we get the second equality in the following expression:

$$\begin{aligned} \nabla \cdot (\mathbf{E}_1 \times \mathbf{H}_2) &= \mathbf{H}_2 \cdot \nabla \times \mathbf{E}_1 - \mathbf{E}_1 \cdot \nabla \times \mathbf{H}_2 \\ &= -j\omega \mathbf{H}_2 \cdot \overline{\boldsymbol{\mu}} \cdot \mathbf{H}_1 - j\omega \mathbf{E}_1 \cdot \overline{\boldsymbol{\varepsilon}} \cdot \mathbf{E}_2 - \mathbf{H}_2 \cdot \mathbf{M}_1 - \mathbf{E}_1 \cdot \mathbf{J}_2 \end{aligned} \tag{30.1.8}$$

By the same token,

$$\nabla \cdot (\mathbf{E}_2 \times \mathbf{H}_1) = -j\omega \mathbf{H}_1 \cdot \overline{\boldsymbol{\mu}} \cdot \mathbf{H}_2 - j\omega \mathbf{E}_2 \cdot \overline{\boldsymbol{\varepsilon}} \cdot \mathbf{E}_1 - \mathbf{H}_1 \cdot \mathbf{M}_2 - \mathbf{E}_2 \cdot \mathbf{J}_1 \tag{30.1.9}$$

If one assumes that

$$\overline{\boldsymbol{\mu}} = \overline{\boldsymbol{\mu}}^t, \qquad \overline{\boldsymbol{\varepsilon}} = \overline{\boldsymbol{\varepsilon}}^t \tag{30.1.10}$$

or when the tensors are symmetric, then $\mathbf{H}_1 \cdot \overline{\boldsymbol{\mu}} \cdot \mathbf{H}_2 = \mathbf{H}_2 \cdot \overline{\boldsymbol{\mu}} \cdot \mathbf{H}_1$ 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 (30.1.8) and (30.1.9), one gets

$$\nabla \cdot (\mathbf{E}_1 \times \mathbf{H}_2 - \mathbf{E}_2 \times \mathbf{H}_1) = -\mathbf{H}_2 \cdot \mathbf{M}_1 - \mathbf{E}_1 \cdot \mathbf{J}_2 + \mathbf{H}_1 \cdot \mathbf{M}_2 + \mathbf{E}_2 \cdot \mathbf{J}_1 \tag{30.1.11}$$

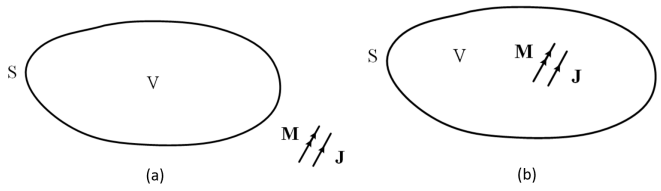


Figure 30.3: The geometry for proving reciprocity theorem when the surface S: (a) does not enclose the sources, and (b) encloses the sources. In the figure, the sources are supposed to be either $(\mathbf{M}_1, \mathbf{J}_1)$ producing fields $(\mathbf{E}_1, \mathbf{H}_1)$ or $(\mathbf{M}_2, \mathbf{J}_2)$ producing fields $(\mathbf{E}_2, \mathbf{H}_2)$.

Now, integrating (30.1.11) over a volume V bounded by a surface S, and invoking Gauss' divergence theorem, we have the reciprocity theorem that

$$\begin{aligned} &\oiint_S d\mathbf{S} \cdot (\mathbf{E}_1 \times \mathbf{H}_2 - \mathbf{E}_2 \times \mathbf{H}_1) \\ &\qquad = -\iiint_V dV [\mathbf{H}_2 \cdot \mathbf{M}_1 + \mathbf{E}_1 \cdot \mathbf{J}_2 - \mathbf{H}_1 \cdot \mathbf{M}_2 - \mathbf{E}_2 \cdot \mathbf{J}_1] \end{aligned} \tag{30.1.12}$$

[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 $\mathbf{a} \cdot \mathbf{b}$.

When the volume V contains no sources (see Figure 30.3), the reciprocity theorem reduces to

$$\oiint_S d\mathbf{S} \cdot (\mathbf{E}_1 \times \mathbf{H}_2 - \mathbf{E}_2 \times \mathbf{H}_1) = 0 \tag{30.1.13}$$

The above is also called Lorentz reciprocity theorem by some authors.[4]

Next, when the surface S contains all the sources (see Figure 30.3), then the right-hand side of (30.1.12) 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. Moreover, under the plane-wave approximation, $\omega\mu_0\mathbf{H}_2 = \boldsymbol{\beta} \times \mathbf{E}_2$, $\omega\mu_0\mathbf{H}_1 = \boldsymbol{\beta} \times \mathbf{E}_1$, then

$$\mathbf{E}_1 \times \mathbf{H}_2 \sim \mathbf{E}_1 \times (\boldsymbol{\beta} \times \mathbf{E}_2) = \mathbf{E}_1(\boldsymbol{\beta} \cdot \mathbf{E}_2) - \boldsymbol{\beta}(\mathbf{E}_1 \cdot \mathbf{E}_2) \tag{30.1.14}$$

$$\mathbf{E}_2 \times \mathbf{H}_1 \sim \mathbf{E}_2 \times (\boldsymbol{\beta} \times \mathbf{E}_1) = \mathbf{E}_2(\boldsymbol{\beta} \cdot \mathbf{E}_1) - \boldsymbol{\beta}(\mathbf{E}_2 \cdot \mathbf{E}_1) \tag{30.1.15}$$

But $\boldsymbol{\beta} \cdot \mathbf{E}_2 = \boldsymbol{\beta} \cdot \mathbf{E}_1 = 0$ in the far field and the $\boldsymbol{\beta}$ vectors are parallel to each other. Therefore, the two terms on the left-hand side of (30.1.12) 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 as the surface area S is growing larger and proportional to r^2.)

As a result, when $S \to \infty$, (30.1.12) can be rewritten simply as

$$\int_V dV[\mathbf{E}_2 \cdot \mathbf{J}_1 - \mathbf{H}_2 \cdot \mathbf{M}_1] = \int_V dV[\mathbf{E}_1 \cdot \mathbf{J}_2 - \mathbf{H}_1 \cdot \mathbf{M}_2] \tag{30.1.16}$$

The inner product symbol is often used to rewrite the above as

$$\langle \mathbf{E}_2, \mathbf{J}_1 \rangle - \langle \mathbf{H}_2, \mathbf{M}_1 \rangle = \langle \mathbf{E}_1, \mathbf{J}_2 \rangle - \langle \mathbf{H}_1, \mathbf{M}_2 \rangle \tag{30.1.17}$$

where the inner product $\langle \mathbf{A}, \mathbf{B} \rangle = \int_V dV \mathbf{A}(\mathbf{r}) \cdot \mathbf{B}(\mathbf{r})$.

The above inner product is also called **reaction**, a concept introduced by Rumsey [177]. The above is also called the **Rumsey reaction theorem**. Sometimes, the above is rewritten more succinctly and tersely as

$$\langle 2, 1 \rangle = \langle 1, 2 \rangle \tag{30.1.18}$$

where

$$\langle 2, 1 \rangle = \langle \mathbf{E}_2, \mathbf{J}_1 \rangle - \langle \mathbf{H}_2, \mathbf{M}_1 \rangle \tag{30.1.19}$$

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 Hertzian dipoles.

[4] Harrington, Time-Harmonic Electric Field [50].

30.2 Conditions for Reciprocity

It is seen that the above proof hinges on (30.1.10). In other words, the anisotropic medium has to be described by symmetric tensors. They include conductive media, but not gyrotropic media which is non-reciprocal. A ferrite biased by a magnetic field is often used in electronic circuits, and it corresponds to a gyrotropic, non-reciprocal medium.[5] Also, our starting equations (30.1.1) to (30.1.4) 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.

30.3 Application to a Two-Port Network and Circuit Theory

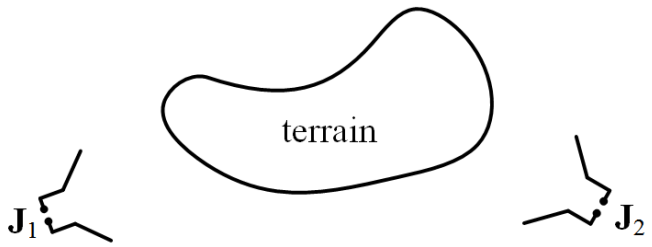


Figure 30.4: A geometry for proving the circuit relationship between two antennas using reciprocity theorem. Circuit relationship is possible when the ports of the antennas are small compared to wavelength.

The reciprocity theorem can be used to distill and condense the interaction 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.[6] In Figure 30.4, 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 antennas are made from reciprocal media, such as conductive media. Since the system is linear time invariant, it can be written as the interaction between two ports as in circuit theory as shown in Figure 30.5. Assuming that these two ports are small compared to wavelengths, then we can apply circuit concepts like potential theory by letting $\mathbf{E} = -\nabla\Phi$ in the neighborhood of the ports. Thus, we can define voltages and currents at these ports, and V-I relationships can be established in the manner of circuit theory.

[5]Non-reciprocal media are important for making isolators in microwave. Microwave signals can travel from Port 1 to Port 2, but not vice versa.

[6]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.

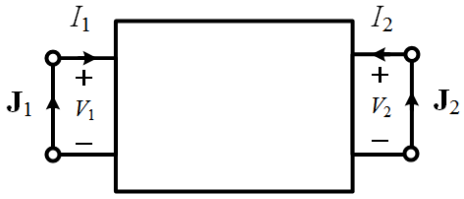


Figure 30.5: The interaction between two antennas in the far field of each other can be reduced to a circuit theory description since the input and output ports of the antennas are small compared to wavelength.

Focusing on a two-port network as shown in Figure 30.5, we have from circuit theory that [178]

$$\begin{bmatrix} V_1 \\ V_2 \end{bmatrix} = \begin{bmatrix} Z_{11} & Z_{12} \\ Z_{21} & Z_{22} \end{bmatrix} \begin{bmatrix} I_1 \\ I_2 \end{bmatrix} \tag{30.3.1}$$

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. 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 [7] 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 port. Consequently, the reaction

$$\langle \mathbf{E}_2, \mathbf{J}_1 \rangle = \int_V dV (\mathbf{E}_2 \cdot \mathbf{J}_1) = I_1 \int_{\text{Port 1}} \mathbf{E}_2 \cdot d\mathbf{l} = -I_1 V_1^{oc} \tag{30.3.2}$$

Even though port 1 is assumed to be off, the $\mathbf{J}_1$ is the impressed current to be used above is the $\mathbf{J}_1$ when port 1 is turned on. Given that the port is in the circuit physics regime, we assume the currents in 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 above, where $d\mathbf{l} = \hat{l} dl$. Since $\int_{\text{Port 1}} \mathbf{E}_2 \cdot d\mathbf{l} = -V_1^{oc}$, we have the last equality above.

We can repeat the derivation with port 2 to arrive at the reaction

$$\langle \mathbf{E}_1, \mathbf{J}_2 \rangle = I_2 \int_{\text{Port 2}} \mathbf{E}_1 \cdot d\mathbf{l} = -I_2 V_2^{oc} \tag{30.3.3}$$

[7] This is the same as the current source concept in circuit theory.

Reciprocity requires these two reactions to be equal, and hence,

$$I_1 V_1^{oc} = I_2 V_2^{oc}$$

But from (30.3.1), we can set the pertinent currents to zero to find these open circuit voltages. 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 area miles apart, as long as the transmit and receive ports of the linear system can be characterized by a linear relation expounded by (30.3.1), and the ports small enough compared to wavelength so that circuit theory prevails, 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.

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 [179], and the partial element equivalence circuit (PEEC) method [180]. Circuit theory is so simple that many people fall in love with it!

30.4 Voltage Sources in Electromagnetics

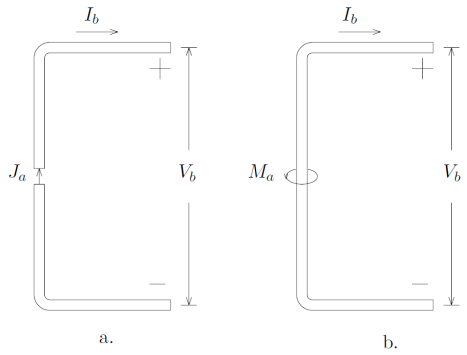


Figure 30.6: Two ways to model voltage sources: (i) An impressed current source $\mathbf{J}_a$ driving a very short antenna, and (ii) An impressed magnetic frill source (loop source) $\mathbf{M}_a$ driving a very short antenna (courtesy of Kong, Electromagnetic Wave Theory [32]).

In the above discussions, we have used current sources in reciprocity theorem to derive certain circuit concepts. Before we end this section, it is prudent to mention how voltage sources are

modeled in electromagnetic theory. The use of the impressed currents so that circuit concepts can be applied is shown in Figure 30.6. The antenna in (a) is driven by a current source. But a magnetic current can be used as a voltage source in circuit theory as shown by Figure 30.6b. By using duality concept, an electric field has to curl around a magnetic current just in 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.[8]

30.5 Hind Sight

The proof of reciprocity theorem for Maxwell's equations is very deeply related to the symmetry of the operator involved. We can see this from linear algebra. Given a matrix equation driven by two different sources $\mathbf{b}_1$ and $\mathbf{b}_2$ with solutions $\mathbf{x}_1$ and $\mathbf{x}_2$, they can be written succinctly as

$$\overline{\mathbf{A}} \cdot \mathbf{x}_1 = \mathbf{b}_1 \tag{30.5.1}$$

$$\overline{\mathbf{A}} \cdot \mathbf{x}_2 = \mathbf{b}_2 \tag{30.5.2}$$

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

$$\mathbf{x}_2^t \cdot \overline{\mathbf{A}} \cdot \mathbf{x}_1 = \mathbf{x}_2^t \cdot \mathbf{b}_1 \tag{30.5.3}$$

$$\mathbf{x}_1^t \cdot \overline{\mathbf{A}} \cdot \mathbf{x}_2 = \mathbf{x}_1^t \cdot \mathbf{b}_2 \tag{30.5.4}$$

If $\overline{\mathbf{A}}$ is symmetric, the left-hand side of both equations are equal to each other.[9] Therefore, we can equate their right-hand side to arrive at

$$\mathbf{x}_2^t \cdot \mathbf{b}_1 = \mathbf{x}_1^t \cdot \mathbf{b}_2 \tag{30.5.5}$$

The above is analogous to the statement of the reciprocity theorem which is

$$\langle \mathbf{E}_2, \mathbf{J}_1 \rangle = \langle \mathbf{E}_1, \mathbf{J}_2 \rangle \tag{30.5.6}$$

where the reaction inner product, as mentioned before, is $\langle \mathbf{E}_i, \mathbf{J}_j \rangle = \int_V \mathbf{E}_i(\mathbf{r}) \cdot \mathbf{J}_j(\mathbf{r})$. 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.[10] So if the operators in Maxwell's equations are symmetrical, then reciprocity theorem applies.

[8] More can be found in Jordain and Balmain, Electromagnetic Waves and Radiation Systems [54].

[9] This can be easily proven by taking the transpose of a scalar, and taking the transpose of the product of matrices.

[10] Such spaces are called Hilbert space.

30.6 Transmit and Receive Patterns of an Antennna

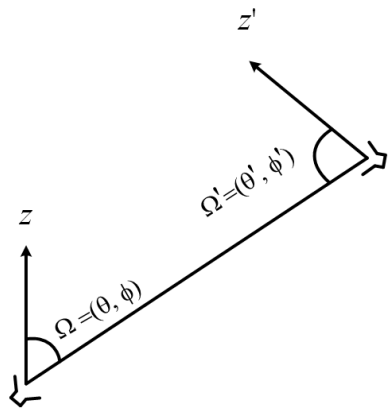


Figure 30.7: The schematic diagram for studying the transmit and receive properties of antennas. The two antennas are assumed to be identical, and each switches between transmit and receive modes in this study.

Reciprocity also implies that the transmit and receive properties of an antenna is similar to each other. The transmit property of an antenna is governed by the gain function, while its receive property is governed by the effective area or aperture. The effective aperture is also a function of angle of the incident wave with respect to to the antenna. The gain function of an antenna is related to its effective aperture by a constant as we shall argue.

Consider an antenna in the transmit mode. Then the time-average radiation power density that it will yield around the antenna, in accordance to (25.3.5), is[11]

$$\langle S_{\rm rad}\rangle = \frac{P_t}{4\pi r^2} G(\theta,\phi) \tag{30.6.1}$$

where P_t is the total power radiated by the transmit antenna, and $G(\theta,\phi)$ is its directive gain pattern or function. It is to be noted that in the above $\int_{4\pi} d\Omega G(\theta,\phi) = 4\pi$. The above is valid when the antenna is lossless.

Effective Gain versus Directive Gain

At this juncture, it is important to introduce the concept of effective gain versus directive gain. The effective gain, also called the power gain, is

$$G_e(\theta,\phi) = f_e G(\theta,\phi) \tag{30.6.2}$$

where f_e is the efficiency of the antenna, a factor less than 1. It accounts for the fact that not all power pumped into the antenna is delivered as radiated power. For instance, power

[11]The author is indebted to inspiration from E. Kudeki of UIUC for this part of the lecture notes [139].

can be lost in the circuits and mismatch of the antenna. Therefore, the correct formula the radiated power density is

$$\langle S_{\text{rad}} \rangle = \frac{P_t}{4\pi r^2} G_e(\theta, \phi) = f_e \frac{P_t}{4\pi r^2} G(\theta, \phi) \tag{30.6.3}$$

This radiated power resembles that of a plane wave when one is far away from the transmitter. Thus if a receive antenna is placed in the far-field of the transmit antenna, it will see this power density as coming from a plane wave. Thus the receive antenna will see an incident power density as

$$\langle S_{\text{inc}} \rangle = \langle S_{\text{rad}} \rangle = \frac{P_t}{4\pi r^2} G_e(\theta, \phi) \tag{30.6.4}$$

Effective Aperture

The effective area or the aperture of a receive antenna is used to characterize its receive property. The power received by such an antenna is then, by using the concept of effective aperture expounded in (26.1.23)

$$P_r = \langle S_{\text{inc}} \rangle A_e(\theta', \phi') \tag{30.6.5}$$

where (θ', ϕ') are the angles at which the plane wave is incident upon the receiving antenna (see Figure 30.7). Combining the above formulas (30.6.4) and (30.6.5), we have

$$P_r = \frac{P_t}{4\pi r^2} G_e(\theta, \phi) A_e(\theta', \phi') \tag{30.6.6}$$

Now assuming that the transmit and receive antennas are identical. Next, we swap their roles of transmit and receive, and also the circuitries involved in driving the transmit and receive antennas. Then,

$$P_r = \frac{P_t}{4\pi r^2} G_e(\theta', \phi') A_e(\theta, \phi) \tag{30.6.7}$$

We also assume that the receive antenna, that now acts as the transmit antenna is transmitting in the (θ', ϕ') direction. Moreover, the transmit antenna, that now acts as the receive antenna is receiving in the (θ, ϕ) direction (see Figure 30.7).

By reciprocity, these two powers are the same, because $Z_{12} = Z_{21}$. Furthermore, since these two antennas are identical, $Z_{11} = Z_{22}$. So by swapping the transmit and receive electronics, the power transmitted and received will not change. A simple transmit-receive circuit diagram is shown in Figure 30.8

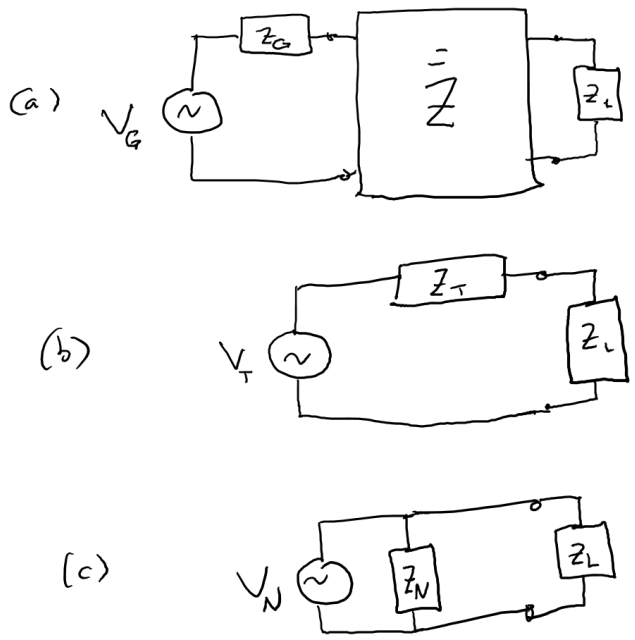


Figure 30.8: The schematic of the circuit for a transmit-receive antenna pair. Because the mutual interaction between the two antennas can be described by the impedance matrix $\overline{\mathbf{Z}}$, circuit theory can be applied to model their mutual interaction as indicated in (a). Moreover, at the receive end, one can even further simplify the circuit by using a Thevenin equivalence (b), or a Norton equivalence (c).

Consequently, we conclude that

$$G_e(\theta,\phi)A_e(\theta',\phi') = G_e(\theta',\phi')A_e(\theta,\phi) \tag{30.6.8}$$

The above implies that

$$\frac{A_e(\theta,\phi)}{G_e(\theta,\phi)} = \frac{A_e(\theta',\phi')}{G_e(\theta',\phi')} = \text{constant} \tag{30.6.9}$$

The above Gedanken experiment is carried out for arbitrary angles. Therefore, the constant is independent of angles. Moreover, this constant is independent of the size, shape, and efficiency of the antenna, as we have not stipulated their shapes, sizes, and efficiency in the above discussion.

To find this constant in (30.6.9), one can repeat the above for a Hertzian dipole, wherein the mathematics of calculating P_r and P_t is a lot simpler. This constant is found to be

$\lambda^2/(4\pi)$.[12] Therefore, an interesting relationship between the effective aperture (or area) and the directive gain function is that

$$A_e(\theta, \phi) = \frac{\lambda^2}{4\pi} G_e(\theta, \phi) \tag{30.6.10}$$

One amusing point about the above formula is that the effective aperture, say of a Hertzian dipole, becomes very large when the frequency is low, or the wavelength is very long. Of course, this cannot be physically true, and I will let you meditate on this paradox and muse over this point.

[12] See Kong [32][p. 700]. The derivation is for 100% efficient antenna. A thermal equilibrium argument is used in [139] and Wikipedia [140] as well.

Lecture 31

Equivalence Theorems, Huygens' Principle

Electromagnetic equivalence theorems are useful for simplifying solutions to many problems. Also, they offer physical insight into the behaviour of electromagnetic fields of a Maxwellian system. They are closely related to the uniqueness theorem and 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 [32, 50, 54, 65, 181]. Some authors also call it Love's equivalence principles [182] and credit has been given to Schelkunoff as well [172].

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 those in electromagnetics. One thing they have in common is that they are all derived by using Gedanken experiment (thought experiment) , involving no math. But in this lecture, we will show that electromagnetic equivalence theorems are also derivable using mathematics, albeit with more work.

31.1 Equivalence Theorems or Equivalence Principles

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.

31.1.1 Inside-Out Case

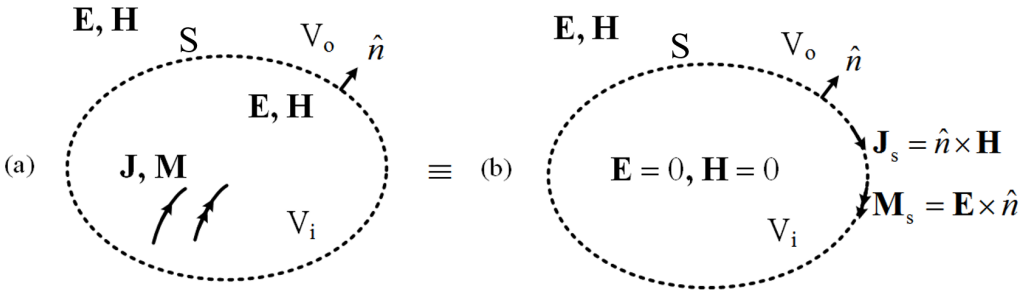


Figure 31.1: The inside-out problem where the two cases in (a) and (b) are equivalent. In (b), equivalence currents are impressed on the surface S so as to produce the same fields outside in V_o in both cases, cases (a) and (b).

In this case, we let $\mathbf{J}$ and $\mathbf{M}$ be the time-harmonic radiating sources inside a surface S radiating into a region $V = V_o \cup V_i$. They produce $\mathbf{E}$ and $\mathbf{H}$ everywhere. We can construct an equivalence problem 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.

Apparently, in case (b) in Figure 31.1, 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 for field discontinuities are satisfied on S. Clearly, surface currents needed for these field discontinuities. By virtue of the boundary conditions and the jump conditions in electromagnetics, these surface currents to be impressed on S are

$$\mathbf{J}_s = \hat{n} \times \mathbf{H}, \qquad \mathbf{M}_s = \mathbf{E} \times \hat{n} \tag{31.1.1}$$

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, when fictitious magnetic current is added to Faraday's law in Section 5.3 for mathematical symmetry, 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 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 $\hat{\times}\mathbf{H}$.

Consequetly, we can convince ourselves that $\hat{n} \times \mathbf{H}$ and $\mathbf{E} \times \hat{n}$ just outside S in both cases are the same. Furthermore, we are persuaded that the above is a bona fide solution to Maxwell's equations.

- The boundary conditions on the surface S satisfy the boundary conditions required of Maxwell's equations.

- By the uniqueness theorem, only the equality of one of them $\mathbf{E} \times \hat{n}$, or $\hat{n} \times \mathbf{H}$ ons S, will guarantee that $\mathbf{E}$ and $\mathbf{H}$ outside S are the same in both cases (a) and (b).

The fact that these equivalence currents generate zero fields inside S is known as the **extinction theorem**. This equivalence theorem can also be proved mathematically, as shall be shown.

31.1.2 Outside-in Case

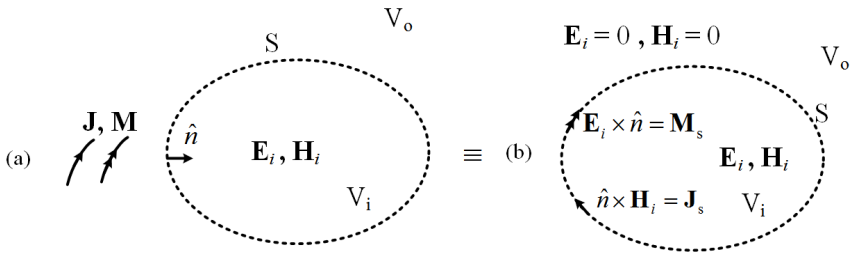


Figure 31.2: The outside-in problem where equivalence currents are impressed on the surface S to produce the same fields inside in both cases.

Similar to before, we find an equivalence problem (b) where the fields inside S in V_i is 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 are needed to account for these discontinuities.

Then by the uniqueness theorem,[1] 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.

31.1.3 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 31.3:

[1] We can add infinitesimal loss to ensure that uniqueness theorem is satisfied in this enclosed volume V_i.

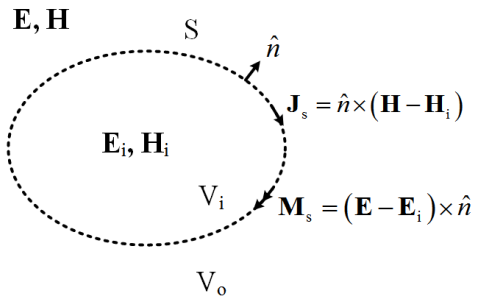


Figure 31.3: The general case where the fields are non-zero both inside and outside the surface S. Equivalence currents are needed on the surface S to support the jump discontinuities in the fields.

31.2 Electric Current on a PEC

Using the equivalence problems in the previous section, we can derive other corollaries of equivalence theorems. We shall show them next.

First, from reciprocity theorem, it is quite easy to prove that an impressed current on the PEC cannot radiate. We can start with the inside-out equivalence problem. Since the fields inside S is zero for the inside-out problem, using a Gedanken experiment, one can insert an PEC object inside S without disturbing the fields $\mathbf{E}$ and $\mathbf{H}$ outside since the field is zero inside S. As the PEC object grows to snugly fit the surface S, then the electric current $\mathbf{J}_s = \hat{n} \times \mathbf{H}$ does not radiate by reciprocity. Only one of the two currents is radiating, namely, the magnetic current $\mathbf{M}_s = \mathbf{E} \times \hat{n}$ is radiating, and $\mathbf{J}_s$ in Figure 31.4 can be removed. This is commensurate with the uniqueness theorem that only the knowledge of $\mathbf{E} \times \hat{n}$ is needed to uniquely determine the fields outside S.

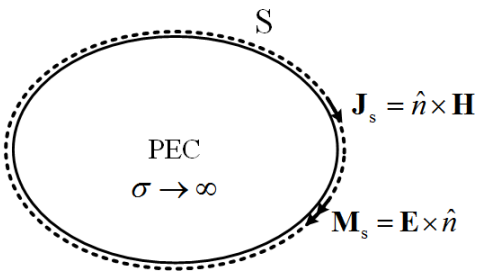


Figure 31.4: On a PEC surface, only one of the two currents is needed since an electric current does not radiate on a PEC surface.

31.3 Magnetic Current on a PMC

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 electric current $\mathbf{J}_s = \hat{n} \times \mathbf{H}$ radiates, and the 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.

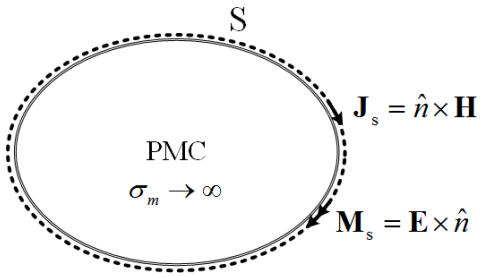


Figure 31.5: Similarly, on a PMC surface, only an electric current is needed to produce the field outside the surface S.

31.4 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 [183]. But the mathematical expression of this idea was due to George Green [2] 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 discuss the scalar wave case first, followed by the electromagnetic vector wave case.

[2] George Green (1793-1841) was self educated and the son of a miller in Nottingham, England [184].

31.4.1 Scalar Waves Case

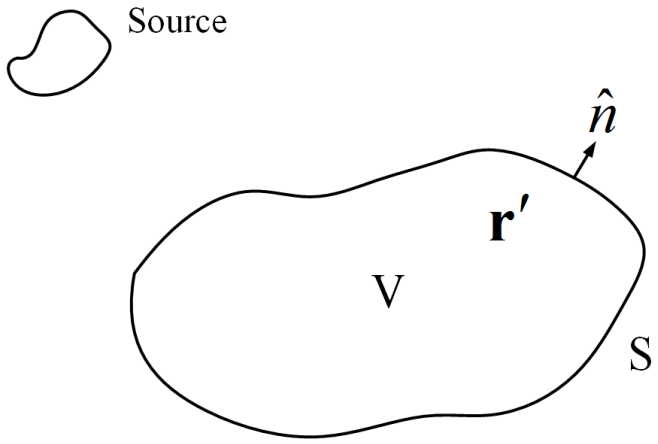


Figure 31.6: The geometry for deriving Huygens' principle for scalar wave equation.

For a $\psi(\mathbf{r})$ that satisfies the scalar wave equation

$$(\nabla^2 + k^2)\,\psi(\mathbf{r}) = 0, \tag{31.4.1}$$

the corresponding scalar Green's function $g(\mathbf{r}, \mathbf{r}')$ satisfies

$$(\nabla^2 + k^2)\,g(\mathbf{r}, \mathbf{r}') = -\delta(\mathbf{r} - \mathbf{r}'). \tag{31.4.2}$$

Next, we multiply (31.4.1) by $g(\mathbf{r}, \mathbf{r}')$ and (31.4.2) by $\psi(\mathbf{r})$. And then, we subtract the resultant equations and integrating over a volume V as shown in Figure 31.6. There are two cases to consider: when $\mathbf{r}'$ is in V, or when $\mathbf{r}'$ is outside V. Thus, we have

$$\left.\begin{array}{ll} \text{if } \mathbf{r}' \in V, & \psi(\mathbf{r}') \\ \text{if } \mathbf{r}' \notin V, & 0 \end{array}\right\} = \int_V d\mathbf{r}\,[g(\mathbf{r}, \mathbf{r}')\nabla^2\psi(\mathbf{r}) - \psi(\mathbf{r})\nabla^2 g(\mathbf{r}, \mathbf{r}')], \tag{31.4.3}$$

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 left-hand side of (31.4.3) can be rewritten using Gauss' divergence theorem, giving[3]

$$\left.\begin{array}{ll} \text{if } \mathbf{r}' \in V, & \psi(\mathbf{r}') \\ \text{if } \mathbf{r}' \notin V, & 0 \end{array}\right\} = \oint_S dS\,\hat{n} \cdot [g(\mathbf{r}, \mathbf{r}')\nabla\psi(\mathbf{r}) - \psi(\mathbf{r})\nabla g(\mathbf{r}, \mathbf{r}')], \tag{31.4.4}$$

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

[3]The equivalence of the volume integral in (31.4.3) to the surface integral in (31.4.4) is also known as Green's theorem [89].

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. The first term on the right-hand side radiates via the Green's function $g(\mathbf{r}, \mathbf{r}')$. 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, the second term corresponds to sources that radiate like dipoles pointing normally to the surface S. These sources ara also called double layer (or dipole layer) sources. These terminologies are prevalent in acoustics. The above mathematical expression also embodies the extinction theorem that says if $\mathbf{r}'$ is outside V, the left-hand side evaluates to zero.

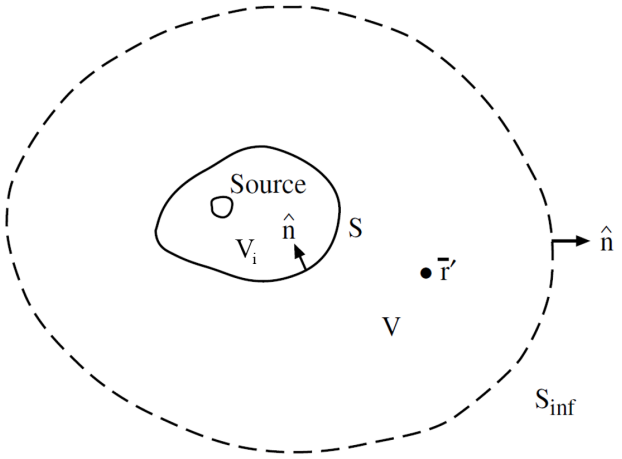


Figure 31.7: The geometry for deriving Huygens' principle for scalar wave. The radiation from the source can be replaced by equivalence sources on the surface S, and the field outside S can be calculated using (31.4.4). Also, the field is zero inside S from (31.4.4). This is the extinction theorem.

If the volume V is bounded by S and $S_{\rm inf}$ as shown in Figure 31.7, then the surface integral in (31.4.4) should include an integral over $S_{\rm inf}$. But when $S_{\rm inf} \to \infty$, all fields look like plane wave, and $\nabla \to -\hat{r}jk$ on $S_{\rm inf}$. Furthermore, $g(\mathbf{r}-\mathbf{r}') \sim O(1/r)$,[4] when $r \to \infty$, and $\psi(\mathbf{r}) \sim O(1/r)$, when $r \to \infty$, if $\psi(\mathbf{r})$ is due to a source of finite extent. Then, the integral over $S_{\rm inf}$ in (31.4.4) vanishes, and (31.4.4) is valid for the case shown in Figure 31.7 as well but with the surface integral over surface S only. Here, the field outside S at $\mathbf{r}'$ is expressible in terms of the field on S. This is similar to the inside-out equivalence principle we have discussed previously Section 31.1.1, albeit this is for scalar wave case.

Notice that in deriving (31.4.4), $g(\mathbf{r}, \mathbf{r}')$ has only to satisfy (31.4.2) 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

[4]The symbol "O" means "of the order."

that $g(\mathbf{r}, \mathbf{r}') = 0$ for $\mathbf{r} \in S$, then (31.4.4) becomes

$$\psi(\mathbf{r}') = -\oint_S dS\, \psi(\mathbf{r})\, \hat{n} \cdot \nabla g(\mathbf{r}, \mathbf{r}'), \qquad \mathbf{r}' \in V. \tag{31.4.5}$$

On the other hand, if require additionally that $g(\mathbf{r}, \mathbf{r}')$ satisfies (31.4.2) with the boundary condition $\hat{n} \cdot \nabla g(\mathbf{r}, \mathbf{r}') = 0$ for $\mathbf{r} \in S$, then (31.4.4) becomes

$$\psi(\mathbf{r}') = \oint_S dS\, g(\mathbf{r}, \mathbf{r}')\, \hat{n} \cdot \nabla \psi(\mathbf{r}), \qquad \mathbf{r}' \in V. \tag{31.4.6}$$

Equations (31.4.4), (31.4.5), and (31.4.6) 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 (31.4.5) and (31.4.6) 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 31.2 and 31.3. (Note that in the above derivation, k^2 could be a function of position as well.)

31.4.2 Electromagnetic Waves Case

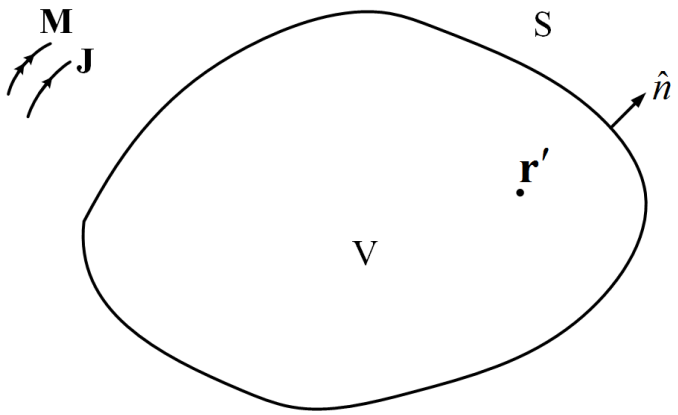


Figure 31.8: Derivation of the Huygens' principle for the electromagnetic case. One only needs to know the surface fields on surface S in order to determine the field at $\mathbf{r}'$ inside V.

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 laborious and complicated vector algebra!

In a source-free homogeneous region, an electromagnetic wave satisfies the vector wave equation

$$\nabla \times \nabla \times \mathbf{E}(\mathbf{r}) - k^2\, \mathbf{E}(\mathbf{r}) = 0. \tag{31.4.7}$$

The analogue of the scalar Green's function for the scalar wave equation is the dyadic Green's function for the electromagnetic wave case [1, 32, 185, 186]. Moreover, the dyadic Green's function satisfies the equation[5]

$$\nabla \times \nabla \times \overline{\mathbf{G}}(\mathbf{r}, \mathbf{r}') - k^2\, \overline{\mathbf{G}}(\mathbf{r}, \mathbf{r}') = \overline{\mathbf{I}}\, \delta(\mathbf{r} - \mathbf{r}'). \tag{31.4.8}$$

It can be shown by direct back substitution that the dyadic Green's function in free space is [186]

$$\overline{\mathbf{G}}(\mathbf{r}, \mathbf{r}') = \left(\overline{\mathbf{I}} + \frac{\nabla\nabla}{k^2}\right) g(\mathbf{r} - \mathbf{r}') \tag{31.4.9}$$

The above allows us to derive the vector Green's theorem [1, 32, 185].

Then, after post-multiplying (31.4.7) by $\overline{\mathbf{G}}(\mathbf{r}, \mathbf{r}')$, pre-multiplying (31.4.8) by $\mathbf{E}(\mathbf{r})$, subtracting the resultant equations and integrating the difference over volume V, considering two cases as we did for the scalar wave case, we have

$$\left.\begin{array}{ll} \text{if } \mathbf{r}' \in V, & \mathbf{E}(\mathbf{r}') \\ \text{if } \mathbf{r}' \notin V, & 0 \end{array}\right\} = \int\limits_V dV\, \left[\mathbf{E}(\mathbf{r}) \cdot \nabla \times \nabla \times \overline{\mathbf{G}}(\mathbf{r}, \mathbf{r}') - \nabla \times \nabla \times \mathbf{E}(\mathbf{r}) \cdot \overline{\mathbf{G}}(\mathbf{r}, \mathbf{r}')\right]. \tag{31.4.10}$$

Next, using the vector identity that[6]

$$\begin{aligned} -\nabla \cdot \left[\mathbf{E}(\mathbf{r}) \times \nabla \times \overline{\mathbf{G}}(\mathbf{r}, \mathbf{r}') + \nabla \times \mathbf{E}(\mathbf{r}) \times \overline{\mathbf{G}}(\mathbf{r}, \mathbf{r}')\right] \\ = \mathbf{E}(\mathbf{r}) \cdot \nabla \times \nabla \times \overline{\mathbf{G}}(\mathbf{r}, \mathbf{r}') - \nabla \times \nabla \times \mathbf{E}(\mathbf{r}) \cdot \overline{\mathbf{G}}(\mathbf{r}, \mathbf{r}'), \end{aligned} \tag{31.4.11}$$

then the integrand of (31.4.10) can be written as a divergence. With the help of Gauss' divergence theorem, it can be written as

$$\begin{aligned} \left.\begin{array}{ll} \text{if } \mathbf{r}' \in V, & \mathbf{E}(\mathbf{r}') \\ \text{if } \mathbf{r}' \notin V, & 0 \end{array}\right\} &= -\oint\limits_S dS\, \hat{n} \cdot \left[\mathbf{E}(\mathbf{r}) \times \nabla \times \overline{\mathbf{G}}(\mathbf{r}, \mathbf{r}') + \nabla \times \mathbf{E}(\mathbf{r}) \times \overline{\mathbf{G}}(\mathbf{r}, \mathbf{r}')\right] \\ &= -\oint\limits_S dS\, \left[-\mathbf{E}(\mathbf{r}) \times \hat{n} \cdot \nabla \times \overline{\mathbf{G}}(\mathbf{r}, \mathbf{r}') + i\omega\mu\, \hat{n} \times \mathbf{H}(\mathbf{r}) \cdot \overline{\mathbf{G}}(\mathbf{r}, \mathbf{r}')\right]. \end{aligned} \tag{31.4.12}$$

[5] 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.

[6] 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 (31.4.11) act on an aribitrary constant vector to convert the dyad into a vector before applying this identity. The equality of the volume integral in (31.4.10) to the surface integral in (31.4.12) is also known as the vector Green's theorem [32, 185]. Earlier form of this theorem was known as Franz formula [187].

The above is just the vector analogue of (31.4.4). We have used the cyclic relation of dot and cross products to rewrite the last expression. Since $\hat{n} \times \mathbf{E}$ and $\hat{n} \times \mathbf{H}$ are associated with surface magnetic current $\mathbf{M}_s$ and surface electric current $\mathbf{J}_s$, respectively, the above can be thought of having these equivalence surface currents radiating via the dyadic Green's function. Again, notice that (31.4.12) is derived via the use of (31.4.8), 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. The above is similar to the outside-in equivalence theorem we have derived in Section 31.1.2 using a Gedanken experiment. Now we have a mathematical derivation of the same theorem.

Now, 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 radiating in the presence of a PEC surface. Then (31.4.12) becomes

$$\mathbf{E}(\mathbf{r}') = -\oint_S dS\, \hat{n} \times \mathbf{E}(\mathbf{r}) \cdot \nabla \times \overline{\mathbf{G}}(\mathbf{r}, \mathbf{r}'), \qquad \mathbf{r}' \in V \tag{31.4.13}$$

for 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 (31.4.12) is zero. On the other hand, if we require that $\hat{n} \times \nabla \times \overline{\mathbf{G}}(\mathbf{r}, \mathbf{r}') = 0$ for $\mathbf{r} \in S$, then (31.4.12) becomes

$$\mathbf{E}(\mathbf{r}') = -i\omega\mu \oint_S dS\, \hat{n} \times \mathbf{H}(\mathbf{r}) \cdot \overline{\mathbf{G}}(\mathbf{r}, \mathbf{r}'), \qquad \mathbf{r}' \in V \tag{31.4.14}$$

Equations (31.4.13) and (31.4.14) 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 in the previous sections, Sections 31.2 and 31.3.

The dyadic Green's functions in (31.4.13) and (31.4.14) are for a closed cavity since boundary conditions are imposed on S for them. But the dyadic Green's function for an unbounded, homogeneous medium, given in (31.4.10) can be written as

$$\overline{\mathbf{G}}(\mathbf{r}, \mathbf{r}') = \frac{1}{k^2}[\nabla \times \nabla \times \bar{\mathbf{I}}\, g(\mathbf{r} - \mathbf{r}') - \bar{\mathbf{I}}\, \delta(\mathbf{r} - \mathbf{r}')], \tag{31.4.15}$$

$$\nabla \times \overline{\mathbf{G}}(\mathbf{r}, \mathbf{r}') = \nabla \times \bar{\mathbf{I}}\, g(\mathbf{r} - \mathbf{r}'). \tag{31.4.16}$$

Then, (31.4.12), for $\mathbf{r}' \in V$ and $\mathbf{r}' \neq \mathbf{r}$, becomes

$$\mathbf{E}(\mathbf{r}') = -\nabla' \times \oint_S dS\, g(\mathbf{r} - \mathbf{r}')\, \hat{n} \times \mathbf{E}(\mathbf{r}) + \frac{1}{i\omega\epsilon} \nabla' \times \nabla' \times \oint_S dS\, g(\mathbf{r} - \mathbf{r}')\, \hat{n} \times \mathbf{H}(\mathbf{r}). \tag{31.4.17}$$

The above can be applied to the geometry in Figure 31.7 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 (31.4.4). Then, (31.4.17) 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 31.1.1). 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. Hence, this is also related to the uniqueness theorem.

Lecture 32

Shielding, Image Theory

The physics 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, EMC/EMI become an increasingly challenging issue. These problems have to be solved using heuristics with a high dosage of physical insight.

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 quite different. Moreover, since there are no magnetic charges around, the shielding of magnetic field is quite 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 imping field, until inside the conductive medium $\mathbf{E} = 0$. In this case, electric current ceases to flow in the conductor.

In other words, when the field reaches the quiescent state, the charges 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. Figure 32.1 shows the static electric field, in the quiescent state, between two conductors (even though they are not PECs), and the electric field has to be normal to the conductor surfaces. Moreover, since $\mathbf{E} = 0$ inside the conductor, $\nabla\Phi = 0$ implying that the potential is a constant inside a conductor at equilibrium.

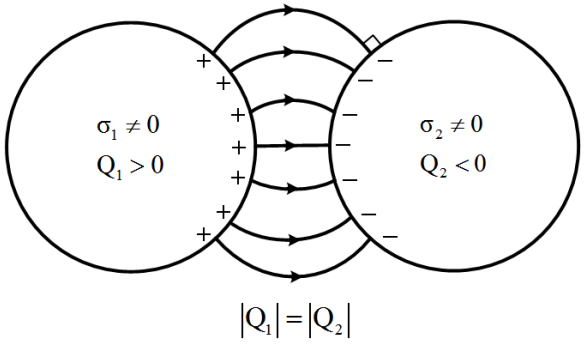


Figure 32.1: The objects can just be conductors, and in the quiescent state (static state), the tangential electric field will be zero on their surfaces. Also, $\mathbf{E} = 0$ inside the conductor, or $\nabla\Phi = 0$, or Φ is a constant.

32.1.2 Relaxation Time

The time it takes for the charges to move around until they reach their quiescent distribution such that $\mathbf{E}(t) = 0$ 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 32.2). It can be proven that this relaxation time is related to ε/σ, but the proof is beyond the scope of this course and 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.

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 32.3) [188].

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 32.4. 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

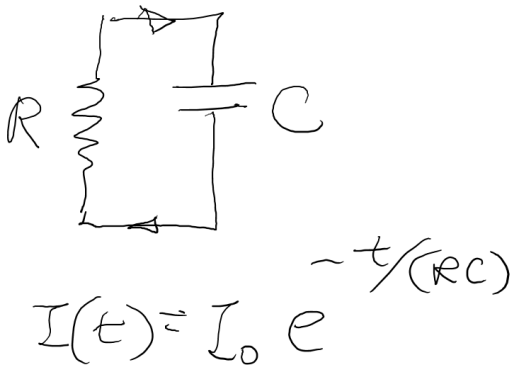


Figure 32.2: The relaxation (or disappearance of accumulated charges) in a conductive object is similar to the relaxation of charges from a charged capacitance in an RC circuit as shown.

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 large the penetration depth.

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 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 chain of events if it is true. Moreover, it will generate infinite energy in the system, which is not physical. Hence, the only 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 electric fields from outside are time varying. In other words, the relaxation time ε/σ is zero. As a consequence, the time-varying electric field $\mathbf{E}$ is always zero inside PEC, and therefore, $\hat{n} \times \mathbf{E} = 0$ on the surface of the PEC, even for time-varying fields.

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 and has a lot of symmetry. These closed form solutions in turn

Figure 32.3: Faraday cage demonstration on volunteers in the Palais de la Découverte in Paris (courtesy of Wikipedia). When the cage is grounded, charges will surge from the ground to the cage surface so as to make the field inside the cage zero, or the potential is constant. Therefore, a grounded Faraday cage effectively shields the external fields from entering the cage.

offer physical insight into the problems. This theory or method is also discussed in many textbooks [1, 43, 54, 65, 79, 181, 189].

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.

By the principle of linear superposition, the tangential static electric field can be canceled by putting an image charge of opposite sign at the mirror location of the original charge. This is shown in Figure 32.5. 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 32.6. 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

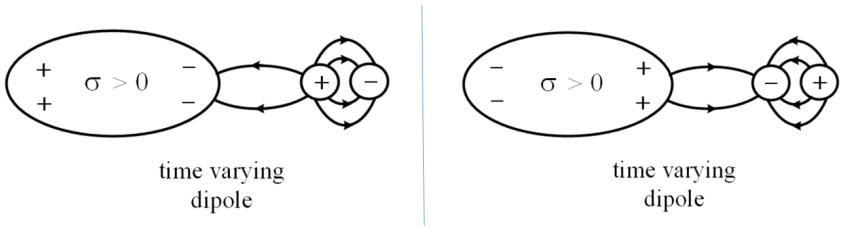


Figure 32.4: If the source that induces the charges on the conductor is time varying, the current in the conductor is always nonzero so that the charges can move around to respond to the external time-varying charges. The two figures above show the orientation of the charges for two snap-shot in time. In other words, a time-varying field can penetrate the conductor to approximately within a skin-depth $\delta = \sqrt{2/(\omega\mu\sigma)}$.

charge. Using 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 32.7.

If this electric dipole is a Hertzian dipole whose field is time-varying, then one needs a PEC half-space to shield out the electric field. Also, the image charges will follow the original dipole charges instantaneously. Then the image theory for static electric dipoles over a half-space still holds true if the dipoles now become Hertzian dipoles, but the surface will have to be a PEC surface so that the fields can be shielded out instantaneously.

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, repels the magnetic field from the conductive medium.[1]

Now, consider a static magnetic field penetrating into a perfect electric conductor, an 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 32.8).

[1]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.

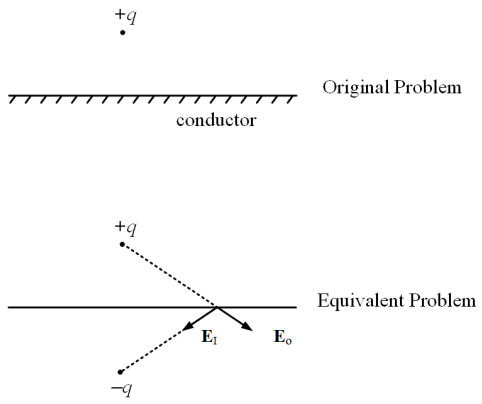


Figure 32.5: The use of image theory to solve the BVP of a point charge on top of a conductor. The boundary condition is that $\hat{n} \times \mathbf{E} = 0$ on the conductor surface. By placing a negative charge with respect to the original charge, by the principle of linear superposition, both of them produce a total field with no tangential component at the interface.

Now, assuming 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 32.8. 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 32.9. 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. Hence, a dipolar bar magnet can be levitated by a superconductor when this magnet is placed closed to it. This is also known as the Meissner effect [190], which is shown in Figure 32.10.

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 32.9.

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 a PMC surface.

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 32.11 [192]. The mushroom structure

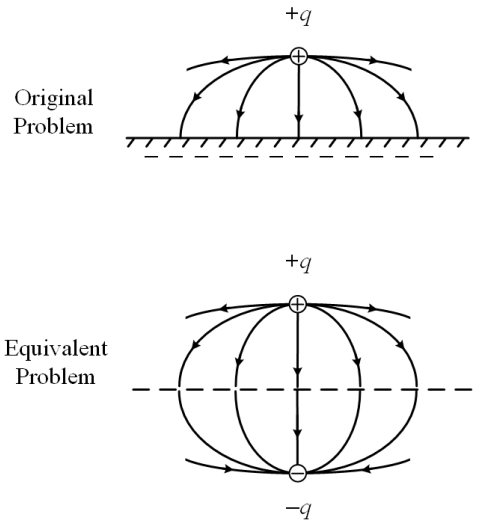


Figure 32.6: By image theory, the total electric of the original problem and the equivalent problem when we add the total electric field due to the original charge and the image charge.

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, because with no surface electric current on this surface, the tangential magnetic field is small, the hallmark of a good magnetic conductor, by the duality principle.

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 32.12. One can go through Gedanken experiments and verify that the reflection rules are as shown in the figure.

32.2.4 Multiple Images

For the geometry shown in Figure 32.13, 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 $\mathbf{E}$ and $\mathbf{H}$ fields. Again, one can repeat the above exercise for magnetic charges, magnetic dipoles, and PMC surfaces.

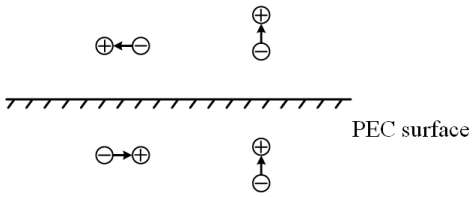


Figure 32.7: By image theory, on a conductor surface, a horizontal static dipole reflects to one of opposite polarity, while a static vertical dipole reflects to one of the same polarity. If the dipoles are time-varying, then a PEC will have a same reflection rule.

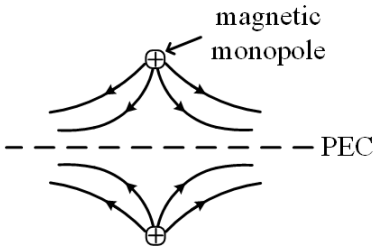


Figure 32.8: On a PEC surface, the requisite boundary condition is $\hat{n} \cdot \mathbf{B} = 0$. Hence, a magnetic monopole on top of a PEC surface will have magnetic field distributed as shown. By image theory, such a field distribution can be obtained by adding a magnetic monopole of the same polarity at its image point.

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 32.14.[2] A charge of $+Q$ reflects to a charge of $-Q_I$ inside the sphere. 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.

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 [35]. It can also be extended to multiple interfaces. But image theory cannot be used for the electrodynamic case due to the different speed of light in different media, giving rise to different retardation effects.

[2]This is worked out in p. 48 and p. 49, Ramo et al [31].

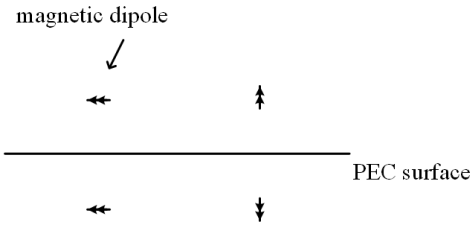


Figure 32.9: Using the rule of how magnetic monopole reflects itself on a PEC surface, the reflection rules for magnetic dipoles can be ascertained. Magnetic dipoles are often denoted by double arrows.

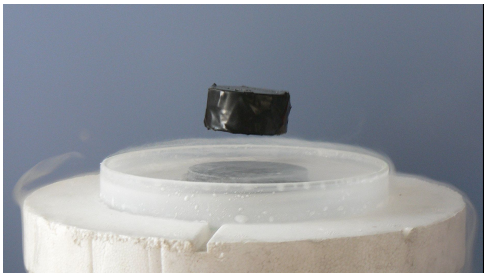


Figure 32.10: On a PEC (superconducting) surface, a vertical magnetic dipole (formed by a small permanent bar magnet here) reflects to one of opposite polarity. Hence, the magnetic dipoles repel each other displaying the Meissner effect. The magnet, because of the repulsive force from its image, levitates above the superconductor (courtesy of Wikipedia [191]).

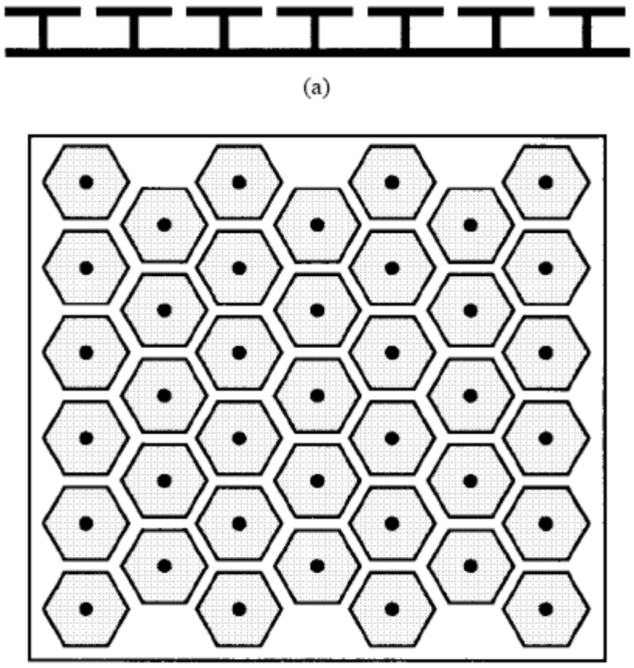


Figure 32.11: A mushroom structure operates like an LC tank circuit. At the right resonant frequency, the surface resembles an open-circuit surface where no current can flow. Hence, tangential magnetic field is zero resembling perfect magnetic conductor (courtesy of Sievenpiper [192]).

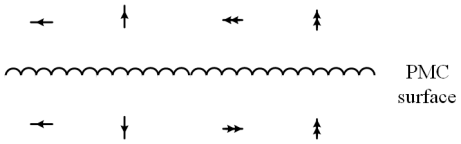


Figure 32.12: Reflection rules for electric and magnetic dipoles over a PMC surface. Magnetic dipoles are denoted by double arrows.

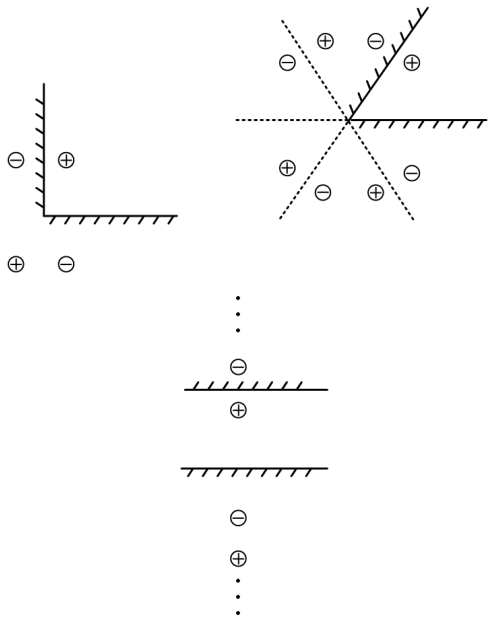


Figure 32.13: Image theory for multiple images [31].

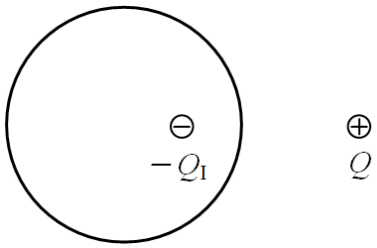


Figure 32.14: Image theory for a point charge near a cylinder or a sphere can be found in closed form [31].

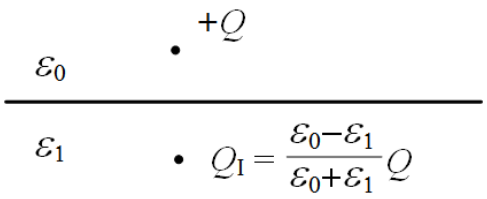


Figure 32.15: A static charge over a dielectric interface can be found in closed form using Fourier transform technique. The solution is beyond the scope of this course.

Lecture 33

High Frequency Solutions, Gaussian Beams

When the frequency is very high, the wavelength of electromagnetic wave becomes very short. In this limit, the 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 equations from those of electricity and magnetism. Thus 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. 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 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 27.2.1), 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. This will simplify the solutions in many instances. For instance, ray tracing can be used to track how these rays can propagate, bounce, or "richohet" in a complex environment. In fact, it is now done in a movie industry to give "realism" to simulate the nuances of how light ray will bounce around in a room, and reflecting off objects.

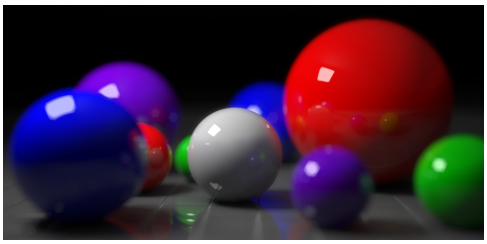


Figure 33.1: Ray-tracing technique can be used in the movie industry to produce realism in synthetic images (courtesy of Wikipedia).

33.1 Tangent Plane Approximations

We have learnt that reflection and transmission of waves at a flat surface can be solved in closed form. The important point here is the physics of phase matching. Due to phase matching, we have the law of reflection, transmission and Snell's law [58].[1]

When a surface is not flat anymore, there is no closed form solution. But when a surface is curved, an approximate solution can be found. This is obtained by using a local tangent-plane approximation when the radius of curvature is much larger than the wavelength. Hence, this 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 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 [32, 194]. In GO, light waves are replaced by light rays. As mentioned before, a light ray is a part of a spherical wave where locally, it 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 lens or ray optics from which lens technology is derived (see Figure 33.3). It is also the basis for ray tracing for high-frequency solutions [195, 196].[2]

Many real world problems do not have closed-form solutions, and have to be treated with approximate methods. In addition to geometrical 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 [197–201].

[1] This law is also known in the Islamic world in 984 [193].

[2] 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.

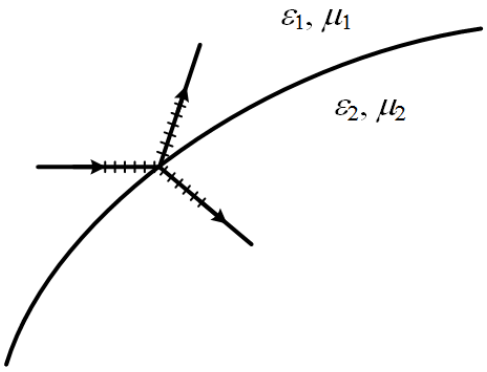


Figure 33.2: In the tangent plane approximation, the surface where reflection and refraction occur is assumed to be locally flat. Thus, phase-matching is approximately satisfied, and hence, the law of reflection, transmittion, and Snell's law are satisfied locally.

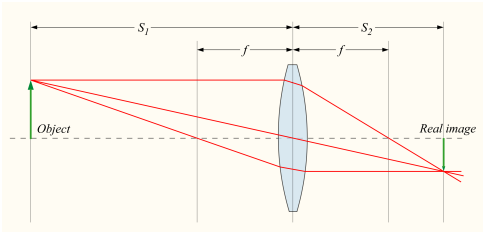


Figure 33.3: Tangent plane approximations can also be made at dielectric interfaces so that Fresnel reflection and transmission coeffients can be used to ascertain the interaction of light rays with a lens. Also, one can use ray tracing to understand the physics of an optical lens (courtesy of Wikipedia).

33.2 Fermat's Principle

Fermat's principle (1600s) [58,202] says that a light ray follows the path that takes the shortest time between two points.[3] 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 stated that a plane wave follows the path that has a minimal phase shift. This principle can be used to derive law of reflection, transmission, and refraction for light rays. It can be used as the guiding principle for ray tracing as well.

[3]This eventually give rise to the principle of least action.

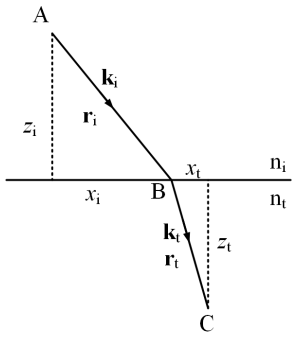


Figure 33.4: In Fermat's principle, a light ray, when propagating from point A to point C, takes the path of least delay in time in the time domain, and delay in phase in the frequency domain.

Given two points A and C 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[4]

$$P = \mathbf{k}_i \cdot \mathbf{r}_i + \mathbf{k}_t \cdot \mathbf{r}_t \tag{33.2.1}$$

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. 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 of the path length or phase shift. Specializing (33.2.1) to a 2D picture, then the phase shift as a function of x_i is stationary. In this 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,[5]

$$\frac{\partial P}{\partial x_i} = 0 = k_{ix} - k_{tx} \tag{33.2.2}$$

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 have been shown earlier. This law was also known in the Islamic world to Ibn Sahl in 984 [193].

[4]In this course, for wavenumber, we use k and β interchangeably, where k is prevalent in optics and β used in microwaves.

[5]One can show that as the separations between A, B, and C are large, and 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 to x_i, one would also get the same answer.

33.2.1 Generalized Snell's Law

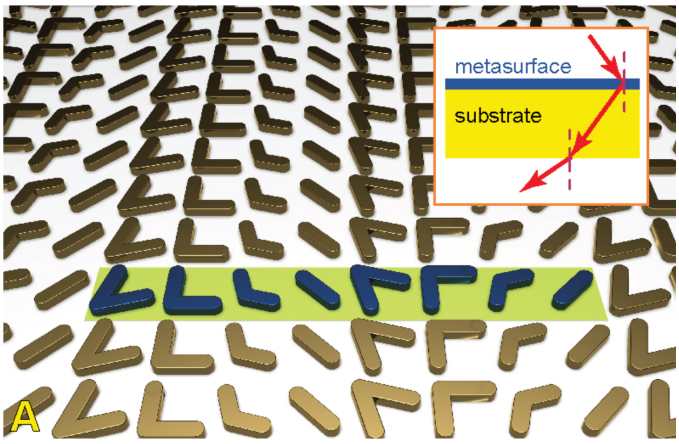


Figure 33.5: A phase screen which is position dependent can be made using nano-fabrication and design with commercial software for solving Maxwell's equations. In such a case, one can derive a generalized Snell's law to describe the diffraction of a wave by such a surface (courtesy of Capasso's group [203]).

Metasurfaces are prevalent these days due to our ability for 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 due to advent in nano-fabrication technology and commercial software for numerical simulation.

To model this phase screen, we can 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 [203] 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. [6]

[6] Such research is being pursued by V. Shalaev and A. Boltasseva's group at Purdue U [204].

33.3 Gaussian Beam

We have seen previously that in a source-free medium

$$\nabla^2 \mathbf{A} + \omega^2 \mu\varepsilon \mathbf{A} = 0 \tag{33.3.1}$$

$$\nabla^2 \Phi + \omega^2 \mu\varepsilon \Phi = 0 \tag{33.3.2}$$

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[7]

$$A_x(\mathbf{r}) = \Psi(\mathbf{r}) e^{-j\beta z} \tag{33.3.5}$$

where $\Psi(\mathbf{r}) = \Psi(x, y, z)$ is a 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, 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 containing β^2 on the right-hand side of (33.3.6) cancels with the term coming from $\omega^2\mu\varepsilon\mathbf{A}$ of (33.3.1). 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,

$$\left|2j\beta\frac{\partial}{\partial z}\Psi\right| \gg \left|\frac{\partial^2}{\partial z^2}\Psi\right| \tag{33.3.8}$$

[7]Also, the wave becomes a transverse wave in the far field, and keeping the transverse component suffices.

where we have assumed that Ψ is a slowly varying function of z such that $\beta\Psi \gg \partial/\partial z\Psi$. 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.[8] 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

A closed form solution to the paraxial wave equation can be obtained by a simple trick.[9] 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 as $\mathbf{r}'$ 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) yield

$$A_x(\mathbf{r}) \approx \frac{e^{-j\beta(z+jb)}}{4\pi(z+jb)} e^{-j\beta\frac{x^2+y^2}{2(z+jb)}} \approx e^{-j\beta z}\Psi(\mathbf{r}) \tag{33.3.13}$$

By comparing the above with (33.3.5), we can identify

$$\Psi(x,y,z) \cong A_0 \frac{jb}{z+jb} e^{-j\beta\frac{x^2+y^2}{2(z+jb)}} \tag{33.3.14}$$

[8] The paraxial wave equation, the diffusion equation and the Schrodinger equation are all classified as parabolic equations in mathematical parlance [35, 49, 205, 206].

[9] Introduced by Georges A. Deschamps of UIUC [207].

where A_0 is used to absorbed the constant to simplify the expression. 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, viz.,

$$\frac{jb}{z+jb} = \frac{1}{\sqrt{1+z^2/b^2}} e^{j\tan^{-1}(\frac{z}{b})} \tag{33.3.16}$$

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 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, and

$$w^2 = \frac{2b}{\beta}\left(1+\frac{z^2}{b^2}\right), \qquad R = \frac{z^2+b^2}{z}, \qquad \psi = \tan^{-1}\left(\frac{z}{b}\right) \tag{33.3.19}$$

For a fixed z, the parameters w, R, and ψ are constants. It is seen that the beam is Gaussian tapered in the x and y directions. Hence, w is the beam waist which varies with z, and it is smallest when $z = 0$, or $w = w_0 = \sqrt{\frac{2b}{\beta}}$.

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, $x^2+y^2 \ll z^2$ 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.

The phase ψ defined in (33.3.19) changes linearly with z for small z, and saturates to a constant for large z. Then the phase of the entire wave is due to the $\exp(-j\beta z)$ in (33.3.13). A cross section of the electric field due to a Gaussian beam is shown in Figure 33.6.

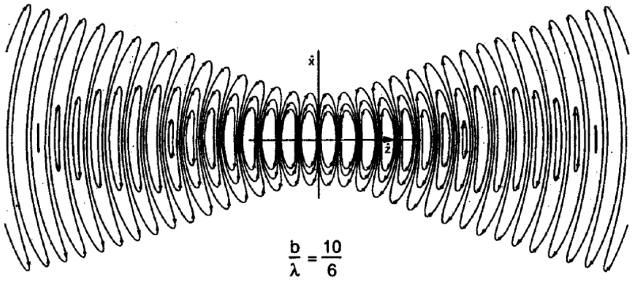


Figure 33.6: Electric field of a Gaussian beam in the x-z plane frozen in time. The wave moves to the right as time increases; here, $b/\lambda = 10/6$ (courtesy of Haus, Electromagnetic Noise and Quantum Optical Measurements [82]). The narrowest beam waist is given by $w_0/\lambda = \sqrt{b/(\lambda\pi)}$.

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). But it is a partial differential equation which can be solved by the separation of variables just like the Helmholtz wave equation. Therefore, in general, it has solutions of the form[10]

$$\Psi_{nm}(x,y,z) = \left(\frac{2}{\pi n!m!}\right)^{1/2} 2^{-N/2} \left(\frac{1}{w}\right) e^{-(x^2+y^2)/w^2} e^{-j\frac{\beta}{2R}(x^2+y^2)} e^{j(m+n+1)\Psi} \cdot H_n\left(x\sqrt{2}/w\right) H_m\left(y\sqrt{2}/w\right) \quad (33.3.21)$$

where $H_n(\xi)$ is a Hermite polynomial of order n. The solutions can also be expressed in terms of Laguere polynomials , namely,

$$\Psi_{nm}(x,y,z) = \left(\frac{2}{\pi n!m!}\right)^{1/2} \min(n,m)! \frac{1}{w} e^{-j\frac{\beta}{2R}\rho^2} - e^{-\rho^2/w^2} e^{+j(n+m+1)\Psi} e^{jl\phi} (-1)^{min(n,m)} \left(\frac{\sqrt{2}\rho}{w}\right) L^{n-m}_{\min(n,m)}\left(\frac{2\rho^2}{w^2}\right) \quad (33.3.22)$$

where $L^k_n(\xi)$ is the associated Laguerre polynomial.

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 due to the complex transverse field distribution of the beams.[11] They

[10] See F. Pampaloni and J. Enderlein [208].

[11] See D.L. Andrew, Structured Light and Its Applications and articles therein [209].

harbor potential for optical communications as well as optical tweezers to manipulate trapped nano-particles. Figure 33.7 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 [209].

Laguerre–Gaussian Beams and Orbital Angular Momentum

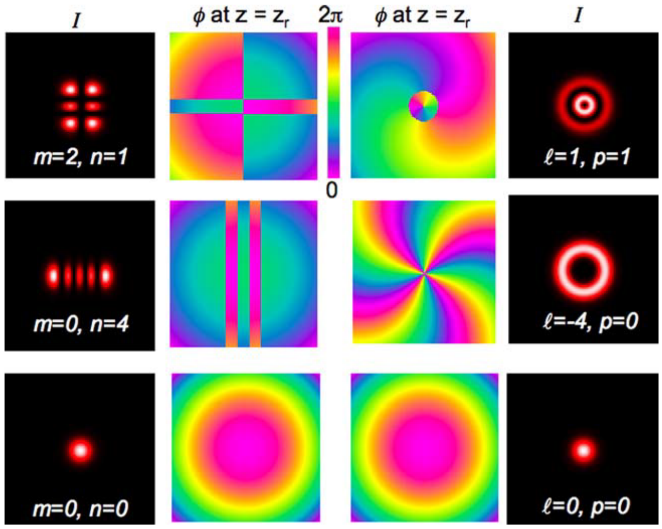


Figure 1.1 Examples of the intensity and phase structures of Hermite–Gaussian modes (*left*) an Laguerre–Gaussian modes (*right*), plotted at a distance from the beam waist equal to the Rayleig range. See color insert.

Figure 33.7: 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 [209]).

Lecture 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! 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. In the optical regime, the scattered light allows us to see objects, as well as admire all hues and colors that are observed of objects. In microwave, the scatterers cause the loss of energy carried by a plane wave. A proper understanding of scattering theory allows us to understand many physical phenomena around us. We will begin by studying Rayleigh scattering, which is scattering by small objects compared to wavelength. With Rayleigh scattering of a simple sphere, we can understand 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 wavelength of light. The size of water molecule is about 0.25 nm, while the wavelength of blue light is about 500 nm. Then a simple solution can be found by using quasi-static analysis in the vicinity of the small particle. This simple scattering solution can be used to explain a number of physical phenomena in nature (see Figure 34.1). For instance, why the sky is blue, the sunset so magnificently beautiful, how birds and insects can navigate themselves without the help of a compass. By the same token, it can also be used to explain why 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 in ancient times.

Figure 34.1: The magnificent beauty of nature can be partly explained by Rayleigh scattering [210, 211].

When a ray of light impinges on an object, we model the incident light as a plane electromagnetic wave (see Figure 34.2). 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 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. In other words, the particle feels an quasi-electrostatic field in the incident field. The incident field polarizes the particle, making it radiate 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).

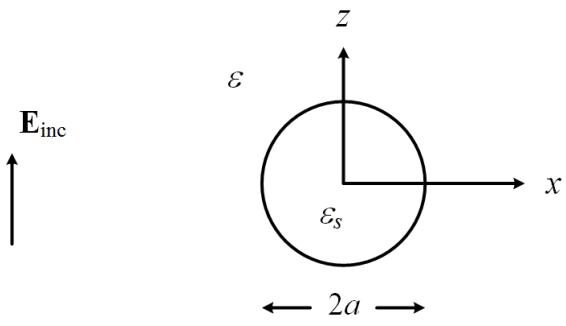


Figure 34.2: Geometry for studying the Rayleigh scattering problem.

34.1.1 Scattering by a Small Spherical Particle

The incident field polarizes the particle making it look like an 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 the solutions together is known as asymptotic matching.)

A Hertzian dipole can be approximated by a small current source so that

$$\mathbf{J}(\mathbf{r}) = \hat{z} Il \delta(\mathbf{r}) \tag{34.1.1}$$

Without loss of generality, we have assumed the Hertzian dipole to be at the origin. In the above, we can 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 $p = ql$. The vector potential $\mathbf{A}$ due to a Hertzian dipole, after substituting (34.1.1), is

$$\begin{aligned} \mathbf{A}(\mathbf{r}) &= \frac{\mu}{4\pi} \iiint_V d\mathbf{r}' \frac{\mathbf{J}(\mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|} e^{-j\beta|\mathbf{r}-\mathbf{r}'|} \\ &= \hat{z} \frac{\mu Il}{4\pi r} e^{-j\beta r} = \hat{z} \frac{j\omega\mu ql}{4\pi r} e^{-j\beta r} \end{aligned} \tag{34.1.3}$$

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.

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 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 need to 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. 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 (23.2.23)) 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 can use a quasi-static approximation about the potential.[1] 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 dipole.

The 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 incident 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

$$\Phi_{sca} = E_s\frac{a^3}{r^2}\cos\theta \tag{34.1.9}$$

which is the potential due to a static dipole. The electrostatic boundary value problem (BVP) has been previously solved and[2]

$$E_s = \frac{\varepsilon_s - \varepsilon}{\varepsilon_s + 2\varepsilon}E_i \tag{34.1.10}$$

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 that

$$p = ql = 4\pi\varepsilon\frac{\varepsilon_s - \varepsilon}{\varepsilon_s + 2\varepsilon}a^3E_i = \alpha E_i \tag{34.1.12}$$

where α is the polarizability of the small particle.

[1] This is the same as ignoring retardation effect.

[2] It was one of the homework problems. See also Section 8.3.6.

Far Field

Now that we have learnt that a small particle is polarized by the incident field, which can be treated as a constant **E** field in its vicinity. In other words, the incident field induces a small dipole moment on the small particle. 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 is quite different from its near field (see Section 25.2). In the far field of the Hertzian dipole, we can start with

$$\mathbf{E} = -j\omega\mathbf{A} - \nabla\Phi = -j\omega\mathbf{A} - \frac{1}{j\omega\mu\varepsilon}\nabla\nabla\cdot\mathbf{A} \tag{34.1.13}$$

In the above, Lorenz gauge (see (23.2.23)) has been used to relate Φ to the vector potential **A**. But when we are in the far field, **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 26.1.2.

34.1.2 Scattering Cross Section

From (34.1.3), and making use of (34.1.2), we see that $A_\phi = 0$ while

$$A_\theta = -\frac{j\omega\mu ql}{4\pi r}e^{-j\beta r}\sin\theta \tag{34.1.15}$$

Consequently, using (34.1.12) for ql, we have in the far field that[3]

$$E_\theta \cong -j\omega A_\theta = -\frac{\omega^2\mu ql}{4\pi r}e^{-j\beta r}\sin\theta = -\omega^2\mu\varepsilon\left(\frac{\varepsilon_s - \varepsilon}{\varepsilon_s + 2\varepsilon}\right)\frac{a^3}{r}E_i e^{-j\beta r}\sin\theta \tag{34.1.16}$$

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 $\langle\mathbf{S}\rangle = 1/2\Re e\{\mathbf{E}\times\mathbf{H}^*\}$. 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

$$P_s = \frac{1}{2}\int_0^\pi r^2\sin\theta d\theta\int_0^{2\pi} d\phi E_\theta H_\phi^* = \frac{1}{2\eta}\int_0^\pi r^2\sin\theta d\theta\int_0^{2\pi} d\phi|E_\theta|^2 \tag{34.1.18}$$

$$= \frac{1}{2\eta}\beta^4\left|\frac{\varepsilon_s - \varepsilon}{\varepsilon_s + 2\varepsilon}\right|^2\frac{a^6}{r^2}|E_i|^2 r^2\left(\int_0^\pi \sin^3\theta d\theta\right)2\pi \tag{34.1.19}$$

[3]The ω^2 dependence of the following function implies that the radiated electric field in the far zone is proportional to the acceleration of the charges on the dipole.

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

$$P_s = \frac{4\pi}{3\eta}\left|\frac{\varepsilon_s - \varepsilon}{\varepsilon_s + 2\varepsilon_s}\right|^2 \beta^4 a^6 |E_i|^2 \tag{34.1.21}$$

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 (7.1.9) to arrive at a Gauss' law for lossy dispersive media, viz., $\nabla \cdot \underset{\sim}{\varepsilon}\mathbf{E} = 0$ 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

$$\Sigma_s = \frac{P_s}{\langle S_{\text{inc}}\rangle} = \frac{8\pi a^2}{3}\left|\frac{\varepsilon_s - \varepsilon}{\varepsilon_s + 2\varepsilon}\right|^2 (\beta a)^4 \tag{34.1.22}$$

where we have used the local plane-wave approximation that

$$\langle S_{\text{inc}}\rangle = \frac{1}{2\eta}|E_i|^2 \tag{34.1.23}$$

The above also implies that

$$P_s = \langle S_{\text{inc}}\rangle \cdot \Sigma_s$$

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.

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).

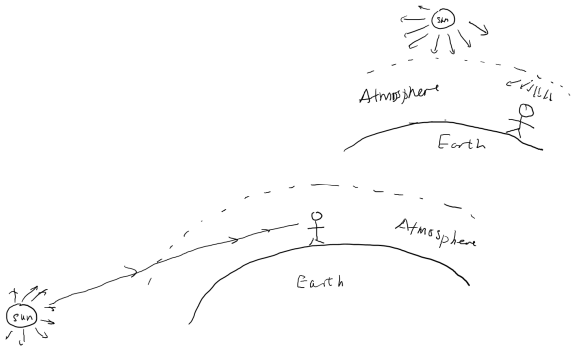


Figure 34.3: During the day time, when we look at the sky, we mainly see scattered sunlight. Since high-frequency light is scattered more, the sky appears blue. At sunset, the sunlight has to go through a thicker atmosphere. Thus, the blue light is scattered away, leaving the red light that reaches the eyes. Hence, the sunset appears red.

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 can be very small giving rise to strongly scattered light (see Section 8.3.6).

Furthermore, since the far field scattered power density of this particle is

$$\langle S \rangle = \frac{1}{2\eta} E_\theta H_\phi^* \sim \sin^2 \theta \tag{34.1.24}$$

the scattering 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.[4] Such a "sunstone" for direction finding is shown in Figure 34.4.

[4]You can go through a Gedanken experiment to convince yourself of such.

Figure 34.4: A sunstone can indicate the polarization of the scattered light. From that, one can deduce where the sun is located (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 PEC sphere.[5] 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)[6] of a PEC sphere as illustrated in Figure 34.5.

[5]Note that there is no PEC at optics. A metal behaves more like a plasma medium at optical frequencies.
[6]Scattering cross section in microwave range is called an RCS due to its prevalent use in radar technology.

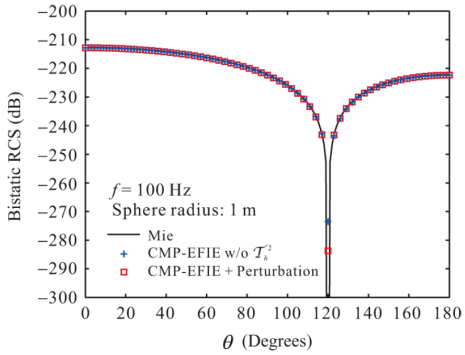


Figure 34.5: RCS (radar scattering cross section) of a small PEC scatterer (courtesy of Sheng et al. [212]).

34.2 Mie Scattering

When the size of the dipole 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 [213,214]. 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.

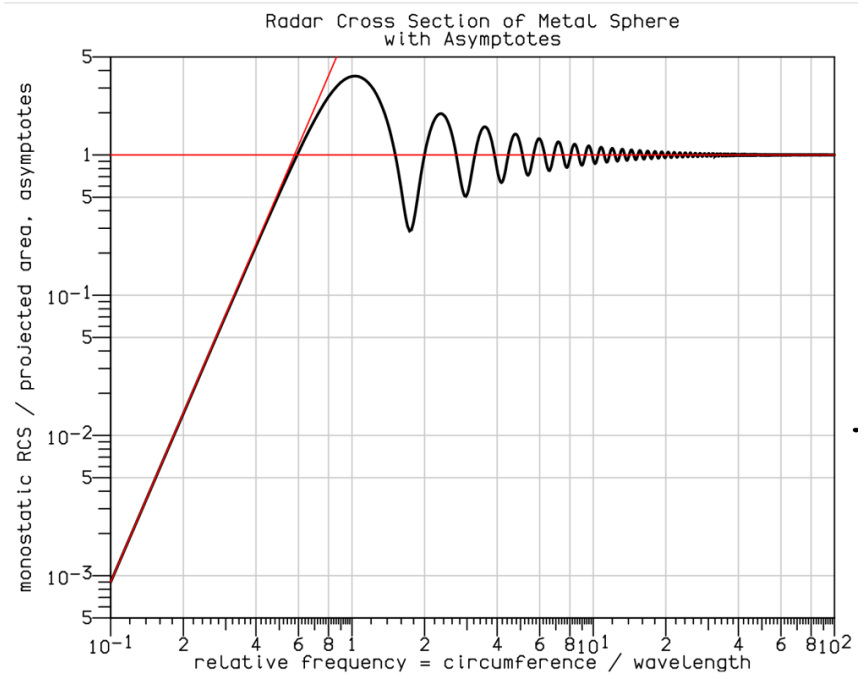


Figure 34.6: Radar cross section (RCS) calculated using Mie scattering theory [214].

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 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 the proof of this is given in J.D. Jackson's book [48].

The true physical reason for this is power orthogonality. Two plane waves cannot interact or exchange power with each other unless they share the same $\mathbf{k}$ or $\boldsymbol{\beta}$ vector, where $\boldsymbol{\beta}$ is both the plane wave direction of the incident wave as well as the forward scattered wave. This is similar to power orthogonality in a waveguide, and it happens for orthogonal modes in waveguides [83, 178].

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 save 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.

In a nutshell, the scattering theorem is intuitively obvious for high-frequency scattering. The amazing part about this theorem is that it is true for all frequencies.

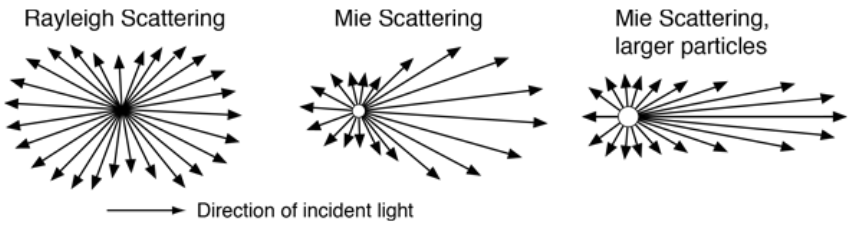


Figure 34.7: A particle scatters increasingly more in the forward direction as the frequency increases (Courtesy of hyperphysics.phy-astr.gsu.edu).

34.2.2 Mie Scattering by Spherical Harmonic Expansions

As mentioned before, as the wavelength becomes shorter, we need to solve the boundary value problem in its full glory without making any approximations. 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 is beyond the scope of this course.[7] The separation of variables in spherical coordinates is not the only useful for Mie scattering, it is also useful for analyzing spherical cavity. So we will present the precursor knowledge so that you can read further into Mie scattering theory if you need to in the future.

34.2.3 Separation of Variables in Spherical Coordinates[8]

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. Hence, the Helmholtz wave equation becomes[9]

$$\left(\frac{1}{r^2}\frac{\partial}{\partial r}r^2\frac{\partial}{\partial r}+\frac{1}{r^2\sin\theta}\frac{\partial}{\partial\theta}\sin\theta\frac{\partial}{\partial\theta}+\frac{1}{r^2\sin^2\theta}\frac{\partial^2}{\partial\phi^2}+\beta^2\right)\Psi(\mathbf{r})=0 \tag{34.2.1}$$

Noting the $\partial^2/\partial\phi^2$ derivative, by using separation of variables technique, we assume $\Psi(\mathbf{r})$ to be of the form

$$\Psi(\mathbf{r})=F(r,\theta)e^{jm\phi} \tag{34.2.2}$$

[7] But it is treated in J.A. Kong's book [32] and Chapter 3 of W.C. Chew, Waves and Fields in Inhomogeneous Media [35] and many other textbooks [48, 65, 181].

[8] May be skipped on first reading.

[9] By quirk of mathematics, it turns out that the first term on the right-hand side below can be simplified by observing that $\frac{1}{r^2}\frac{\partial}{\partial r}r^2=\frac{1}{r}\frac{\partial}{\partial r}r$.

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

$$\left(\frac{1}{r^2}\frac{\partial}{\partial r}r^2\frac{\partial}{\partial r} + \frac{1}{r^2\sin\theta}\frac{\partial}{\partial\theta}\sin\theta\frac{\partial}{\partial\theta} - \frac{m^2}{r^2\sin^2\theta} + \beta^2\right)F(r,\theta) = 0 \tag{34.2.3}$$

Again, by using the separation of variables, and letting further that

$$F(r,\theta) = b_n(\beta r)P_n^m(\cos\theta) \tag{34.2.4}$$

where we require that

$$\left\{\frac{1}{\sin\theta}\frac{d}{d\theta}\sin\theta\frac{d}{d\theta} + \left[n(n+1) - \frac{m^2}{\sin^2\theta}\right]\right\}P_n^m(\cos\theta) = 0 \tag{34.2.5}$$

when $P_n^m(\cos\theta)$ is the associate Legendre polynomial. 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 separation constant.

Consequently, $b_n(kr)$ 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. They are related to the cylindrical functions via [35, 49] [10]

$$b_n(\beta r) = \sqrt{\frac{\pi}{2\beta r}}B_{n+\frac{1}{2}}(\beta r) \tag{34.2.7}$$

It is customary to define the spherical harmonic as [48, 110]

$$Y_{nm}(\theta,\phi) = \sqrt{\frac{2n+1}{4\pi}\frac{(n-m)!}{(n+m)!}}P_n^m(\cos\theta)e^{jm\phi} \tag{34.2.8}$$

The above is 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[11] like Fourier series, so that

$$\sum_{n=0}^{\infty}\sum_{m=-n}^{n}Y_{nm}^*(\theta',\phi')Y_{nm}(\theta,\phi) = \delta(\phi-\phi')\delta(\cos\theta-\cos\theta') \tag{34.2.11}$$

[10] 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.

[11] 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.

Lecture 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 reflections off a flat surface or a planarly layered medium can be found easily. When we have a source like a point source, it generates a spherical wave. We do not know how to reflect exactly a spherical wave off a planar interface. But by expanding a spherical wave in terms of sum of plane waves and evanescennt 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. 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 integrals to elucidate their physics. From this, we can see ray theory emerging from the complicated mathematics. It reminds me 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. If you can assimilate them, it will be an exhilarating experience.

35.1 Spectral Representations of Sources

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 waves can be expanded in terms of 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 plane waves summation. This topic is found in many textbooks [1, 32, 35, 95, 96, 181, 205, 215].

35.1.1 A Point Source

There are a number of ways to derive the plane wave expansion of a point source. We will illustrate one of the ways. The 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 with 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 then be solved in the spherical coordinates, yielding the solution given in the previous lecture, namely, Green's function with the source point at the origin, or[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 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 (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

$$\nabla^2 = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2} = -k_x^2 - k_y^2 - k_z^2$$

Then, together with the Fourier representation of the delta function, which is[4]

$$\delta(x)\,\delta(y)\,\delta(z) = \frac{1}{(2\pi)^3}\iiint_{-\infty}^{\infty} dk_x dk_y dk_z\, e^{ik_x x + ik_y y + ik_z z} \tag{35.1.4}$$

we convert (35.1.1) into

$$\iiint_{-\infty}^{\infty} dk_x dk_y dk_z\, [k_0^2 - k_x^2 - k_y^2 - k_z^2]\tilde{\phi}(k_x,k_y,k_z) e^{ik_x x + ik_y y + ik_z z} \tag{35.1.5}$$

$$= -\iiint_{-\infty}^{\infty} dk_x dk_y dk_z\, e^{ik_x x + ik_y y + ik_z z}. \tag{35.1.6}$$

[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 [110]).

[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.

Since the above is equal for all x, y, and z, we can Fourier inverse transform the above to get

$$\tilde{\phi}(k_x, k_y, k_z) = \frac{-1}{k_0^2 - k_x^2 - k_y^2 - k_z^2}. \tag{35.1.7}$$

Consequently, we have

$$\phi(x, y, z) = \frac{-1}{(2\pi)^3} \iiint_{-\infty}^{\infty} d\mathbf{k} \, \frac{e^{ik_x x + ik_y y + ik_z z}}{k_0^2 - k_x^2 - k_y^2 - k_z^2}. \tag{35.1.8}$$

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 $k_x^2 + k_y^2 + k_z^2 \neq k_0^2$.

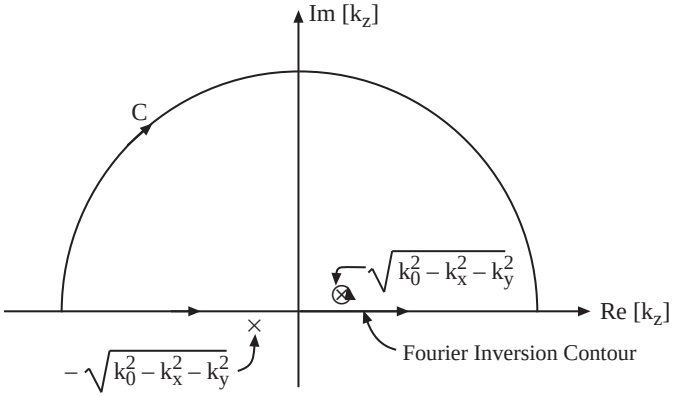


Figure 35.1: The integration along the real axis is equal to the integration along C plus the residue of the pole at $(k_0^2 - k_x^2 - k_y^2)^{1/2}$, by invoking Jordan's lemma.

Weyl Identity

To make the plane waves in (35.1.8) into physical plane waves, we have to massage it into a different form. We rearrage the integrals in (35.1.8) so that the dk_z integral is performed first. In other words,

$$\phi(\mathbf{r}) = \frac{1}{(2\pi)^3} \iint_{-\infty}^{\infty} dk_x dk_y e^{ik_x x + ik_y y} \int_{-\infty}^{\infty} dk_z \frac{e^{ik_z z}}{k_z^2 - (k_0^2 - k_x^2 - k_y^2)} \tag{35.1.9}$$

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

[5]In (35.1.8), 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 [216].

(35.1.8) undefined. However, if a small loss is assumed in k_0 such that $k_0 = k_0' + ik_0''$, then the poles are 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 [52]: The manipulation that leads to (35.1.8) is not strictly correct. Second, the introduction of a small loss also guarantees the radiation condition and the uniqueness of the solution to (35.1.1), and therefore, the equality of (35.1.2) and (35.1.8) [35].

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 [89], the integration for k_z over the contour C as shown in Figure 35.1 vanishes. Then, by Cauchy's theorem [89], the integration over the Fourier inversion contour on the real axis is the same as integrating over 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, we have

$$\phi(x,y,z) = \frac{i}{2(2\pi)^2} \iint_{-\infty}^{\infty} dk_x dk_y \, \frac{e^{ik_x x + ik_y y + ik_z' z}}{k_z'}, \quad z > 0, \tag{35.1.10}$$

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, one can deform the contour to pick up the pole contribution. Hence, 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). As such, the result for all z can be written as

$$\phi(x,y,z) = \frac{i}{2(2\pi)^2} \iint_{-\infty}^{\infty} dk_x dk_y \, \frac{e^{ik_x x + ik_y y + ik_z' |z|}}{k_z'}, \quad \text{all } z. \tag{35.1.11}$$

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 (35.1.11), yielding the identity

$$\frac{e^{ik_0 r}}{r} = \frac{i}{2\pi} \iint_{-\infty}^{\infty} dk_x dk_y \, \frac{e^{ik_x x + ik_y y + ik_z |z|}}{k_z}, \tag{35.1.12}$$

where $k_x^2 + k_y^2 + k_z^2 = k_0^2$, or $k_z = (k_0^2 - k_x^2 - k_y^2)^{1/2}$. The above is known as the ***Weyl identity*** (Weyl 1919). 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 (35.1.12) could be interpreted as an integral summation of plane waves propagating in all directions, including evanescent waves. It is the plane-wave expansion (including evanescent wave) of a spherical wave.

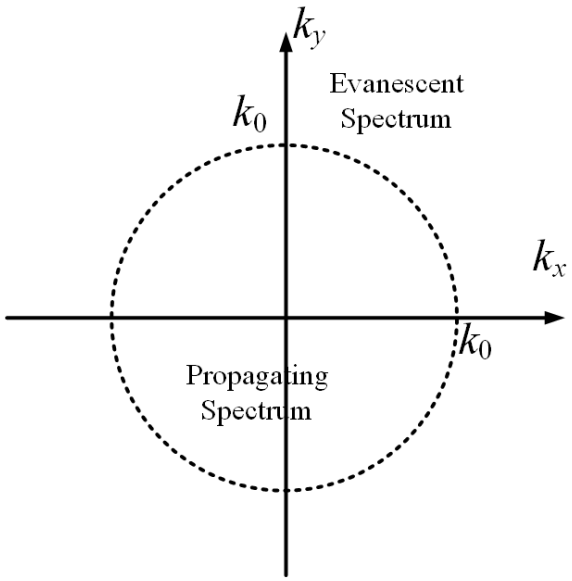


Figure 35.2: The wave is propagating for $\mathbf{k}_\rho$ vectors inside the disk, while the wave is evanescent for $\mathbf{k}_\rho$ outside the disk.

One can also interpret the above as a 2D surface integral in the Fourier space over the k_x and k_y plane or variables. When $k_x^2 + k_y^2 < k_0^2$, or the spatial spectrum 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. The high spectral components, which are related to the evanescent waves, are important for reconstructing the singularity of the Green's function.

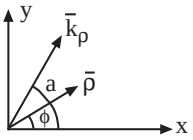


Figure 35.3: The $\mathbf{k}_\rho$ and the $\boldsymbol{\rho}$ vector on the xy plane.

Sommerfeld Identity

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 integral. First, in (35.1.12), 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), and $dk_x dk_y = k_\rho dk_\rho\, d\alpha$. Then, $k_x x + k_y y = \mathbf{k}_\rho \cdot \boldsymbol{\rho} =$

$k_\rho \cos(\alpha - \phi)$, and with the appropriate change of variables, we have

$$\frac{e^{ik_0 r}}{r} = \frac{i}{2\pi} \int_0^\infty k_\rho dk_\rho \int_0^{2\pi} d\alpha \frac{e^{ik_\rho \rho \cos(\alpha-\phi)+ik_z|z|}}{k_z}, \tag{35.1.13}$$

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, $k_\rho^2 = k_x^2 + k_y^2$. Then, using the integral identity for Bessel functions given by[6]

$$J_0(k_\rho \rho) = \frac{1}{2\pi} \int_0^{2\pi} d\alpha \, e^{ik_\rho \rho \cos(\alpha-\phi)}, \tag{35.1.14}$$

(35.1.13) becomes

$$\frac{e^{ik_0 r}}{r} = i \int_0^\infty dk_\rho \, \frac{k_\rho}{k_z} J_0(k_\rho \rho) e^{ik_z|z|}. \tag{35.1.15}$$

The above is also known as the ***Sommerfeld identity*** (Sommerfeld 1909 [109]; [205][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 evanescent in the $\pm z$ direction when $k_\rho > k_0$.

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 $H_0^{(1)}(e^{i\pi} x) = -H_0^{(2)}(x)$, a variation of the above identity can be derived as

$$\frac{e^{ik_0 r}}{r} = \frac{i}{2} \int_{-\infty}^\infty dk_\rho \, \frac{k_\rho}{k_z} H_0^{(1)}(k_\rho \rho) e^{ik_z|z|}. \tag{35.1.16}$$

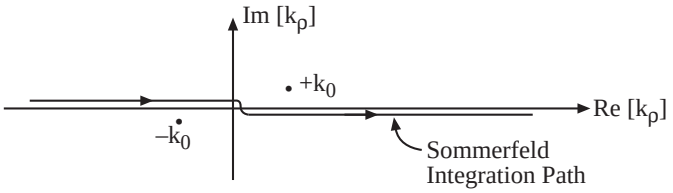


Figure 35.4: Sommerfeld integration path.

Since $H_0^{(1)}(x)$ has a logarithmic branch-point singularity at $x = 0$,[7] and $k_z = (k_0^2 - k_\rho^2)^{1/2}$ has algebraic branch-point singularities at $k_\rho = \pm k_0$, the integral in Equation (35.1.16) is

[6] See Chew [35], or Whitaker and Watson(1927) [217].

[7] $H_0^{(1)}(x) \sim \frac{2i}{\pi} \ln(x)$, see Chew [35][p. 14], or Abromawitz or Stegun [117].

undefined unless we stipulate also the path of integration. Hence, 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 bronach 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 14.1 as well as through a layered medium in Section 16.1. 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 $H_z = 0$, $E_z \neq 0$.[8] 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 its field in terms of plane waves of TE-type and TM-type. Then, the nonzero component of E_z characterizes TM waves, while the nonzero component of H_z characterizes TE 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) [109] 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) [218] 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

[8] *Chew, Waves and Fields in Inhomogeneous Media* [35]; *Kong, Electromagnetic Wave Theory* [32].

shown to be given by

$$\mathbf{H}(\mathbf{r}) = \nabla \times \hat{\alpha} I\ell \, \frac{e^{ikr}}{4\pi r}. \tag{35.2.2}$$

With the above fields, their TM and TE components can be extracted easily in anticipation of their plane wave expansions for propagation through layered media.

(a) Vertical Electric Dipole (VED)

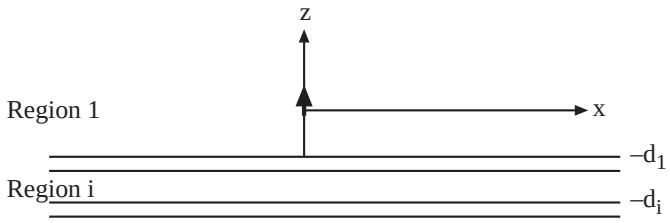


Figure 35.5: A vertical electric dipole over a layered medium.

A vertical electric dipole shown in Figure 35.5 has $\hat{\alpha} = \hat{z}$; hence, in anticipation of their plane wave expansions, the TM 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 field.

Next, using the Sommerfeld identity (35.1.16) in the above, and after exchanging the order of integration and differentiation, we have[9]

$$E_z = \frac{-I\ell}{4\pi\omega\epsilon} \int_0^\infty dk_\rho \, \frac{k_\rho^3}{k_z} J_0(k_\rho \rho) e^{ik_z|z|}, \; |z| \neq 0 \tag{35.2.5}$$

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. Since cylindrical waves actually are linear superpositions of plane waves, because we can work backward from (35.1.16) to (35.1.12) to see this. As such, the integrand in

[9]By using (35.1.16) in (35.2.3), the $\partial^2/\partial z^2$ operating on $e^{ik_z|z|}$ produces a Dirac delta function singularity. Detail discussion on this can be found in the chapter on dyadic Green's function in *Chew, Waves and Fields in Inhomogeneous Media* [35].

(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.[10]

Consequently, for a VED on top of a stratified medium as shown, the downgoing plane wave from the point source will be reflected like TM waves with the generalized reflection coefficient $\tilde{R}_{12}^{TM}$. Hence, 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

$$\epsilon_i E_{iz} = \frac{-I\ell}{4\pi\omega} \int_0^\infty dk_\rho \frac{k_\rho^3}{k_{1z}} J_0(k_\rho\rho) A_i \left[e^{-ik_{iz}z} + \tilde{R}_{i,i+1}^{TM} e^{ik_{iz}z+2ik_{iz}d_i}\right]. \tag{35.2.7}$$

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.[11] Therefore, $\tilde{R}_{i,i+1}^{TM}$ and A_i could be obtained using the methods discussed in *Chew, Waves and Fields in Inhomogeneous Media* [110].

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)

The HED is more complicated. Unlike the VED that excites only the TM waves, an HED will excite both TE and TM 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 and the TE components, in anticipation of their plane wave expansions, as

$$E_z = \frac{iI\ell}{4\pi\omega\epsilon} \frac{\partial^2}{\partial z \partial x} \frac{e^{ikr}}{r}, \tag{35.2.8}$$

$$H_z = -\frac{I\ell}{4\pi} \frac{\partial}{\partial y} \frac{e^{ikr}}{r}. \tag{35.2.9}$$

[10] One can perform a sanity check on the odd and even symmetry of the fields' z-component by sketching the fields of a static horizontal electric dipole.

[11] See *Chew, Waves and Fields in Inhomogeneous Media* [35], p. 46, (2.1.6) and (2.1.7). Or we can gather from (14.1.6) to (14.1.7) 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}.

Then, with the Sommerfeld identity (35.1.16), we can expand the above as

$$E_z = \pm\frac{iI\ell}{4\pi\omega\epsilon}\cos\phi \int_0^\infty dk_\rho\, k_\rho^2 J_1(k_\rho\rho)e^{ik_z|z|} \tag{35.2.10}$$

$$H_z = i\frac{I\ell}{4\pi}\sin\phi \int_0^\infty dk_\rho\, \frac{k_\rho^2}{k_z} J_1(k_\rho\rho)e^{ik_z|z|}. \tag{35.2.11}$$

Now, Equation (35.2.10) represents the wave expansion of the TM field, while (35.2.11) represents the wave expansion of the TE 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 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).

35.3 Stationary Phase Method—Fermat's Principle

Sommerfeld integrals are rather complex, and by themselves, they do not offer much physical insight into the physics of the field. To elucidate the physics, we can apply the stationary phase method to find approximations of these integrals when the frequency is high, or kr is large, or the observation point is many wavelengths away from the source point. It turns out that this method is initmately related to Fermat's principle.

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,

$$\frac{e^{ik_0 r}}{r} = \frac{i}{2\pi}\iint_{-\infty}^{\infty} dk_x dk_y \frac{e^{ik_x x + ik_y y + ik_z|z|}}{k_z} = i\int_0^\infty dk_\rho \frac{k_\rho}{k_z} J_0(k_\rho\rho)e^{ik_z|z|}. \tag{35.3.1}$$

We can just focus our attention on the reflected wave term in (35.2.6) and rewrite it in cartesian coordinates to get

$$\begin{aligned} E_{1z}^R &= \frac{-I\ell}{8\pi^2\omega\epsilon_1} \iint_{-\infty}^{\infty} dk_x dk_y \frac{k_x^2 + k_y^2}{k_{1z}} R_{12}^{TM} e^{ik_x x + ik_y y + ik_{1z}(z+2d_1)} \\ &= \iint_{-\infty}^{\infty} dk_x dk_y \frac{1}{k_{1z}} F(k_x, k_y) e^{ik_x x + ik_y y + ik_{1z}(z+2d_1)} \end{aligned} \tag{35.3.2}$$

where

$$F(k_x, k_y) = \frac{-I\ell}{8\pi^2\omega\epsilon_1}(k_x^2 + k_y^2) R_{12}^{TM}$$

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 $i = 1, 2$. Now the problem reduces to finding the approximation of the following integral:

$$E_{1z}^R = \iint_{-\infty}^{\infty} dk_x dk_y \frac{1}{k_{1z}} F(k_x, k_y) e^{irh(k_x, k_y)} \tag{35.3.3}$$

where

$$rh(k_x, k_y) = r\left(k_x \frac{x}{r} + k_y \frac{y}{r} + k_{1z} \frac{z}{r}\right), \tag{35.3.4}$$

We want to approximate the above integral when kr is large. For simplicity, we have set $d_1 = 0$ to begin with.

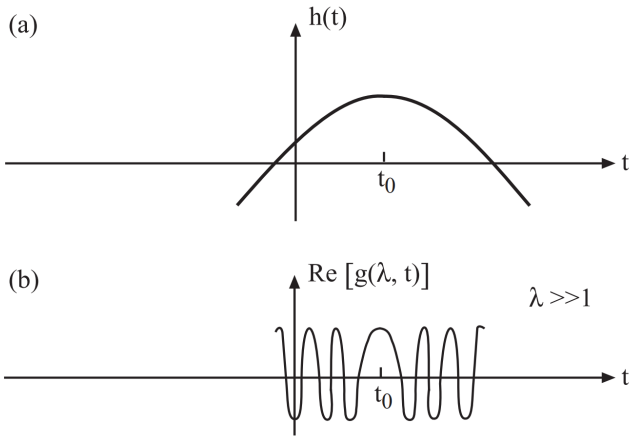


Figure 35.6: In this figure, t can represent k_x or k_y when one of them is varying. Around the stationary phase point, the function $h(t)$ is slowly varying. In this figure, $\lambda = r$, and $g(k_x, k_y, \lambda) = e^{i\lambda h(k_x,k_y)} = e^{irh(k_x,k_y)}$. When $\lambda = r$ is large, the function $g(\lambda, k_x, k_y)$ is rapidly varying with respect to either k_x or k_y. Hence, most of the contributions to the integral comes from around the stationary phase point.

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.[12] 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. Otherwise, the integrand is rapidly varying away from this point, and the integration will destructively cancel with each other, while around the stationary point, they will add constructively.

The stationary point in the k_x and k_y plane is found by setting the derivatives of $h(k_x, k_y)$ with respect to to k_x and k_y to zero. By so doing

$$\frac{\partial h}{\partial k_x} = \frac{x}{r} - \frac{k_x}{k_{1z}}\frac{z}{r} = 0, \qquad \frac{\partial h}{\partial k_y} = \frac{y}{r} - \frac{k_y}{k_{1z}}\frac{z}{r} = 0 \tag{35.3.5}$$

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

$$k_{xs} = k_1\sin\theta\cos\phi, \quad k_{ys} = k_1\sin\theta\sin\phi \tag{35.3.6}$$

with the corresponding $k_{1zs} = k_1\cos\theta$.

[12]The yardstick in wave physics is always wavelength. Large distance is also synonymous to increasing the frequency or reducing the wavelength.

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 equate $F(k_x, k_y)$ to a constant equal to its value at the stationary phase point, and say that

$$E^R_{1z} \simeq F(k_{xs}, k_{ys}) \iint_{-\infty}^{\infty} \frac{1}{k_{1z}} e^{ik_x x + ik_y y + ik_{1z} z} dk_x dk_y = 2\pi F(k_{xs}, k_{ys}) \frac{e^{ik_1 r}}{ir} \tag{35.3.7}$$

In the above, the integral can be performed in closed form using the Weyl identity. The above expression has two important physical interpretations.

(i) Even though a source is emanating plane waves in all directions in accordance to (35.1.12), at the observation point r far away from the 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 is 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.

(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. If we were to make $d_1 \neq 0$ again in the above analysis, the math remains similar except that now, we replace r with $r_I = \sqrt{x^2 + y^2 + (z + 2d_1)^2}$. Due to the reflecting half-space, the source point has an image point as shown in Figure 35.7 This physical picture is shown in the figure where r_I now is the distance of the observation point to the image point. The stationary phase method extract a ray 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 is at the angle at which the incident ray impinges on the half-space.

(iii) At the stationary point, the ray is formed by the **k**-vector where $\mathbf{k} = \hat{x}k_1 \sin\theta\cos\phi + \hat{y}k_1 \sin\theta\sin\phi + \hat{z}k_1 \cos theta$. This ray points in the same direction as the position vector of the observation point $\mathbf{r} = \hat{x}r\sin\theta\cos\phi + \hat{y}r\sin\theta\sin\phi + \hat{z}r\cos theta$. In other words, the **k**-vector and the **r**-vector point in the same direction. This is reminiscent of Fermat principle, because when this happens, the ray propagates with the minimum phase between the source point and the observation point. When $z \to z + 2d_1$, the ray for the image source is altered to that shown in Figure 35.7 where it the ray is minimum phase from the image source to the obervation point. Hence, the stationary phase method is initimately related to Fermat's principle.

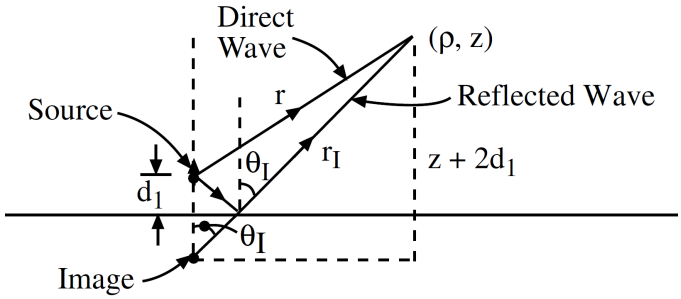


Figure 35.7: At high frequencies, the source point and the observation point are connected by a ray. The ray represents a bundle of plane waves that interfere constructively. This even true for a bundle of plane waves that reflect off an interface. So ray theory or ray optics prevails here, and the ray bounces off the interface according to the reflection coefficient of a plane wave impinging at the interface with θ_I.

35.4 Riemann Sheets and Branch Cuts[13]

The Sommerfeld integrals will have integrands that are multi-value or double value. Proper book keeping is needed so that the evaluation of these integrals can be performed unambiguously. The function $k_z = (k_0^2 - k_\rho^2)^{1/2}$ in (35.1.15) and (35.1.16) are double-value functions because, in taking the square root of a number, two values are possible. In particular, k_z is a double-value function of k_ρ. Consequently, for every point on a complex k_ρ plane in Figure 35.4, there are two possible values of k_z. Therefore, as an example, the integral (35.1.11) is undefined unless we stipulate which of the two values of k_z is adopted in performing the integration.

A multivalue function is denoted on a complex plane with the help of ***Riemann sheets*** [35, 89]. For instance, a double-value function such as k_z is assigned two Riemann sheets to a single complex plane. On one of these Riemann sheets, k_z assumes a value just opposite in sign to the value on the other Riemann sheet. The correct sign for k_z is to pick the square root solution so that $\Im m(k_z) > 0$. This will ensure a decaying wave from the source.

35.5 Some Remarks[14]

Even though we have arrived at the solutions of a point source on top of a layered medium by heuristic arguments of plane waves propagating through layered media, they can also be derived more rigorously. For example, Equation (35.2.6) can be arrived at by matching boundary conditions at every interface. The reason why a more heuristic argument is still valid is due to the completeness of Fourier transforms. It is best explained by putting a source over a half space and a scalar problem.

[13]This may be skipped on first reading.
[14]This may be skipped on first reading.

We can expand the scalar field in the upper region as

$$\Phi_1(x,y,z) = \iint_{-\infty}^{\infty} dk_x dk_y \tilde{\Phi}_1(k_x,k_y,z) e^{ik_x x + ik_y y} \tag{35.5.1}$$

and the scalar field in the lower region as

$$\Phi_2(x,y,z) = \iint_{-\infty}^{\infty} dk_x dk_y \tilde{\Phi}_2(k_x,k_y,z) e^{ik_x x + ik_y y} \tag{35.5.2}$$

If we require that the two fields be equal to each other at $z = 0$, then we have

$$\iint_{-\infty}^{\infty} dk_x dk_y \tilde{\Phi}_1(k_x,k_y,z=0) e^{ik_x x + ik_y y} = \iint_{-\infty}^{\infty} dk_x dk_y \tilde{\Phi}_2(k_x,k_y,z=0) e^{ik_x x + ik_y y} \tag{35.5.3}$$

In order to remove the integral, and replace it with a simple scalar problem, one has to impose the above equation for all x and y. Then by the completeness of Fourier transform implies that[15]

$$\tilde{\Phi}_1(k_x,k_y,z=0) = \tilde{\Phi}_2(k_x,k_y,z=0) \tag{35.5.4}$$

The above equation is much simpler than that in (35.5.3). In other words, due to the completeness of Fourier transform, one can match a boundary condition spectral-component by spectral-component. If the boundary condition is matched for all spectral components, than (35.5.3) is also true.

[15]Or that we can perform a Fourier inversion on the above integrals.

Lecture 36

Computational Electromagnetics, Numerical Methods

Due to the 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 [219], 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 (courtesy of Wikipedia [219]).

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 spurr 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 [220]): The need for a faster computer is obvious. Unfortunately, in the last half century or so, we have been using a large part of the gift of technology to destroy God's greatest gift, life, in warfare!

Figure 36.2: A woman working as a computer shortly after the second world war (courtesy of Wikipedia [220]).

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 is also known as ***computational electromagnetics***. Numerical methods exploit the blinding speed of modern digital computers to perform calculations, and hence to solve large system of equations.

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 derived from Maxwell's equations.

36.2 Examples of Differential Equations

An example of differential equations written in terms of sources are 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, 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. They have 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, and f is the unknown, and g is the driving source. Differential equations, or partial differential equations, as mentioned before, 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).

36.3 Examples of Integral Equations

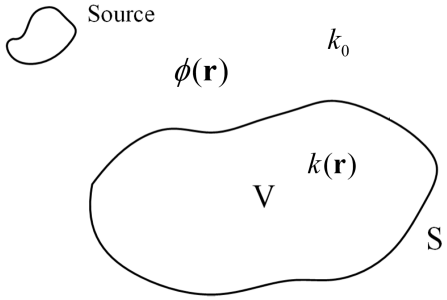


Figure 36.3: Geometry for the derivation of the volume-integral equation for scalar waves. The wavenumber $k(\mathbf{r})$ is assumed to be inhomogeneous and hence, a function of position $\mathbf{r}$ inside the scatterer, but a constant k_0 outside the scatterer.

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. 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

$$[\nabla^2 + k_0^2]g(\mathbf{r}, \mathbf{r}') = -\delta(\mathbf{r} - \mathbf{r}'). \tag{36.3.2}$$

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

$$\phi(\mathbf{r}) = -\int\limits_{V_s} dV' g(\mathbf{r}, \mathbf{r}')Q(\mathbf{r}') + \int\limits_{V} dV' g(\mathbf{r}, \mathbf{r}')[k^2(\mathbf{r}') - k_0^2]\phi(\mathbf{r}'). \tag{36.3.4}$$

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

$$\phi(\mathbf{r}) = \phi_{inc}(\mathbf{r}) + \int_V dV' g(\mathbf{r}, \mathbf{r}')[k^2(\mathbf{r}') - k_0^2]\phi(\mathbf{r}'). \tag{36.3.5}$$

In the above equation, if the total field $\phi(\mathbf{r}')$ inside the volume V 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 domain. Consequently, (36.3.5) becomes the desired integral equation after rearrangement as

$$\phi_{inc}(\mathbf{r}) = \phi(\mathbf{r}) - \int_V dV' g(\mathbf{r}, \mathbf{r}')[k^2(\mathbf{r}') - k_b^2]\phi(\mathbf{r}'), \quad \mathbf{r} \in V. \tag{36.3.6}$$

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 as

$$\phi_{inc}(\mathbf{r}) = \phi(\mathbf{r}) - \mathcal{G}(\mathbf{r}, \mathbf{r}')\phi(\mathbf{r}'), \quad \mathbf{r} \in V, \tag{36.3.7}$$

where $\mathcal{G}$ is the integral operator in (36.3.6).[1] 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 as

$$\mathscr{L} f = g \tag{36.3.8}$$

where $\mathscr{L}$ is a linear operator, while f represents the unknown function $\phi(\mathbf{r})$ and g is the known function $\phi_{inc}(\mathbf{r})$.

[1] Sometimes, this is called the kernel of the integral equation.

36.3.2 Surface Integral Equation

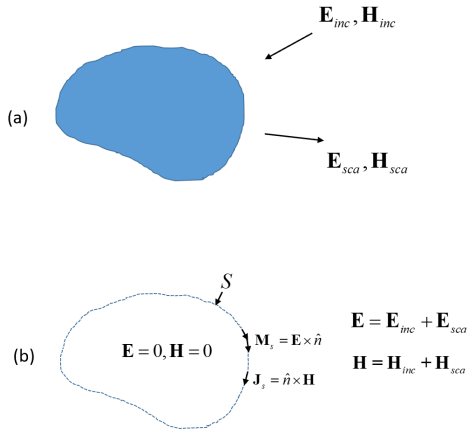


Figure 36.4: Geometry for the derivation of the surface-integral equation for vector electromagnetics waves. (a) The original electromagnetics scattering problem. (b) The equivalent electromagnetics problem by invoking equivalence principle.

The surface integral equation method is rather popular in a number of applications, because it employs a homogeneous-medium Green's function which is simple in form, and the unknowns reside on a surface rather than in a volume.[2]

The surface integral equation for vector electromagnetic field can be derived using the equivalence theorem also called the Love's equivalence theorem [182]. 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 31.1.

In this figure, the total field outside the scatterer, $\mathbf{E} = \mathbf{E}_{inc} + \mathbf{E}_{sca}$ 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! One can verify that this is the case by performing Gedanken experiments.

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 31.2. There, only the magnetic surface current

[2] These are sometimes called boundary integral equations method [221, 222].

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 (31.4.17)

$$\mathbf{E}_{sca}(\mathbf{r}) = \frac{1}{i\omega\epsilon}\nabla\times\nabla\times\oint_S dS'\, g(\mathbf{r}-\mathbf{r}')\,\hat{n}'\times\mathbf{H}(\mathbf{r}') = \frac{1}{i\omega\epsilon}\nabla\times\nabla\times\oint_S dS'\, g(\mathbf{r}-\mathbf{r}')\,\mathbf{J}_s(\mathbf{r}') \quad (36.3.9)$$

In the above, we have swapped $\mathbf{r}'$ and $\mathbf{r}$ compared to (31.4.17). 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 $\hat{n}\times\mathbf{E}_{sca} = -\hat{n}\times\mathbf{E}_{inc}$

$$-\hat{n}\times\mathbf{E}_{inc}(\mathbf{r}) = \hat{n}\times\frac{1}{i\omega\epsilon}\nabla\times\nabla\times\oint_S dS'\, g(\mathbf{r}-\mathbf{r}')\,\mathbf{J}_s(\mathbf{r}'). \qquad \mathbf{r}\in S \quad (36.3.10)$$

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 $\mathbf{J}_s$. 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

$$\mathscr{L}f = g \quad (36.4.1)$$

In the above, f is a functional vector which is the analogy of the vector $\mathbf{f}$ in matrix theory or linear algebra. 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 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. However, the index in this case is a continuum, and countably infinite. Hence, it corresponds to a vector of infinite dimension and it lives in an infinite dimensional space.[3]

To make such functions economical in storage, for instance, 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.[4] 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 hence, $f(\mathbf{r})$ can be thought of as a vector as well.

[3]When these functions are square integrable implying finite "energy", these infinite dimensional spaces are called Hilbert spaces.

[4]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.

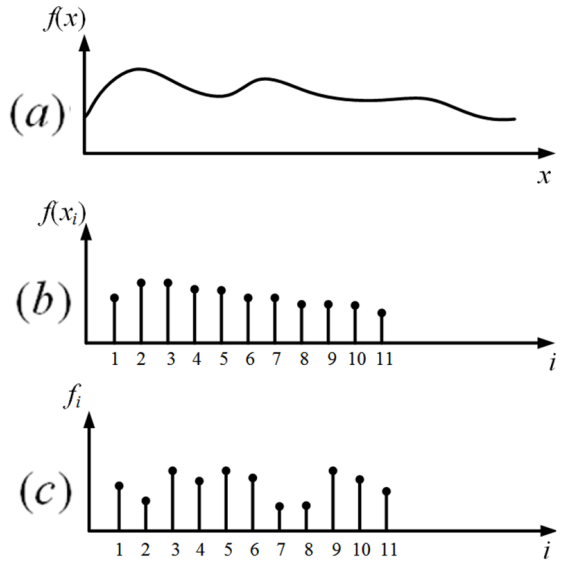


Figure 36.5: A function can be thought of as a vector. (a) A continuum function $f(x)$ plotted as a function of x. (b) A digitally sampled values of the same function. (c) When stored in a computer, it will be stored as an array vector.

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. It maps a function f defined in a Hilbert space V to g 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 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 $\mathscr{L}$.

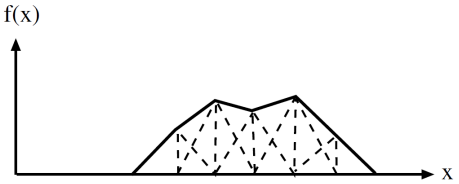
36.6 Approximating Operator Equations with Matrix Equations

36.6.1 Subspace Projection Methods

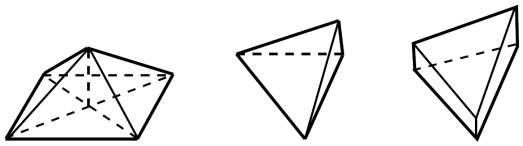
One main task of numerical method is first to approximate an operator equation $\mathscr{L}f = g$ by a matrix equation $\overline{\mathbf{L}} \cdot \mathbf{f} = \mathbf{g}$. To achieve the above, we first let

$$f \cong \sum_{n=1}^{N} a_n f_n = g \tag{36.6.1}$$

In the above, $f_n, n, \ldots, N$ are known functions called basis functions. 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.



(a) One Dimension



(b) Two Dimensions

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.

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.[5] For PDEs these functions are defined over a 3D coordinate space, while in SIEs, these functions are defined over a surface. In a 1D problems, these functions are defined over a 1D coordinate space.

36.6.2 Dual Spaces

The functions $w_m, m = 1, \ldots, N$ 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.

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

$$\begin{aligned} \left[\overline{\mathbf{L}}\right]_{mn} &= \langle w_m, \mathscr{L} f_n \rangle \\ [\mathbf{a}]_n = a_n, [\mathbf{g}]_m &= \langle w_m, g \rangle \end{aligned} \tag{36.6.6}$$

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 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.

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 [222].) When finite discrete basis are used to represent the surface unknowns, it is also called the boundary element method (BEM) [223]. But when this method is applied to solve PDEs, it is called the finite element method (FEM) [224–227], which is a rather popular method due to its simplicity.

[5] This is known as the reaction inner product [35, 50, 131]. As oppose to most math and physics literature, the energy inner product is used [131] where $\langle f_1, f_2 \rangle = \int d\mathbf{r} f_1^*(\mathbf{r}) f_2(\mathbf{r})$.

36.6.4 Mesh Generation

In order to approximate a function defined on an arbitrary shaped surface or volume by a finite sum of basis functions, it is best to mesh (tessellate or discretize) the surface and volume by meshes. In 2D, all shapes can be tessellated by unions of triangles, while a 3D volume can be meshed (tessellated) by unions of tetrahedrons. Such meshes are used not only in CEM, but in other fields such as solid mechanics. Hence, there are many "solid modeling" commercial software available to generate sophisticated meshes.

When a surface is curved, or of arbitrary shape, it can be meshed by union of triangles as shown in Figure 36.7. When a volume is of arbitrary shape or a volume is around an arbitrary shape object, it can be meshed by tetrahedrons as shown in Figure 36.8. Then basis functions as used in (36.6.1) are defined to interpolate the field between nodal values or values defined on the edges of a triangle or a tetrahedron.

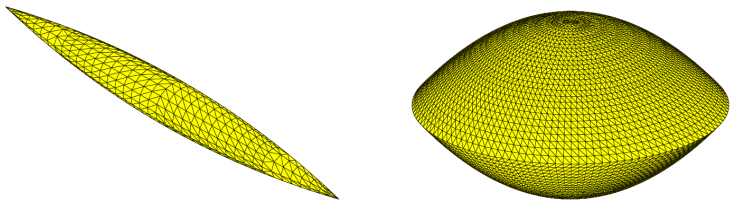


Figure 36.7: An arbitrary surface can be meshed by a union of triangles.

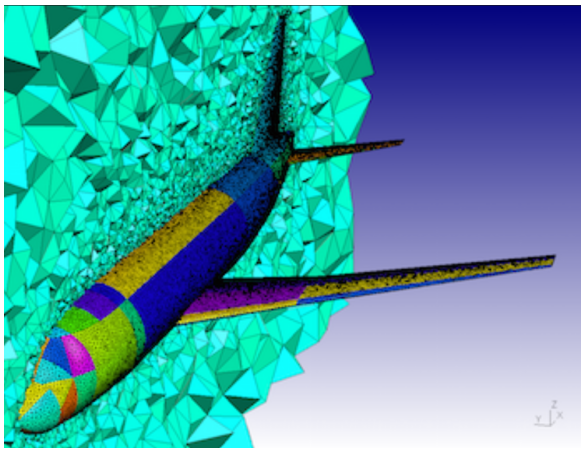


Figure 36.8: A volume region can be meshed by a union of tetrahedra. But the surface of the aircraft is meshed with a union of trianlges (courtesy of gmsh.info).

36.6.5 Differential Equation Solvers versus Integral Equation Solvers

As have been shown, the two classes of numerical solvers for Maxwell's equations consist of 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. The fields permeate all of space, and hence, the unknowns are volumetrically distributed. When the fields are digitized 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. That is integral equations are derived from Maxwell's equations using Green's function, where the unknowns now are surface sources such as surface electric and magnetic currents. Therefore, the unknowns are generally smaller, 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. Thus, the matrix systems generally are smaller. Once the currents are found, then the fields they generate can also be computed.

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. 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].

36.7 Solving Matrix Equation by Optimization

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 [228] or Kramer's rule [229]) have computational complexity[6] 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.

To this end, it is better to convert the solving of a matrix equation into an optimization problem. 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.[7] First, we will figure out how to find such a functional.

Consider a matrix equation given by

$$\overline{\mathbf{L}} \cdot \mathbf{f} = \mathbf{g} \tag{36.7.1}$$

[6]The scaling of computer time with respect to the number of unknowns (degrees of freedom) is known in the computer parlance as computational complexity.

[7]Functional is usually defined as a function of a function [35, 49]. Here, we include a function of a vector to be a functional as well.

For simplicity, we consider $\overline{\mathbf{L}}$ as a symmetric matrix.[8] Then the corresponding functional 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 called a quadratic functional because it is analogous to $I = Lx^2 - 2xg$, which is quadratic, in its simplest 1D rendition.

Taking 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. Then 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 symmetric, the first two terms are the same, and the above 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 $\delta I = 0$. Thus we conclude that at the optimal point (or the stationary point),

$$\overline{\mathbf{L}} \cdot \mathbf{f}_o = \mathbf{g} \tag{36.7.5}$$

Hence, the optimal point to the functional I in (36.7.2) is the solution to (36.7.1) or (36.7.5).

36.7.1 Gradient of a Functional

The above method, when applied to an infinite dimensional Hilbert space problem, is called 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 [230]. In this notation, the functional first variation δI in (36.7.4) becomes

$$\delta I = 2\delta f_j L_{ij} f_i - 2\delta f_j g_j \tag{36.7.6}$$

Also, 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 notation, it is easier to see what a functional derivative is. We can differentiate the above with respect to f_j easily to arrive at

$$\frac{\partial I}{\partial f_j} = 2L_{ij} f_i - 2g_j \tag{36.7.7}$$

[8] Functional for the asymmetric case can be found in *Chew, Waves and Fields in Inhomogeneous Media*, Chapter 5 [35].

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

$$\frac{\delta I}{\delta \mathbf{f}} = \nabla_{\mathbf{f}} I = 2\overline{\mathbf{L}} \cdot \mathbf{f} - 2\mathbf{g} \tag{36.7.8}$$

The left-hand side is a notation for the functional derivative or the gradient of a functional in a multi-dimensional space whichi 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. 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.[9] 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 $N_{iter}N^{\alpha}$ 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 solutions, to obtain an approximate solution where the error is acceptable, $N_{\text{iter}} \ll N$. The total solution time or solve time which is $N_{iter}N^{\alpha} \ll NN^{\alpha} \ll N^3$, resulting in great savings in computer time, especially if $\alpha = 1$. This is the case for FEM [227], [231], [232], [233], [226], and fast multipole algorithm [234], [235].

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.[10]

The value I is also called the cost function, and its minimum is sought in the seeking of the solution by gradient search methods. Detail discussion of these methods is given in [236]. Figure 36.9 shows the contour plot of a cost function in 2D. When the condition number[11] of the matrix $\overline{\mathbf{L}}$ is large (implying that the matrix is ill-conditioned), the contour plot will resemble a deep valley. And hence, the gradient search method will tend to zig-zag along the way as it finds the optimal solution. Therefore, convergence is slow for matrices with large condition numbers

[9] This is the case for many differential equation solvers such as finite-element method or finite-difference method.

[10] Chew et al, *Fast and Efficient Algorithms in CEM* [9].

[11] This is the ratio of the largest eigenvalue of the matrix to its smallest eigenvalue.

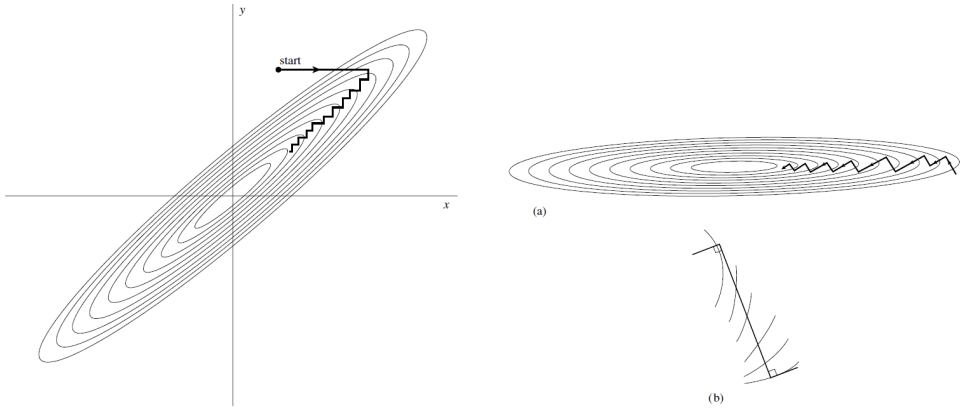


Figure 36.9: Plot of a 2D cost function, $I(x, y)$ for an ill-conditioned system (courtesy of Numerical Recipe [236]). A higher dimensional plot of this cost function will be difficult.

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.

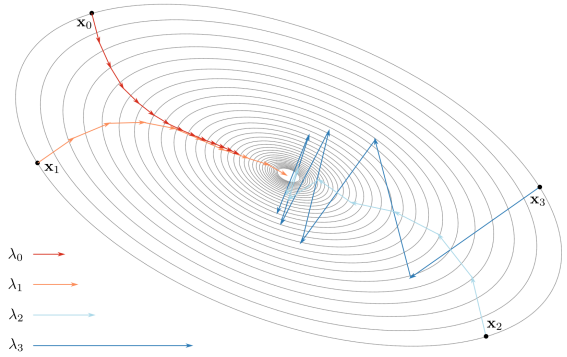


Figure 36.10: Gradient search or gradient descent method is finding an optimal point (courtesy of Y. Ioannou: https://blog.yani.io/sgd/).

Lecture 37

Finite Difference Method, Yee Algorithm

In this lecture, we are going to introduce one of the simplest methods to solve Maxwell's equations numerically. This is the finite-difference time-domain method first proposed by Yee [237] and popularized by Taflove [238]. Because of its simplicity, and that a simple Maxwell solver can be coded in one afternoon, 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 method 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 (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. Moreover, it is quite versatile, and given the present computer technology, it has been used with great success in solving many practical problems. This method is based on a simple Yee algorithm [237] and has been vastly popularized by Taflove [238, 239].

In the finite-difference method, continuous space-time is replaced with a discrete space-time. Then, 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 solve large matrices, resulting in great savings in computer time. Moreover, the matrix for the system of equations is never generated making this a matrix-free method. There is no need for matrix management as one writes this numerical solver. More recently,

[1] "rule" is used as a verb.

the development of parallel processor architectures in computers has also further enhanced the efficiency of the finite-difference time-domain scheme [240].

The finite-difference method is also described in numerous works (see, for example, Potter 1973 [241]; Taflove 1988 [238]; Ames 2014 [242]; and Morton 2019 [243].

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 [35, 244].

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. Then, the time derivative can be approximated in many ways. For example, a derivative can be approximated by forward, backward, and central finite difference formulas [245].

$$\text{Forward difference:}\quad \frac{\partial\phi(\mathbf{r},t)}{\partial t} \approx \frac{\phi(\mathbf{r},t+\Delta t)-\phi(\mathbf{r},t)}{\Delta t}, \tag{37.1.2}$$

$$\text{Backward difference:}\quad \frac{\partial\phi(\mathbf{r},t)}{\partial t} \approx \frac{\phi(\mathbf{r},t)-\phi(\mathbf{r},t-\Delta t)}{\Delta t}, \tag{37.1.3}$$

$$\text{Central difference:}\quad \frac{\partial\phi(\mathbf{r},t)}{\partial t} \approx \frac{\phi(\mathbf{r},t+\frac{\Delta t}{2})-\phi(\mathbf{r},t-\frac{\Delta t}{2})}{\Delta t}, \tag{37.1.4}$$

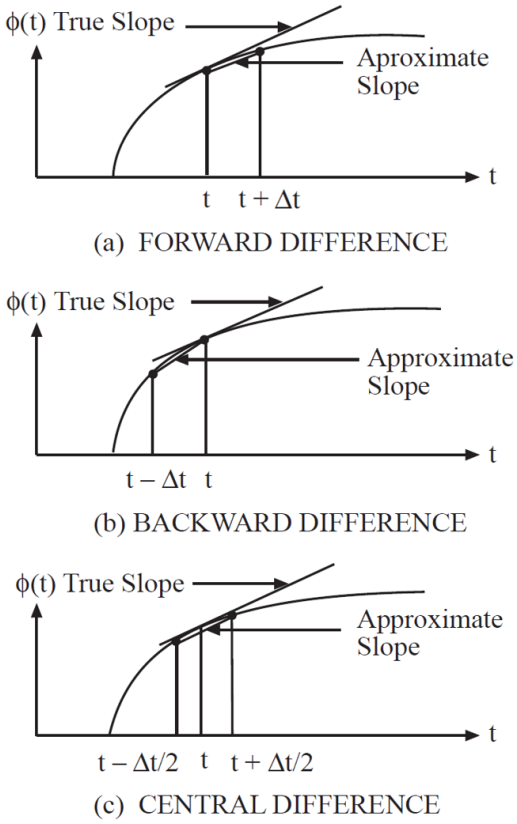


Figure 37.1: Different finite-difference approximations for the time derivative. One can eyeball that the central difference formula is the best. This can be further confirmed by a Taylor series analysis.

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 37.1. 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

$$\frac{\partial^2}{\partial t^2}\phi(\mathbf{r},t) \approx \frac{\partial}{\partial t}\left[\frac{\phi(\mathbf{r},t+\frac{\Delta t}{2}) - \phi(\mathbf{r},t-\frac{\Delta t}{2})}{\Delta t}\right] \tag{37.1.5}$$

$$\approx \frac{\phi(\mathbf{r},t+\Delta t) - 2\phi(\mathbf{r},t) + \phi(\mathbf{r},t-\Delta t)}{(\Delta t)^2}. \tag{37.1.6}$$

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 $\phi(\mathbf{r},t) = \phi(\mathbf{r},l\Delta t) = \phi^l(\mathbf{r})$, where l is an integer is 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 $t = l\Delta t$ and $t = (l-1)\Delta t$ for all $\mathbf{r}$, one can deduce $\phi(\mathbf{r},t)$ at $t = (l+1)\Delta t$. 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 $(x,y,z) = (m\Delta x, n\Delta y, p\Delta z)$, and then

$$\begin{aligned}\frac{\partial}{\partial z}\mu^{-1}\frac{\partial}{\partial z}\phi(x,y,z) \approx \frac{1}{(\Delta z)^2}\Big[&\mu^{-1}_{m,n,p+\frac{1}{2}}\phi_{m,n,p+1}\\ &-\left(\mu^{-1}_{m,n,p+\frac{1}{2}}+\mu^{-1}_{m,n,p-\frac{1}{2}}\right)\phi_{m,n,p}+\mu^{-1}_{m,n,p-\frac{1}{2}}\phi_{m,n,p-1}\Big].\end{aligned} \tag{37.1.12}$$

This cumbersome and laborious looking equation can be abbreviated if we define a central difference operator as[2]

$$\bar{\partial}_z \phi_m = \frac{1}{\Delta z}\left(\phi_{m+\frac{1}{2}} - \phi_{m-\frac{1}{2}}\right) \tag{37.1.13}$$

Then the right-hand side of the (37.1.12) can be written succinctly as

$$\frac{\partial}{\partial z}\mu^{-1}\frac{\partial}{\partial z}\phi(x,y,z) \approx \bar{\partial}_z \mu_{m,n,p}\bar{\partial}_z \phi_{m,n,p} \tag{37.1.14}$$

With similar approximations to the other terms in (37.1.9), (37.1.8) is now compactly written as

$$\begin{aligned}\phi^{l+1}_{m,n,p} = (\Delta t)^2 c^2_{m,n,p}\mu_{m,n,p}\left[\bar{\partial}_x \mu_{m,n,p}\bar{\partial}_x + \bar{\partial}_y \mu_{m,n,p}\bar{\partial}_y + \bar{\partial}_z \mu_{m,n,p}\bar{\partial}_z\right]\phi_{m,n,p} \\ + 2\phi^l_{m,n,p} - \phi^{l-1}_{m,n,p}.\end{aligned} \tag{37.1.15}$$

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. Hence, it is customary to store the medium property at the integer grid points for ease of book-keeping, and to approximate

$$\mu_{m+\frac{1}{2},n,p} \simeq \frac{1}{2}(\mu_{m+1,n,p} + \mu_{m,n,p}), \tag{37.1.16}$$

$$\mu_{m+\frac{1}{2},n,p} + \mu_{m-\frac{1}{2},n,p} \simeq 2\mu_{m,n,p}, \tag{37.1.17}$$

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 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 [246].

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 a numerical finite-difference scheme is stable. To do this, one performs the von Neumann stability analysis (von Neumann 1943 [247]) on Equation (37.1.18). We will assume the medium to be homogeneous to simplify the analysis.

As shown previously in Section 35.1, a point source gives rise to a spherical wave that can be expanded in terms of sum of plane waves in different directions. It also implies that any wave emerging from sources can be expanded in terms of sum of plane waves. This is the spirit of the spectral expansion method. So 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 as a trial solution

$$\phi(x,y,z,t) = A(t)e^{ik_x x + ik_y y + ik_z z}, \tag{37.1.19}$$

In discretized form, by letting $\Delta x = \Delta y = \Delta z = \Delta s$, it is just

$$\phi^l_{m,n,p} = A^l e^{ik_x m\Delta s + ik_y n\Delta s + ik_z p\Delta s}. \tag{37.1.20}$$

We can imagine that $A^l = A_0 e^{-i\omega l\Delta t}$, so that the above is actually a Fourier plane wave mode in the frequency domain. Using (37.1.20), it is easy to show that for the x space derivative,

$$\begin{aligned}\phi^l_{m+1,n,p} - 2\phi^l_{m,n,p} + \phi^l_{m-1,n,p} &= 2[\cos(k_x\Delta s) - 1]\phi^l_{m,n,p} \\ &= -4\sin^2\left(\frac{k_x\Delta s}{2}\right)\phi^l_{m,n,p}.\end{aligned} \tag{37.1.21}$$

The space derivatives in y and z directions can be similarly derived.

The time derivative can be similarly approximated, and it is approximately equal to

$$\frac{\partial^2}{\partial t^2}\phi(\mathbf{r},t)(\Delta t)^2 \approx \phi^{l+1}_{m,n,p} - 2\phi^l_{m,n,p} + \phi^{l-1}_{m,n,p}. \tag{37.1.22}$$

Substituting (37.1.20) into the above, we have the second time derivative being proportional to

$$\frac{\partial^2}{\partial t^2}\phi(\mathbf{r},t)(\Delta t)^2 \approx (A^{l+1} - 2A^l + A^{l-1})e^{ik_x m\Delta s + ik_y n\Delta s + ik_z p\Delta s} \tag{37.1.23}$$

To simplify further, one can assume that

$$A^{l+1} = gA^l. \tag{37.1.24}$$

This is commensurate with assuming that

$$A(t) = A_0 e^{-i\omega t} \tag{37.1.25}$$

where ω can be complex. In other words, our trial solution (37.1.19) is also a time-harmonic signal where ω can be complex. If the finite-difference scheme is unstable for such a signal, it is unstable for all signals.

Consequently, the time derivative is proportional to

$$\frac{\partial^2}{\partial t^2}\phi(\mathbf{r},t)(\Delta t)^2 \approx (g-2+g^{-1})\phi^l_{m,n,p} \tag{37.1.26}$$

We need to find the value of g for which the solution (37.1.20) satisfies (37.1.18). To this end, one uses (37.1.21) and (37.1.24) in (37.1.18), and repeating (37.1.21), which is for m variable in the x direction, for the n and p variables in the x and y directions as well, one obtains

$$\begin{aligned}(g-2+g^{-1})\phi^l_{m,n,p} &= -4\left(\frac{\Delta t}{\Delta s}\right)^2 c^2\left[\sin^2\left(\frac{k_x\Delta s}{2}\right)+\sin^2\left(\frac{k_y\Delta s}{2}\right)\right.\\ &\qquad\left.+\sin^2\left(\frac{k_z\Delta s}{2}\right)\right]\phi^l_{m,n,p}\\ &= -4r^2s^2\phi^l_{m,n,p},\end{aligned} \tag{37.1.27}$$

where

$$r=\left(\frac{\Delta t}{\Delta s}\right)c, \qquad s^2=\sin^2\left(\frac{k_x\Delta s}{2}\right)+\sin^2\left(\frac{k_y\Delta s}{2}\right)+\sin^2\left(\frac{k_z\Delta s}{2}\right). \tag{37.1.28}$$

Equation (37.1.27) implies that, for nonzero $\phi^l_{m,n,p}$,[3]

$$g^2-2g+4r^2s^2g+1=0, \tag{37.1.29}$$

or that

$$g=(1-2r^2s^2)\pm 2rs\sqrt{(r^2s^2-1)}\,. \tag{37.1.30}$$

In order for the solution to be stable, it is necessary that $|g|\le 1$. But if

$$r^2s^2<1, \tag{37.1.31}$$

the second term in (37.1.30) is pure imaginary, and

$$|g|^2=(1-2r^2s^2)^2+4r^2s^2(1-r^2s^2)=1, \tag{37.1.32}$$

when (37.1.31) is true. Therefore, stability is ensured. Since from (37.1.28), $s^2\le 3$ for all k_x, k_y, and k_z, from (37.1.31). From (37.1.31), we conclude that

$$r<\frac{1}{s}$$

But we also know that

$$\frac{1}{s}\ge\frac{1}{\sqrt{3}}$$

[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:)

The above two inequalities will be satisfied if the general condition is

$$r < \frac{1}{\sqrt{3}}, \qquad \text{or} \qquad \Delta t < \frac{\Delta s}{c\sqrt{3}}. \tag{37.1.33}$$

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

$$\Delta t < \frac{\Delta s}{c\sqrt{n}}. \tag{37.1.34}$$

One may ponder on the 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 (37.1.34) 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 (37.1.34), instability ensues. The above is also known as the CFL (Courant, Friedrichs, and Lewy 1928 [248]) stability criterion. It could be easily modified for $\Delta x \neq \Delta y \neq \Delta z$ [239].

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, to make things simpler, we will analyze the grid-dispersion error in a homogeneous free space medium.

To ascertain the grid-dispersion error, we assume that the solution is time-harmonic, or that $A^l = A_0 e^{-i\omega l \Delta t}$ in (37.1.20). In this case, the left-hand side of (37.1.27) becomes

$$\left(e^{-i\omega\Delta t} - 2 + e^{+i\omega\Delta t}\right) \phi^l_{m,n,p} = -4 \sin^2\left(\frac{\omega \Delta t}{2}\right) \phi^l_{m,n,p}. \tag{37.1.35}$$

Then, from (37.1.27), it follows that

$$\sin\left(\frac{\omega\Delta t}{2}\right) = rs, \tag{37.1.36}$$

where r and s are given in (37.1.28). Now, (37.1.36) 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 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

$$\omega = c\sqrt{k_x^2 + k_y^2 + k_z^2} = c|\mathbf{k}| = ck. \tag{37.1.37}$$

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 (37.1.36), because of the definition of s as given by (37.1.28), the dispersion relation between ω and $\mathbf{k}$ is not isotropic. This implies that plane waves propagating in different directions will have different phase velocities.

Equation (37.1.36) is the dispersion relation for the approximate solution. It departs from Equation (37.1.37), the exact dispersion relation, as a consequence of the finite-difference approximation. This departure gives rise to errors, which are the consequence of grid dispersion. For example, when c is a constant, (37.1.37) states that the phase velocities of plane waves of different wavelengths and directions are the same. However, this is not true for (37.1.36), as shall be shown.

Assuming s small, (37.1.36), after using Taylor series expansion, can be written as

$$\frac{\omega\Delta t}{2} = \sin^{-1} rs \cong rs + \frac{r^3 s^3}{6}. \tag{37.1.38}$$

When Δs is small, using the small argument approximation for the sine function, one obtains from (37.1.28)

$$s \simeq \frac{\Delta s}{2}(k_x^2 + k_y^2 + k_z^2)^{1/2} \tag{37.1.39}$$

Equation (37.1.38), by taking the higher-order Taylor expansion of (37.1.38), then becomes

$$\frac{\omega\Delta t}{2} \simeq r\frac{\Delta s}{2}(k_x^2 + k_y^2 + k_z^2)^{1/2}\,[1 - \delta] \tag{37.1.40}$$

where (see [35])

$$\delta = \frac{\Delta s^2}{24}\frac{k_x^4 + k_y^4 + k_z^4}{k_x^2 + k_y^2 + k_z^2} - \frac{r^2\Delta s^2}{24}(k_x^2 + k_y^2 + k_z^2) \tag{37.1.41}$$

From the above, one can see that 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. A soothing fact is that when $\Delta s \ll 1$, δ is small.

Since $\mathbf{k}$ is inversely proportional to wavelength λ, then δ in the correction to the above equation is proportional to $\Delta s^2/\lambda^2$. Therefore, to reduce the grid dispersion error, it is necessary for δ to be small or to have

$$\left(\frac{\Delta s}{\lambda}\right)^2 \ll 1. \tag{37.1.42}$$

When this is true, using the fact that $r = c\Delta t/\Delta s$, then (37.1.40) becomes

$$\frac{\omega}{c} \approx \sqrt{k_x^2 + k_y^2 + k_z^2}\,. \tag{37.1.43}$$

which is close to the dispersion relation of free space as indicated in (37.1.37). Consequently, in order for the finite-difference scheme to propagate a certain frequency content accurately, the grid size must be much less than the wavelength of the corresponding frequency. 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 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 [237]) 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 method. To derive it, Maxwell's equations are first written in cartesian coordinates:

$$-\frac{\partial B_x}{\partial t} = \frac{\partial E_z}{\partial y} - \frac{\partial E_y}{\partial z}, \tag{37.2.1}$$

$$-\frac{\partial B_y}{\partial t} = \frac{\partial E_x}{\partial z} - \frac{\partial E_z}{\partial x}, \tag{37.2.2}$$

$$-\frac{\partial B_z}{\partial t} = \frac{\partial E_y}{\partial x} - \frac{\partial E_x}{\partial y}, \tag{37.2.3}$$

$$\frac{\partial D_x}{\partial t} = \frac{\partial H_z}{\partial y} - \frac{\partial H_y}{\partial z} - J_x, \tag{37.2.4}$$

$$\frac{\partial D_y}{\partial t} = \frac{\partial H_x}{\partial z} - \frac{\partial H_z}{\partial x} - J_y, \tag{37.2.5}$$

$$\frac{\partial D_z}{\partial t} = \frac{\partial H_y}{\partial x} - \frac{\partial H_x}{\partial y} - J_z. \tag{37.2.6}$$

Before proceeding any further, it is prudent to rewrite the differential equation form of Maxwell's equations in their integral form. The first equation above can be rewritten as

$$-\frac{\partial}{\partial t}\iint_{\Delta S} B_x dS = \oint_{\Delta C} \mathbf{E}\cdot d\mathbf{l} \tag{37.2.7}$$

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 37.2. 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.

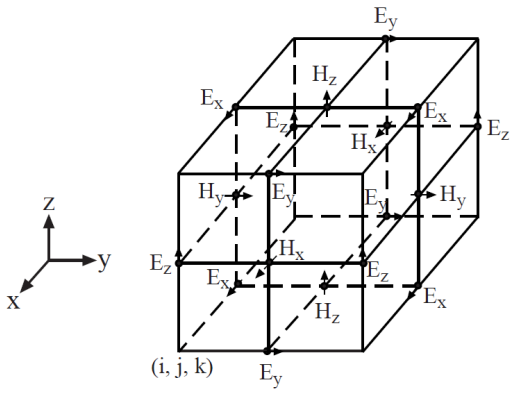


Figure 37.2: The assignment of fields on a grid in the Yee algorithm.

After denoting $f(m\Delta x, n\Delta y, p\Delta z, l\Delta t) = f^l_{m,n,p}$, and replacing derivatives with central finite-differences in accordance with Figure 37.2, (37.2.1) becomes

$$\frac{1}{\Delta t}\left[B^{l+\frac{1}{2}}_{x,m,n+\frac{1}{2},p+\frac{1}{2}} - B^{l-\frac{1}{2}}_{x,m,n+\frac{1}{2},p+\frac{1}{2}}\right] = \frac{1}{\Delta z}\left[E^l_{y,m,n+\frac{1}{2},p+1} - E^l_{y,m,n+\frac{1}{2},p}\right] - \frac{1}{\Delta y}\left[E^l_{z,m,n+1,p+\frac{1}{2}} - E^l_{z,m,n,p+\frac{1}{2}}\right]. \tag{37.2.8}$$

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 37.2, 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.[4]

[4]This algorithm is intimately related to differential forms which has given rise to the area of discrete exterior calculus [249].

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 37.2, and using the right-hand side of (37.2.8) and its equivalent, 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 satisfied. This is important in maintaining the stability of the numerical algorithm [250].

Furthermore, a similar approximation of (37.2.4) leads to

$$\begin{aligned}\frac{1}{\Delta t}\left[D^{l}_{x,m+\frac{1}{2},n,p} - D^{l-1}_{x,m+\frac{1}{2},n,p}\right] &= \frac{1}{\Delta y}\left[H^{l-\frac{1}{2}}_{z,m+\frac{1}{2},n+\frac{1}{2},p} - H^{l-\frac{1}{2}}_{z,m+\frac{1}{2},n-\frac{1}{2},p}\right] \\ &\quad -\frac{1}{\Delta z}\left[H^{l-\frac{1}{2}}_{y,m+\frac{1}{2},n,p+\frac{1}{2}} - H^{l-\frac{1}{2}}_{y,m+\frac{1}{2},n,p-\frac{1}{2}}\right] - J^{l-\frac{1}{2}}_{x,m+\frac{1}{2},n,p}.\end{aligned} \tag{37.2.10}$$

Also, 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. 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 way 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.

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

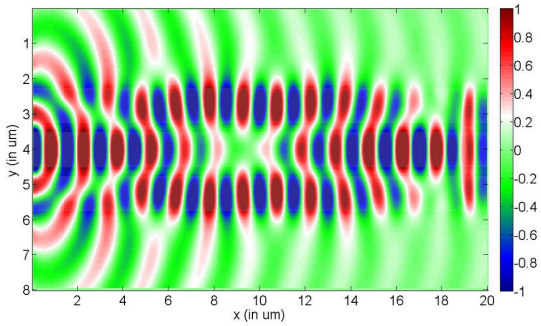


Figure 37.3: The 2D FDTD simulation of complicated optical waveguides (courtesy of Mathworks).

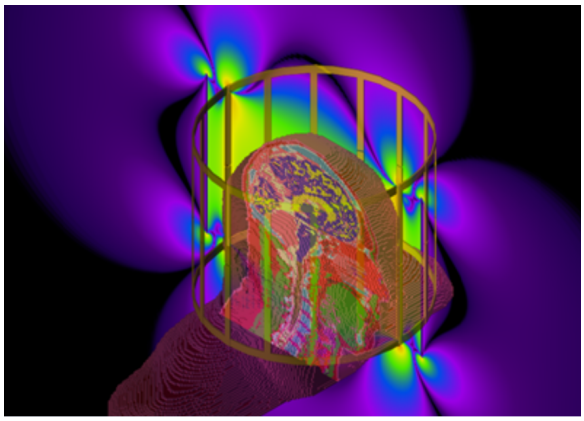


Figure 37.4: FDTD simulation of human head in a squirrel cage of an MRI (magnetic resonance imaging) system (courtesy of REMCOM).

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, 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 $\mathbf{J}$. 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 $\mathbf{x}$.

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 it are not reflected. This mimicks the physics of an infinitely large box.

This is analogous to experments 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 37.5). Figure 37.6 shows an acoustic version of anechoic chamber.

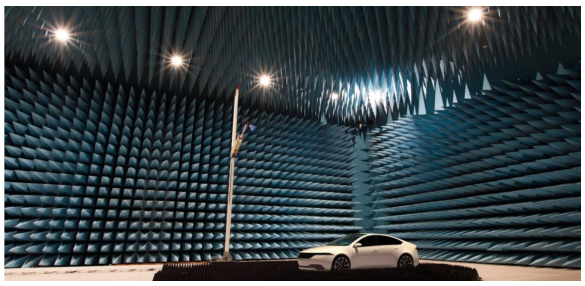


Figure 37.5: An anechoic chamber for radio frequency. In such an electromagnetically quiet chamber, interference from other RF equipment is minimized (courtesy of Panasonic).

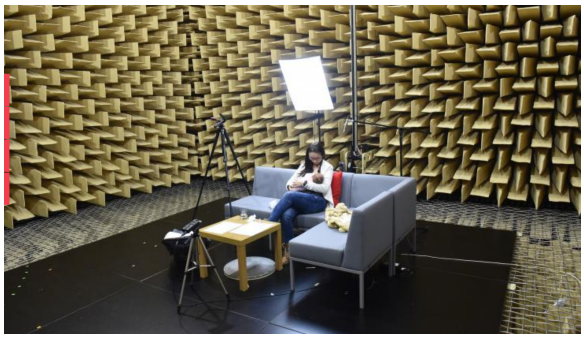


Figure 37.6: An acoustic anechoic chamber. In such a chamber, even the breast-feeding sound of a baby can be heard clearly (courtesy of AGH University, Poland).

By the same token, in order to simulate numerically an infinite 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 [35]).[5]

A recently invented ABC is the perfectly matched layers (PML) [251]. Also, another similar ABC is the stretched coordinates PML [252]. Figure 37.7 shows simulation results with and without stretched coordinates PMLs on the walls of the simulation domain [253].

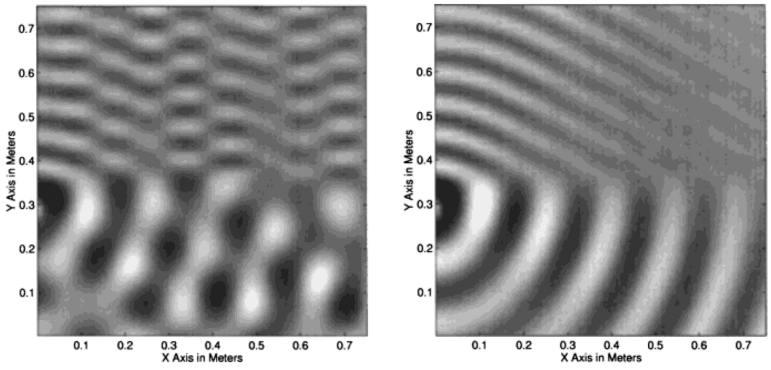


Figure 37.7: Simulation of a source on top of a half-space (left) without stretched coordinates PML; and (right) with stretched coordinats PML [253].

[5]ABCs are beyond the scope of these lecture notes.

Lecture 38

Quantum Theory of Light

The quantum theory of the world is the culmination of a series 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, we finally have the revelation about the perplexing rules. But we are grateful that, aided by experimental data, these laws of nature are deciphered by our predecessors.

It is important to know that with new quantum theory emerges the quantum theory of light. This theory is intimately related to Maxwell's equations as shall be seen. These new theories spawn the possibility for quantum technologies, one of which is quantum computing. Others are quantum communication, quantum crytography, quantum sensing and many more.

38.1 Historical Background on Quantum Theory

That light is a wave has been demonstrated by Newton's ring phenomenon [254] in the eighteenth century (1717) (see Figure 38.1). In 1801, Thomas Young demonstrated the double slit experiment for light [255] that further confirmed its wave nature (see Figure 38.2). But by the beginning of the 20-th century, one has to accept that light is both a particle, called a photon, carrying a quantum of energy with 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.

Theory

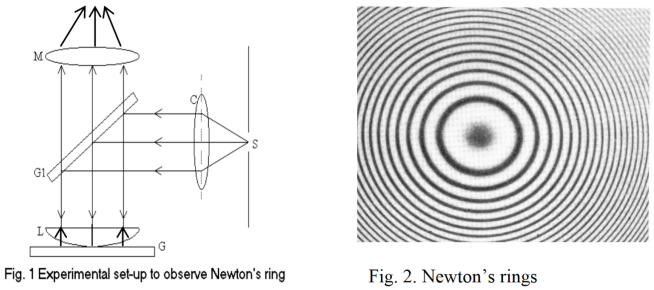


Fig. 1 Experimental set-up to observe Newton's ring

Fig. 2. Newton's rings

Figure 38.1: A Newton's rings experiment (courtesy of [254]).

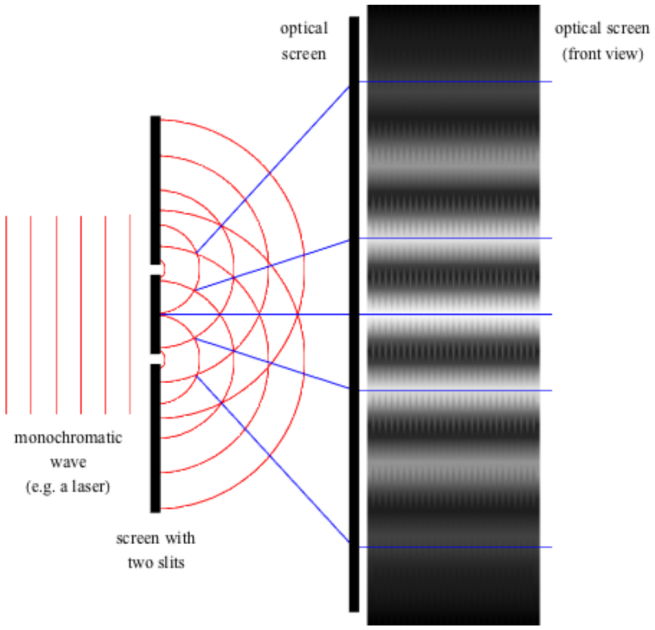


Figure 38.2: A Young's double-slit experiment (courtesy of [256]).

As mentioned above, quantum theory is a major intellectual achievement of the twentieth century, even though we are still discovering new knowledge 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) [257]. So fluid is not infinitely divisible: as when water is divided into smaller pieces,

we will eventually arrive at water molecule, H_2O, which is the fundamental building block of water.

In 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) [258] 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 $\hbar$ is now known as the Planck constant and $\hbar = h/2\pi = 6.626 \times 10^{-34}$ 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 $\approx 3\times10^{-19}$J $\approx 75 \quad k_BT$, where k_BT denotes the thermal energy from thermal law, and k_B is Boltzmann's constant. This is about 25 meV at room temperature.[1] A microwave photon has approximately 1×10^{-5} eV $\approx 10^{-2}$meV.

The second experimental evidence that light is quantized is the photo-electric effect [259]. 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. 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 [255].) In 1924, De Broglie [260] 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 [261]. The old way of describing particle motion is known as classical mechanics, and the new way of describing particle motion is known as quantum mechanics. Quantum mechanics is very much motivated by a branch of classical mechanics called Hamiltonian mechanics. We will first use Hamiltonian mechanics to study a simple pendulum and connect it with electromagnetic oscillations.

[1] 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 at room temperature.

38.2 Connecting Electromagnetic Oscillation to Simple Pendulum

The theory for quantization of electromagnetic field was started by Dirac in 1927 [3]. In the beginning, it was called quantum electrodynamics (QED) important for understanding particle physics phenomena and light-matter interactions [262]. Later on, it became important in quantum optics where quantum effects in electromagnetics technologies first emerged. Now, microwave photons are measurable and are important in quantum computers. Hence, quantum effects are important in the microwave regime as well.

Maxwell's equations originally were inspired by experimental findings of Maxwell's time, and he beautifully put them together using mathematics known during his time. But Maxwell's equations can also be "derived" using Hamiltonian mechanics and energy conservation. First, electromagnetic theory can be regarded as for describing an infinite set of coupled harmonic oscillators. In one dimension, when a wave propagates on a string, or an electromagnetic wave propagates on a transmission line, they can be regarded as propagating on a set of coupled harmonic oscillators as shown in Figure 38.3. Maxwell's equations describe the waves travelling in 3D space due to the coupling between an infinite set of harmonic oscillators. In fact, methods have been developed to solve Maxwell's equations using transmission-line-matrix (TLM) method [263], or the partial element equivalent circuit (PEEC) method [180]. In materials, these harmonic oscillators are atoms or molecules, but vacuum they can be thought of as electron-positron pairs (e-p pairs). Electrons are matters, while positrons are anti-matters. Together, in their quiescent state, they form vacuum or "nothingness". Hence, vacuum can support the propogation of electromagnetic waves through vast distances: we have received light from galaxies many light-years away.

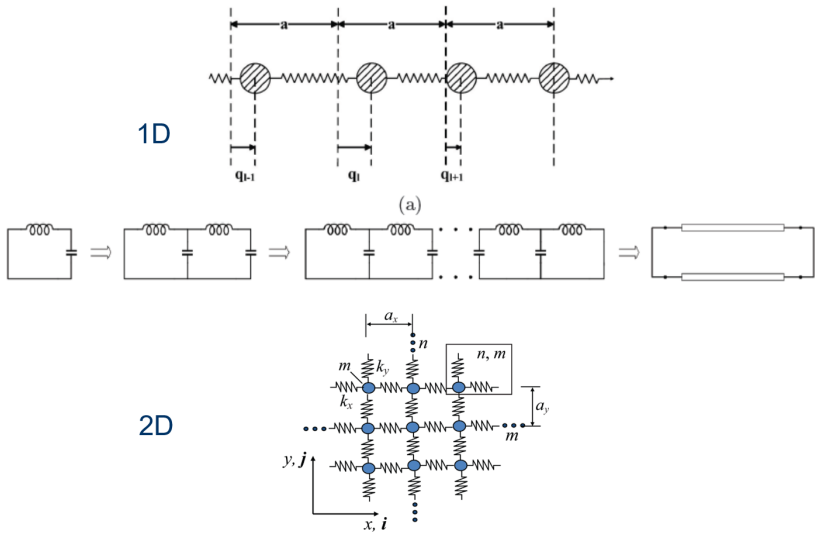


Figure 38.3: Maxwell's equations describe the coupling of harmonic oscillators in a 3D space. This is similar to waves propagating on a string or a 1D transmission line, or a 2D array of coupled oscillators. The saw-tooth symbol in the figure represents a spring.

The cavity modes in electromagnetics are similar to the oscillation of a pendulum in simple harmonic motion. To understand the quantization of electromagnetic field, we start by connecting these cavity-mode oscillations to the oscillations of a simple pendulum. It is to be noted that fundamentally, electromagnetic oscillation exists because of displacement current. Displacement current exists even in vacuum because vacuum is polarizable, namely that $\mathbf{D} = \varepsilon \mathbf{E}$ where for vacuum, $\varepsilon = \varepsilon_0$. Furthermore, displacement current exists because of the $\partial \mathbf{D}/\partial t$ term in the generalized Ampere's law added by Maxwell, namely,

$$\nabla \times \mathbf{H} = \frac{\partial \mathbf{D}}{\partial t} + \mathbf{J} \tag{38.2.1}$$

Together with Faraday's law that

$$\nabla \times \mathbf{E} = -\frac{\partial \mathbf{B}}{\partial t} \tag{38.2.2}$$

(38.2.1) and (38.2.2) together allow for the existence of wave. The coupling between the two equations gives rise to the "springiness" of electromagnetic fields.

Wave exists due to the existence of coupled harmonic oscillators, and at a fundamental level, these harmonic oscillators are electron-positron (e-p) pairs. The fact that they are coupled allows waves to propagate through space, and even in vacuum.

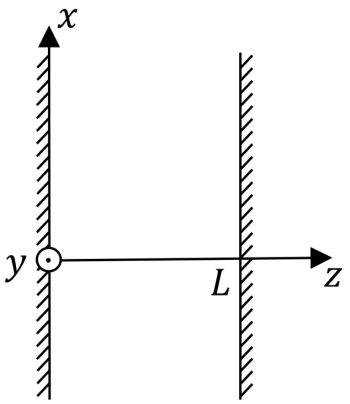


Figure 38.4: A one-dimensional cavity solution to Maxwell's equations is one of the simplest way to solve Maxwell's equations.

To make the problem simpler, we can start by looking at a one dimensional cavity formed by two PEC (perfect electric conductor) plates as shown in Figure 38.4. Assume source-free Maxwell's equations in between the plates and letting $\mathbf{E} = \hat{x}E_x$, $\mathbf{H} = \hat{y}H_y$, then (38.2.1) and (38.2.2) become

$$\frac{\partial}{\partial z}H_y = -\varepsilon\frac{\partial}{\partial t}E_x \tag{38.2.3}$$

$$\frac{\partial}{\partial z}E_x = -\mu\frac{\partial}{\partial t}H_y \tag{38.2.4}$$

The above are similar to the telegrapher's equations. We can combine them to arrive at

$$\frac{\partial^2}{\partial z^2}E_x = \mu\varepsilon\frac{\partial^2}{\partial t^2}E_x \tag{38.2.5}$$

There are infinitely many ways to solve the above partial differential equation. But here, we use separation of variables to solve the above by letting $E_x(z,t) = E_0(t)f(z)$. Then we arrive at two separate equations that

$$\frac{d^2E_0(t)}{dt^2} = -\omega_l^2 E_0(t) \tag{38.2.6}$$

and

$$\frac{d^2f(z)}{dz^2} = -\omega_l^2\mu\varepsilon f(z) \tag{38.2.7}$$

where ω_l^2 is the separation constant. There are infinitely many ways to solve the above equations which are also eigenvalue equations where ω_l^2 and $\omega_l^2\mu\varepsilon$ are eigenvalues for the first

and the second equations, respectively. The general solution for (38.2.7) is that

$$E_0(t) = E_0 \cos(\omega_l t + \psi) \tag{38.2.8}$$

In the above, ω_l, which is related to the separation constant, is yet indeterminate. To make ω_l^2 determinate, we need to impose boundary conditions. A simple way is to impose homogeneous Dirchlet boundary conditions that $f(z) = 0$ at $z = 0$ and $z = L$. This implies that $f(z) = \sin(k_l z)$. In order to satisfy the boundary conditions at $z = 0$ and $z = L$, one deduces that

$$k_l = \frac{l\pi}{L}, \qquad l = 1, 2, 3, \ldots \tag{38.2.9}$$

Then,

$$\frac{\partial^2 f(z)}{\partial z^2} = -k_l^2 f(z) \tag{38.2.10}$$

where $k_l^2 = \omega_l^2 \mu\epsilon$. Hence, $k_l = \omega_l/c$, and the above solution can only exist for discrete frequencies or that

$$\omega_l = \frac{l\pi}{L} c, \qquad l = 1, 2, 3, \ldots \tag{38.2.11}$$

These are the discrete resonant frequencies ω_l of the modes of the 1D cavity.

The above solutions for $E_x(z, t)$ can be thought of as the collective oscillations of coupled harmonic oscillators forming the modes of the cavity. At the fundamental level, these oscillations are oscillators made by electron-positron pairs. But macroscopically, their collective resonances manifest themselves as giving rise to infinitely many electromagnetic cavity modes. The amplitudes of these modes, $E_0(t)$ are simple harmonic oscillations.

The resonance between two parallel PEC plates is similar to the resonance of a transmission line of length L shorted at both ends. One can see that the resonance of a shorted transmission line is similar to the coupling of infnitely many LC tank circuits. To see this, as shown in Figure 38.3, we start with a single LC tank circuit as a simple harmonic oscillator with only one resonant frequency. When two LC tank circuits are coupled to each other, they will have two resonant frequencies. For N of them, they will have N resonant frequencies. For a continuum of them, they will be infinitely many resonant frequencies or modes as indicated by Equation (38.2.9).

What is more important is that the resonance of each of these modes is similar to the resonance of a simple pendulum or a simple harmonic oscillator. For a fixed point in space, the field due to this oscillation is similar to the oscillation of 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

$$m\frac{d^2x}{dt^2} + \kappa x = 0 \tag{38.2.12}$$

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. By letting[2]

$$x = x_0 e^{-i\omega t} \tag{38.2.13}$$

the above becomes

$$-m\omega^2 x_0 + \kappa x_0 = 0 \tag{38.2.14}$$

Again, a non-trivial solution is possible only at the resonant frequency of the oscillator or that when $\omega = \omega_0$ where

$$\omega_0 = \sqrt{\frac{\kappa}{m}} \tag{38.2.15}$$

This is the eigensolution of (38.2.12) with eigenvalue ω_0^2.

38.3 Hamiltonian Mechanics

Equation (38.2.12) can be derived by Newton's law but it can also be derived via Hamiltonian mechanics as well. Since Hamiltonian mechanics motivates quantum mechanics, we will look at the Hamiltonian mechanics view of the equation of motion (EOM) of a simple pendulum given by (38.2.12).

Hamiltonian mechanics, developed by Hamilton (1805-1865) [264], is motivated by energy conservation [265]. The Hamiltonian H of a system is given by its total energy, namely that

$$H = T + V \tag{38.3.1}$$

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

$$T = \frac{1}{2}mv^2 = \frac{1}{2m}m^2v^2 = \frac{p^2}{2m} \tag{38.3.2}$$

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, is given by

$$V = \frac{1}{2}\kappa x^2 = \frac{1}{2}m\omega_0^2 x^2 \tag{38.3.3}$$

Hence, the Hamiltonian is given by

$$H = T + V = \frac{p^2}{2m} + \frac{1}{2}m\omega_0^2 x^2 \tag{38.3.4}$$

[2]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.

At any instant of time t, we assume that $p(t) = mv(t) = m\frac{d}{dt}x(t)$ is independent of $x(t)$.[3] In other words, they can vary independently of each other. But $p(t)$ and $x(t)$ have to time evolve to conserve energy to keep H, the total energy, constant or independent of time. In other words,

$$\frac{d}{dt}H\left(p(t), x(t)\right) = 0 = \frac{dp}{dt}\frac{\partial H}{\partial p} + \frac{dx}{dt}\frac{\partial H}{\partial x} \tag{38.3.5}$$

Therefore, the Hamilton equations of motion are derived to be[4]

$$\frac{dp}{dt} = -\frac{\partial H}{\partial x}, \qquad \frac{dx}{dt} = \frac{\partial H}{\partial p} \tag{38.3.6}$$

From (38.3.4), we gather that

$$\frac{\partial H}{\partial x} = m\omega_0^2 x, \qquad \frac{\partial H}{\partial p} = \frac{p}{m} \tag{38.3.7}$$

Applying (38.3.6), we have[5]

$$\frac{dx}{dt} = \frac{p}{m}, \qquad \frac{dp}{dt} = -m\omega_0^2 x \tag{38.3.8}$$

Combining the two equations in (38.3.8) above, we have

$$m\frac{d^2x}{dt^2} = -m\omega_0^2 x = -\kappa x \tag{38.3.9}$$

which is also derivable by Newton's law.

A typical harmonic oscillator solution to (38.3.9) is

$$x(t) = x_0 \cos(\omega_0 t + \psi) \tag{38.3.10}$$

The corresponding $p(t) = m\frac{dx}{dt}$ is

$$p(t) = -mx_0\omega_0 \sin(\omega_0 t + \psi) \tag{38.3.11}$$

Hence

$$\begin{aligned} H &= \frac{1}{2}m\omega_0^2 x_0^2 \sin^2(\omega_0 t + \psi) + \frac{1}{2}m\omega_0^2 x_0^2 \cos^2(\omega_0 t + \psi) \\ &= \frac{1}{2}m\omega_0^2 x_0^2 = E \end{aligned} \tag{38.3.12}$$

And the total energy E is a constant of motion (physicists parlance for a time-independent variable), it depends only on the amplitude x_0 of the oscillation.

[3] $p(t)$ and $x(t)$ are termed conjugate variables in many textbooks.

[4] 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 [265].

[5] We can also calibrate our clock here so that it agrees with our definition of momentum in the ensuing equation.

38.4 Schrödinger Equation (1925)

Having seen the Hamiltonian mechanics for describing a simple pendulum which is homomorphic to a cavity resonator, we shall next see the quantum mechanics description of the same simple pendulum: In other words, we will look at the quantum pendulum. To this end, we will invoke Schrödinger equation.

Schrödinger equation cannot be derived just as in the case Maxwell's equations. It is a wonderful result of a postulate and a guessing game based on experimental observations [71, 72]. Hamiltonian mechanics says that

$$H = \frac{p^2}{2m} + \frac{1}{2}m\omega_0^2 x^2 = E \tag{38.4.1}$$

where E is the total energy of the oscillator, or pendulum. In classical mechanics, the position x of the particle associated with the pendulum is known with great certainty. But in the quantum world, this position x of the quantum particle is uncertain and is fuzzy. As shall be seen later, x is a random variable.[6]

To build this uncertainty into a quantum harmonic oscillator, we have to look at it from the quantum world. The position of the particle is described by a wave function,[7] 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 and that $E = \hbar\omega$ from Planck's law and the photo-electric effect. Equation (38.4.1) can be written more suggestively as

$$\frac{\hbar^2 k^2}{2m} + \frac{1}{2}m\omega_0^2 x^2 = \hbar\omega \tag{38.4.2}$$

To add more depth to the above equation, one lets the above become an operator equation that operates on a wave function $\psi(x,t)$ so that

$$-\frac{\hbar^2}{2m}\frac{\partial^2}{\partial x^2}\psi(x,t) + \frac{1}{2}m\omega_0^2 x^2\psi(x,t) = i\hbar\frac{\partial}{\partial t}\psi(x,t) \tag{38.4.3}$$

If the wave function is of the form

$$\psi(x,t) \sim e^{ikx-i\omega t} \tag{38.4.4}$$

then upon substituting (38.4.4) back into (38.4.3), we retrieve (38.4.2).

Equation (38.4.3) is Schrödinger 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

$$\psi(x,t) = e^{ikx-i\omega t}A(x,t) \tag{38.4.5}$$

[6]For lack of a better notation, we will use x to both denote a position in classical mechanics as well as a random variable in quantum theory.

[7]Since a function is equivalent to a vector, and this wave function describes the state of the quantum system, this is also called a state vector.

where $A(x,t)$ is a slowly varying function of x and t, compared to $e^{ikx-i\omega t}$.[8] In other words, this is the expression for a wave packet. With this wave packet, the $\partial^2/\partial x^2$ can be again approximated by $-k^2$ in the short-wavelength limit, as has been done in the paraxial wave approximation. Furthermore, if the signal is assumed to be quasi-monochromatic, then $i\hbar\partial/\partial_t\psi(x,t) \approx \hbar\omega$, we again retrieve the classical equation in (38.4.2) from (38.4.3). Hence, the classical equation (38.4.2) is a short wavelength, monochromatic approximation of Schrödinger equation. However, as we shall see, the solutions to Schrödinger equation are not limited to just wave packets described by (38.4.5).

In classical mechanics, the position of a particle is described by the variable x, but in the quantum world, the position of a particle x is a random variable. This property needs to be related to the wavefunction that is the solution to Schrödinger equation.

For this course, we need only to study the one-dimensional Schrödinger equation. The above can be converted into eigenvalue problem, just as in waveguide and cavity problems, using separation of variables, by letting[9]

$$\psi(x,t) = \psi_n(x)e^{-i\omega_n t} \tag{38.4.6}$$

By so doing, (38.4.3) becomes an eigenvalue problem

$$\left[-\frac{\hbar^2}{2m}\frac{d^2}{dx^2} + \frac{1}{2}m\omega_0^2 x^2\right]\psi_n(x) = E_n\psi_n(x) \tag{38.4.7}$$

where $E_n = \hbar\omega_n$ is the eigenvalue while $\psi_n(x)$ is the corresponding eigenfunction.

The parabolic x^2 potential profile is also known as a potential well as it can provide the restoring force to keep the particle bound to the well classically (see Section 38.3 and (38.3.8)). The above equation is also similar to the electromagnetic equation for a dielectric slab waveguide, where the second term is a dielectric profile (mind you, varying in the x direction) that can trap a waveguide mode. Therefore, the potential well is a trap for the particle both in classical mechanics or in wave physics.

The above equation (38.4.7) can be solved in closed form in terms of Hermite-Gaussian functions (1864) [266], or that

$$\psi_n(x) = \sqrt{\frac{1}{2^n n!}}\sqrt{\frac{m\omega_0}{\pi\hbar}}e^{-\frac{m\omega_0}{2\hbar}x^2}H_n\left(\sqrt{\frac{m\omega_0}{\hbar}}x\right) \tag{38.4.8}$$

where $H_n(y)$ is a Hermite polynomial, and the eigenvalues are found in closed form as

$$E_n = \left(n + \frac{1}{2}\right)\hbar\omega_0 \tag{38.4.9}$$

Here, the eigenfunction or eigenstate $\psi_n(x)$ is known as the photon number state (or just a number state) of the solution. It corresponds to having n "photons" in the oscillation. If this is conceived as the collective oscillation of the e-p pairs in a cavity, there are n photons

[8] Recall that this is similar in spirit when we study high frequency solutions of Maxwell's equations and paraxial wave approximation.

[9] Mind you, the following is ω_n, not ω_0.

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 $\psi_0(x)$. This is the zero-point energy state. This state is there even if the system is at its lowest energy state.

It is to be noted that in the quantum world, the position x of the pendulum is random. Moreover, this position $x(t)$ is mapped to the amplitude $E_0(t)$ of the field. Hence, it is the amplitude of an electromagnetic oscillation that becomes uncertain and fuzzy as shown in Figure 38.5.

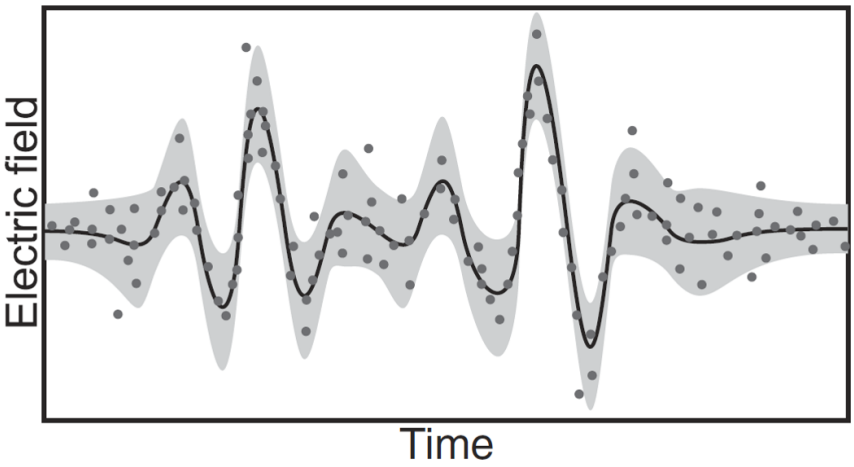


Figure 38.5: Schematic representation of the randomness of measured electric field. The electric field amplitude maps to the displacement (position) of the quantum harmonic oscillator, which is a random variable (courtesy of Kira and Koch [267]).

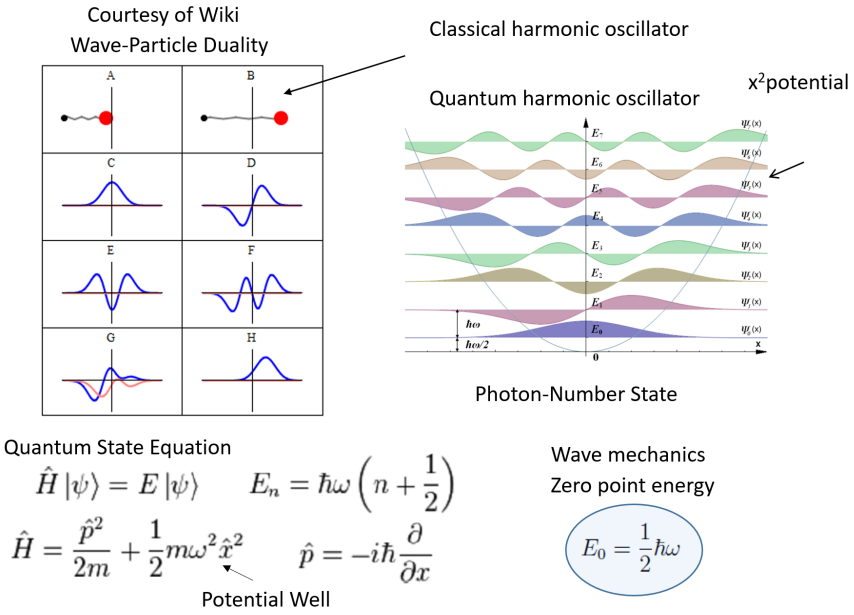


Figure 38.6: Plots of the eigensolutions of the quantum harmonic oscillator (courtesy of Wikipedia [268]).

38.5 Some Quantum Interpretations–A Preview

Schrödinger used this equation with resounding success. He derived a three-dimensional version of this to study the wave function and eigenvalues of a hydrogen atom. These eigenvalues E_n for a hydrogen atom agreed well with experimental observations that had eluded scientists for decades. Schrödinger did not actually understand what these wave functions meant. It was Max Born (1926) who gave a physical interpretation of these wave functions.

As mentioned before, in the quantum world, a position x is now a random variable. There is a probability distribution function (PDF) associated with this random variable x. This PDF for x is related to the a wave function $\psi(x,t)$, and it is given $|\psi(x,t)|^2$. Then according to probability theory, the probability of finding the particles in the interval[10] $[x, x+\Delta x]$ is $|\psi(x,t)|^2\Delta x$. Since $|\psi(x,t)|^2$ is a probability density function (PDF), and it is necessary that

$$\int_{-\infty}^{\infty} dx|\psi(x,t)|^2 = 1 \tag{38.5.1}$$

The average value or expectation value of the random variable x is now given by

$$\int_{-\infty}^{\infty} dx x|\psi(x,t)|^2 = \langle x(t)\rangle = \bar{x}(t) \tag{38.5.2}$$

[10]This is the math notation for an interval [,].

This is not the most ideal notation, since although x is not a function of time, its expectation value with respect to a time-varying function, $\psi(x,t)$, can be time-varying.

Notice that in going from (38.4.1) to (38.4.3), or from a classical picture to a quantum picture, we have let the momentum become p, originally a scalar number in the classical world, become a differential operator, namely that[11]

$$p \to \hat{p} = -i\hbar \frac{\partial}{\partial x} \tag{38.5.3}$$

The momentum p of a particle now also becomes uncertain and is a random varible: its expectation value is given by[12]

$$\int_{\infty}^{\infty} dx \psi^*(x,t) \hat{p} \psi(x,t) = -i\hbar \int_{-\infty}^{\infty} dx \psi^*(x,t) \frac{\partial}{\partial x} \psi(x,t) = \langle \hat{p}(t) \rangle = \bar{p}(t) \tag{38.5.4}$$

The expectation values of position x and the momentum operator $\hat{p}$ are measurable in the laboratory. Hence, they are also called observables.

38.5.1 Matrix or Operator Representations

We have seen in computational electromagnetics that an operator can be projected into a smaller subspace and manifests itself in different representations. Hence, an operator in quantum theory can have different representations depending on the space chosen. For instance, given a matrix equation

$$\overline{\mathbf{P}} \cdot \mathbf{x} = \mathbf{b} \tag{38.5.5}$$

we can find a unitary operator $\overline{\mathbf{U}}$ with the property $\overline{\mathbf{U}}^{\dagger} \cdot \overline{\mathbf{U}} = \overline{\mathbf{I}}$. Then the above equation can be rewritten as

$$\overline{\mathbf{U}} \cdot \overline{\mathbf{P}} \cdot \mathbf{x} = \overline{\mathbf{U}} \cdot \mathbf{b} \to \overline{\mathbf{U}} \cdot \overline{\mathbf{P}} \cdot \overline{\mathbf{U}}^{\dagger} \cdot \overline{\mathbf{U}} \cdot \mathbf{x} = \overline{\mathbf{U}} \cdot \mathbf{b} \tag{38.5.6}$$

Then a new equation is obtained such that

$$\overline{\mathbf{P}}' \cdot \mathbf{x}' = \mathbf{b}', \qquad \overline{\mathbf{P}}' = \overline{\mathbf{U}} \cdot \overline{\mathbf{P}} \cdot \overline{\mathbf{U}}^{\dagger}, \qquad \mathbf{x}' = \overline{\mathbf{U}} \cdot \mathbf{x} \qquad \mathbf{b}' = \overline{\mathbf{U}} \cdot \mathbf{b} \tag{38.5.7}$$

The operators we have encountered thus far in Schrödinger equation are in coordinate space representations.[13] In coordinate space representation, the momentum operator $\hat{p} = -i\hbar \frac{\partial}{\partial x}$, and the variable x can be regarded as a position operator in coordinate space representation. The operator $\hat{p}$ and x do not commute. In other words, it can be shown that

$$[\hat{p}, x] = \left[-i\hbar \frac{\partial}{\partial x}, x \right] = -i\hbar \tag{38.5.8}$$

In the classical world, $[p, x] = 0$, but not in the quantum world. In the equation above, we can elevate x to become an operator by letting $\hat{x} = x\hat{I}$, where $\hat{I}$ is the identity operator. Then

[11] We use $\hat{}$ to denote a quantum operator.

[12] This concept of the average of an operator seldom has an analogue in an intro probality course, but it is called the expectation value of an operator in quantum theory.

[13] Or just coordinate representation.

both $\hat{p}$ and $\hat{x}$ are now operators, and are on the same footing. In this manner, we can rewrite equation (38.5.8) above as

$$[\hat{p}, \hat{x}] = -i\hbar\hat{I} \tag{38.5.9}$$

By performing unitary transformation, it can be shown that the above identity is coordinate independent: it is true in any representation of the operators.

It can be shown easily that when two operators share the same set of eigenfunctions, they commute. When two operators $\hat{p}$ and $\hat{x}$ do not commute, it means that the expectation values of quantities associated with the operators, $\langle\hat{p}\rangle$ and $\langle\hat{x}\rangle$, cannot be determined to arbitrary precision simultaneously. For instance, $\hat{p}$ and $\hat{x}$ correspond to random variables, then the standard deviation of their measurable values, or their expectation values, obey the uncertainty principle relationship that[14]

$$\Delta p \Delta x \geq \hbar/2 \tag{38.5.10}$$

where Δp and Δx are the standard deviation of the random variables p and x.

38.6 Bizarre Nature of the Photon Number States

The photon number states are successful in predicting that the collective e-p oscillations are associated with n photons embedded in the energy of the oscillating modes. However, these number states are bizarre: The expectation values of the position of the quantum pendulum associated these states are always zero. To illustrate further, we form the wave function with a photon-number state

$$\psi(x,t) = \psi_n(x)e^{-i\omega_n t}$$

Previously, since the $\psi_n(x)$ are eigenfunctions, they are mutually orthogonal and they can be orthonormalized meaning that

$$\int_{-\infty}^{\infty} dx\psi_n^*(x)\psi_{n'}(x) = \delta_{nn'} \tag{38.6.1}$$

Then one can easily show that the expectation value of the position of the quantum pendulum in a photon number state is

$$\langle x(t)\rangle = \bar{x}(t) = \int_{-\infty}^{\infty} dxx|\psi(x,t)|^2 = \int_{-\infty}^{\infty} dxx|\psi_n(x)|^2 = 0 \tag{38.6.2}$$

because the integrand is always odd symmetric. In other words, the expectation value of the position x of the pendulum is always zero. It can also be shown that the expectation value of the momentum operator $\hat{p}$ is also zero for these photon number states. Hence, there are no classical oscillations that resemble them. Therefore, one has to form new wave functions by linear superposing these photon number states into a coherent state. This will be the discussion in the next lecture.

[14] The proof of this is quite straightforward but is outside the scope of this course.

Lecture 39

Quantum Coherent State of Light

As mentioned in the previous lecture, the discovery of Schrödinger wave equation was a resounding success. When it was discovered by Schrödinger, and appplied to a very simple hydrogen atom, its eigensolutions, especially the eigenvalues E_n coincide beautifully with spectroscopy experiment of the hydrogen atom. Since the electron wavefunctions inside a hydrogen atom does not have a classical analog, less was known about these wavefunctions. But in QED and quantum optics, the wavefunctions have to be connected with classical electromagnetic oscillations. As seen previously, electromagnetic oscillations resemble those of a pendulum. The original eigenstates of the quantum pendulum were the photon number states also called the Fock states. The connection to the classical pendulum was tenuous, but required by the correspondence principle–quantum phenomena resembles classical phenomena in the high energy limit. This connection was finally established by the establishment of the coherent state.

39.1 The Quantum Coherent State

We have seen that a photon number states[1] of a quantum pendulum do not have a classical correspondence as the average or expectation values of the position and momentum of the pendulum are always zero for all time for this state. Therefore, we have to seek a time-dependent quantum state that has the classical equivalence of a pendulum. This is the coherent state, which is the contribution of many researchers, most notably, Roy Glauber (1925–2018) [269] in 1963, and George Sudarshan (1931–2018) [270]. Glauber was awarded the Nobel prize in 2005.

We like to emphasize again that the modes of an electromagnetic cavity oscillation are homomorphic to the oscillation of classical pendulum. Hence, we first connect the oscillation of a quantum pendulum to a classical pendulum. Then we can connect the oscillation of

[1] In quantum theory, a "state" is synonymous with a state vector or a function.

a quantum electromagnetic mode to the classical electromagnetic mode and then to the quantum pendulum.

39.1.1 Quantum Harmonic Oscillator Revisited

To this end, we revisit the quantum harmonic oscillator or the quantum pendulum with more mathematical depth. Rewriting Schrödinger equation as the eigenequation for the photon number state for the quantum harmonic oscillator, we have

$$\hat{H}\psi(x) = \left[-\frac{\hbar^2}{2m}\frac{d^2}{dx^2} + \frac{1}{2}m\omega_0^2 x^2\right]\psi(x) = E\psi(x). \tag{39.1.1}$$

where $\psi(x)$ is the eigenfunction, and E is the eigenvalue. The above can be changed into a dimensionless form first by dividing $\hbar\omega_0$, and then let $\xi = \sqrt{\frac{m\omega_0}{\hbar}}x$ be a dimensionless variable. The above then becomes

$$\frac{1}{2}\left(-\frac{d^2}{d\xi^2} + \xi^2\right)\psi(\xi) = \frac{E}{\hbar\omega_0}\psi(\xi) \tag{39.1.2}$$

We can define $\hat{\pi} = -i\frac{d}{d\xi}$ and $\hat{\xi} = \hat{I}\xi$ to rewrite the Hamiltonian as

$$\hat{H} = \frac{1}{2}\hbar\omega_0(\hat{\pi}^2 + \hat{\xi}^2) \tag{39.1.3}$$

Furthermore, the Hamiltonian in (39.1.2) looks almost like $A^2 - B^2$, and hence motivates its factorization. To this end, we first show that

$$\frac{1}{\sqrt{2}}\left(-\frac{d}{d\xi} + \xi\right)\frac{1}{\sqrt{2}}\left(\frac{d}{d\xi} + \xi\right) = \frac{1}{2}\left(-\frac{d^2}{d\xi^2} + \xi^2\right) - \frac{1}{2}\left(\frac{d}{d\xi}\xi - \xi\frac{d}{d\xi}\right) \tag{39.1.4}$$

It can be shown easily that as operators (meaning that they will act on a function to their right), the last term on the right-hand side is an identity operator, namely that

$$\left(\frac{d}{d\xi}\xi - \xi\frac{d}{d\xi}\right) = \hat{I} \tag{39.1.5}$$

Therefore

$$\frac{1}{2}\left(-\frac{d^2}{d\xi^2} + \xi^2\right) = \frac{1}{\sqrt{2}}\left(-\frac{d}{d\xi} + \xi\right)\frac{1}{\sqrt{2}}\left(\frac{d}{d\xi} + \xi\right) + \frac{1}{2} \tag{39.1.6}$$

We define the operator

$$\hat{a}^\dagger = \frac{1}{\sqrt{2}}\left(-\frac{d}{d\xi} + \xi\right) \tag{39.1.7}$$

The above is the creation, or raising operator and the reason for its name is obviated later. Moreover, we define

$$\hat{a} = \frac{1}{\sqrt{2}}\left(\frac{d}{d\xi} + \xi\right) \tag{39.1.8}$$

which represents the annihilation or lowering operator. With the above definitions of the raising and lowering operators, it is easy to show that by straightforward substitution that

$$\left[\hat{a}, \hat{a}^\dagger\right] = \hat{a}\hat{a}^\dagger - \hat{a}^\dagger\hat{a} = \hat{I} \tag{39.1.9}$$

Therefore, Schrödinger equation (39.1.2) for quantum harmonic oscillator can be rewritten more concisely as

$$\frac{1}{2}\left(\hat{a}^\dagger\hat{a} + \hat{a}\hat{a}^\dagger\right)\psi = \left(\hat{a}^\dagger\hat{a} + \frac{1}{2}\right)\psi = \frac{E}{\hbar\omega_0}\psi \tag{39.1.10}$$

In mathematics, a function is analogous to a vector. So ψ is the implicit representation of a vector. The operator

$$\left(\hat{a}^\dagger\hat{a} + \frac{1}{2}\right)$$

is an implicit[2] representation of an operator, and in this case, a differential operator. So in the above, (39.1.10), is analogous to the matrix eigenvalue equation $\overline{\mathbf{A}} \cdot \mathbf{x} = \lambda\mathbf{x}$.

Consequently, the Hamiltonian operator can be expressed concisely as

$$\hat{H} = \hbar\omega_0\left(\hat{a}^\dagger\hat{a} + \frac{1}{2}\right) \tag{39.1.11}$$

Equation (39.1.10) above is in implicit math notation. In implicit Dirac notation, it is

$$\left(\hat{a}^\dagger\hat{a} + \frac{1}{2}\right)|\psi\rangle = \frac{E}{\hbar\omega_0}|\psi\rangle \tag{39.1.12}$$

In the above, $\psi(\xi)$ is a function which is a vector in a functional space. It is denoted as ψ in math notation and $|\psi\rangle$ in Dirac notation. This is also known as the "ket". The conjugate transpose of a vector in Dirac notation is called a "bra" which is denoted as $\langle\psi|$. Hence, the inner product between two vectors is denoted as $\langle\psi_1|\psi_2\rangle$ in Dirac notation.[3]

If we denote a photon number state by $\psi_n(x)$ in explicit notation, ψ_n in math notation or $|\psi_n\rangle$ in Dirac notation, then we have

$$\left(\hat{a}^\dagger\hat{a} + \frac{1}{2}\right)|\psi_n\rangle = \frac{E_n}{\hbar\omega_0}|\psi_n\rangle = \left(n + \frac{1}{2}\right)|\psi_n\rangle \tag{39.1.13}$$

where we have used the fact that $E_n = (n + 1/2)\hbar\omega_0$. Therefore, by comparing terms in the above, we have

$$\hat{a}^\dagger\hat{a}|\psi_n\rangle = n|\psi_n\rangle \tag{39.1.14}$$

and the operator $\hat{a}^\dagger\hat{a}$ is also known as the number operator because of the above. It is often denoted as

$$\hat{n} = \hat{a}^\dagger\hat{a} \tag{39.1.15}$$

[2]A notation like $\overline{\mathbf{A}} \cdot \mathbf{x}$, we will call implicit, while a notation $\sum_{i,j} A_{ij}x_j$, we will call explicit.

[3]There is a one-to-one correspondence of Dirac notation to matrix algebra notation. $\hat{A}|x\rangle \leftrightarrow \overline{\mathbf{A}} \cdot \mathbf{x}$, $\langle x| \leftrightarrow \mathbf{x}^\dagger$ $\langle x_1|x_2\rangle \leftrightarrow \mathbf{x}_1^\dagger \cdot \mathbf{x}_2$.

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

$$\hat{a}|\psi_n\rangle = \sqrt{n}|\psi_{n-1}\rangle \qquad \Leftrightarrow \hat{a}|n\rangle \ = \sqrt{n}|n-1\rangle \tag{39.1.16}$$

$$\hat{a}^\dagger|\psi_n\rangle = \sqrt{n+1}|\psi_{n+1}\rangle \Leftrightarrow \hat{a}^\dagger|n\rangle = \sqrt{n+1}|n+1\rangle \tag{39.1.17}$$

hence their names as lowering and raising operator.[4]

39.2 Some Words on Quantum Randomness and Quantum Observables

We saw previously that in classical mechanics, the conjugate variables p and x are deterministic variables. But in the quantum world, they become random variables with means and variance. It was quite easy to see that x is a random variable in the quantum world. But the momentum p is elevated to become a differential operator $\hat{p}$, and it is not clear that it is a random variable anymore.

Quantum theory is a lot richer in content than classical theory. Hence, in quantum theory, conjugate variables like p and x are observables endowed with the properties of mean and variance. For them to be endowed with these properties, they are elevated to become quantum operators, which are the representations of these observables. To be meaningful, a quantum state $|\psi\rangle$ has to be defined for a quantum system, and these operators representing observables act on the quantum state.

Henceforth, we have to extend the concept of the average of a random variable to the "average" of a quantum operator. Now that we know Dirac notation, we can write the expectation value of the operator $\hat{p}$ with respect to a quantum state ψ as

$$\bar{p} = \langle p \rangle = \langle \psi | \hat{p} | \psi \rangle \tag{39.2.1}$$

The above is the elevated way of taking the "average" of an operator which is related to the mean of the random variable p. In the above, the two-some, $\{\hat{p}, |\psi\rangle\}$ endows the quantum observable p with random properties. Here, p becomes a random variable whose mean is given by the above formula.

As mentioned before, Dirac notation is homomorphic to matrix algebra notation. The above is similar to $\boldsymbol{\psi}^\dagger \cdot \overline{\mathbf{P}} \cdot \boldsymbol{\psi} = \bar{p}$. This quantity $\bar{p}$ is always real if $\overline{\mathbf{P}}$ is a Hermitian matrix. Hence, in (39.2.1), the expectation value $\bar{p}$ is always real if $\hat{p}$ is Hermitian. In fact, it can be proved that $\hat{p}$ is Hermitian in the function space that it is defined.

Furthermore, the variance of the random variable p can be derived from the quantum operator $\hat{p}$ with respect to to a quantum state $|\psi\rangle$. It is defined as

$$\sigma_p^2 = \langle \psi | (\hat{p} - \bar{p})^2 | \psi \rangle = \langle \psi | \hat{p}^2 | \psi \rangle - \bar{p}^2 \tag{39.2.2}$$

where σ_p is the standard deviation of the random variable p and σ_p^2 is its variance [71, 72].

[4]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 notation. Namely, $|n\rangle \to |\psi_n\rangle \to \psi_n(\xi)$.

The above implies that the definition of the quantum operators and the quantum states is not unique. One can define a unitary matrix or operator $\overline{\mathbf{U}}$ such that $\overline{\mathbf{U}}^\dagger \cdot \overline{\mathbf{U}} = \overline{\mathbf{I}}$. Then the new quantum state is now given by the unitary transform $\boldsymbol{\psi}' = \overline{\mathbf{U}} \cdot \boldsymbol{\psi}$. With this, we can easily show that

$$\begin{aligned} \bar{p} &= \boldsymbol{\psi}^\dagger \cdot \overline{\mathbf{P}} \cdot \boldsymbol{\psi} = \boldsymbol{\psi}^\dagger \cdot \overline{\mathbf{U}}^\dagger \cdot \overline{\mathbf{U}} \cdot \overline{\mathbf{P}} \cdot \overline{\mathbf{U}}^\dagger \cdot \overline{\mathbf{U}} \cdot \boldsymbol{\psi} \\ &= \boldsymbol{\psi}'^\dagger \cdot \overline{\mathbf{P}}' \cdot \boldsymbol{\psi}' \end{aligned} \tag{39.2.3}$$

where $\overline{\mathbf{P}}' = \overline{\mathbf{U}} \cdot \overline{\mathbf{P}} \cdot \overline{\mathbf{U}}^\dagger$ via unitary transform. Now, $\overline{\mathbf{P}}'$ is the new quantum operator representing the observable p and $\boldsymbol{\psi}'$ is the new quantum state vector.

In the previous section, we have elevated the position variable or observable ξ to become an operator $\hat{\xi} = \xi \hat{I}$. This operator is clearly Hermitian, and hence, the expectation value of this position operator is always real. Here, $\hat{\xi}$ is diagonal in the coordinate space representation, but it need not be in other Hilbert space representations using unitary transformation shown above.

39.3 Derivation of 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. This is the role of the coherent state: It will show the correspondence principle, with a classical pendulum emerging from a quantum pendulum when the energy of the pendulum is large. Hence, it will be interesting to see how the coherent state is derived.

The derivation of the coherent state is more math than physics. Nevertheless, the derivation is interesting. We are going to present it according to the simplest way presented in the literature. There are deeper mathematical methods to derive this coherent state like Bogoliubov transform which is outside the scope of this course.

Now, endowed with the needed mathematical tools, we can derive the coherent state simply. To say succinctly, the coherent state is the eigenstate of the annihilation operator, namely that

$$\hat{a}|\alpha\rangle = \alpha|\alpha\rangle \tag{39.3.1}$$

Here, we use α as an eigenvalue as well as an index or identifier of the state $|\alpha\rangle$.[5] Since the number state $|n\rangle$ is complete, the coherent state $|\alpha\rangle$ can be expanded in terms of the number state $|n\rangle$. Or that

$$|\alpha\rangle = \sum_{n=0}^{\infty} C_n |n\rangle \tag{39.3.2}$$

When the annihilation operator is applied to the above, we have

$$\hat{a}|\alpha\rangle = \sum_{n=0}^{\infty} C_n \hat{a}|n\rangle = \sum_{n=1}^{\infty} C_n \hat{a}|n\rangle = \sum_{n=1}^{\infty} C_n \sqrt{n}|n-1\rangle = \sum_{n=0}^{\infty} C_{n+1}\sqrt{n+1}|n\rangle \tag{39.3.3}$$

[5] 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.

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 (39.3.1), then

$$\sum_{n=0}^{\infty} C_{n+1}\sqrt{n+1}|n\rangle = \alpha \sum_{n=0}^{\infty} C_n|n\rangle \tag{39.3.4}$$

By the orthonormality of the number states $|n\rangle$ and the completeness of the set,

$$C_{n+1} = \alpha C_n/\sqrt{n+1} \tag{39.3.5}$$

Or recursively

$$C_n = C_{n-1}\alpha/\sqrt{n} = C_{n-2}\alpha^2/\sqrt{n(n-1)} = \ldots = C_0\alpha^n/\sqrt{n!} \tag{39.3.6}$$

Consequently, the coherent state $|\alpha\rangle$ is

$$|\alpha\rangle = C_0 \sum_{n=0}^{\infty} \frac{\alpha^n}{\sqrt{n!}}|n\rangle \tag{39.3.7}$$

But due to the probabilistic interpretation of quantum mechanics, the state vector $|\alpha\rangle$ is normalized to one, or that[6]

$$\langle\alpha|\alpha\rangle = 1 \tag{39.3.8}$$

Then

$$\begin{aligned}\langle\alpha|\alpha\rangle &= C_0^* C_0 \sum_{n,n'}^{\infty} \frac{\alpha^n}{\sqrt{n!}} \frac{\alpha^{n'}}{\sqrt{n'!}} \langle n'|n\rangle \\ &= |C_0|^2 \sum_{n=0}^{\infty} \frac{|\alpha|^{2n}}{n!} = |C_0|^2 e^{|\alpha|^2} = 1\end{aligned} \tag{39.3.9}$$

Therefore, $C_0 = e^{-|\alpha|^2/2}$ for normalization, or that

$$|\alpha\rangle = e^{-|\alpha|^2/2} \sum_{n=0}^{\infty} \frac{\alpha^n}{\sqrt{n!}}|n\rangle \tag{39.3.10}$$

In the above, to reduce the double summations into a single summation, we have made use of $\langle n'|n\rangle = \delta_{n'n}$, 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.

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 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.

[6]The expression can be written more explicitly as $\langle\alpha|\alpha\rangle = \langle f_\alpha|f_\alpha\rangle = \int_{-\infty}^{\infty} d\xi f_\alpha^*(\xi) f_\alpha(\xi) = 1$.

39.3.1 Time Evolution of a Quantum State

The Schrödinger equation can be written concisely as

$$\hat{H}|\psi\rangle = i\hbar\partial_t|\psi\rangle \tag{39.3.11}$$

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

$$|\psi(t)\rangle = e^{-i\hat{H}t/\hbar}|\psi(0)\rangle \tag{39.3.12}$$

Meaning of the Function of an Operator

At this juncture, it is prudent to digress to discuss the meaning of a function of an operator, which occurs in (39.3.12): The exponential function is a function of the operator $\hat{H}$. This is best explained by expanding the pertinent function into a Taylor series, namely,

$$f(\overline{\mathbf{A}}) = f(0)\overline{\mathbf{I}} + f'(0)\overline{\mathbf{A}} + \frac{1}{2!}f''(0)\overline{\mathbf{A}}^2 + \ldots + \frac{1}{n!}f^{(n)}(0)\overline{\mathbf{A}}^n + \ldots \tag{39.3.13}$$

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 (39.3.13) to an eigenvector $\mathbf{v}$ of $\overline{\mathbf{A}}$, we have

$$\begin{aligned} f(\overline{\mathbf{A}})\mathbf{v} &= f(0)\mathbf{v} + f'(0)\overline{\mathbf{A}}\mathbf{v} + \frac{1}{2!}f''(0)\overline{\mathbf{A}}^2\mathbf{v} + \ldots + \frac{1}{n!}f^{(n)}(0)\overline{\mathbf{A}}^n\mathbf{v} + \ldots \\ &= f(0)\mathbf{v} + f'(0)\lambda\mathbf{v} + \frac{1}{2!}f''(0)\overline{\lambda}^2\mathbf{v} + \ldots + \frac{1}{n!}f^{(n)}(0)\lambda^n\mathbf{v} + \ldots = f(\lambda)\mathbf{v} \end{aligned} \tag{39.3.14}$$

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

$$e^{\overline{\mathbf{A}}}\mathbf{v} = e^{\lambda}\mathbf{v} \tag{39.3.15}$$

Applying this to the photon number state with $\hat{H}$ being that of the quantum pendulum and that $|n\rangle$ is the eigenvector of $\hat{H}$, then

$$e^{-i\hat{H}t/\hbar}|n\rangle = e^{-i\omega_n t}|n\rangle \tag{39.3.16}$$

where $\omega_n = \left(n + \frac{1}{2}\right)\omega_0$. In particular, $|n\rangle$ is an eigenstate of the Hamiltonian $\hat{H}$ for the quantum pendulum, or that from (39.1.11) and (39.1.14)

$$\hat{H}|n\rangle = \hbar\omega_n|n\rangle = \hbar\omega_0\left(n + \frac{1}{2}\right)|n\rangle \tag{39.3.17}$$

In other words, $|n\rangle$, a shorthand notation for $|\psi_n\rangle$ in (39.1.13), is an eigenvector of $\hat{H}$.

Time Evolution of the Coherent State

Using the above time-evolution operator, then the time dependent coherent state evolves in time as[7]

$$|\alpha,t\rangle = e^{-i\hat{H}t/\hbar}|\alpha\rangle = e^{-|\alpha|^2/2}\sum_{n=0}^{\infty}\frac{\alpha^n e^{-i\omega_n t}}{\sqrt{n!}}|n\rangle \tag{39.3.18}$$

By letting $\omega_n = \omega_0\left(n+\frac{1}{2}\right)$, the above can be written as

$$|\alpha,t\rangle = e^{-i\omega_0 t/2}e^{-|\alpha|^2/2}\sum_{n=0}^{\infty}\frac{\left(\alpha e^{-i\omega_0 t}\right)^n}{\sqrt{n!}}|n\rangle \tag{39.3.19}$$

Now we see that the last factor in (39.3.19) is similar to the expression for a coherent state in (39.3.10). Therefore, we can express the above more succinctly by replacing α in (39.3.10) with $\tilde{\alpha} = \alpha e^{-i\omega_0 t}$ as

$$|\alpha,t\rangle = e^{-i\omega_0 t/2}|\alpha e^{-i\omega_0 t}\rangle = e^{-i\omega_0 t/2}|\tilde{\alpha}\rangle \tag{39.3.20}$$

Consequently,

$$\hat{a}|\alpha,t\rangle = \hat{a}e^{-i\omega_0 t/2}|\alpha e^{-i\omega_0 t}\rangle = e^{-i\omega_0 t/2}\left(\alpha e^{-i\omega_0 t}\right)|\alpha e^{-i\omega_0 t}\rangle = \tilde{\alpha}|\alpha,t\rangle \tag{39.3.21}$$

Therefore, $|\alpha,t\rangle$ is the eigenfunction of the $\hat{a}$ operator with eigenvalue $\tilde{\alpha}$. But now, the eigenvalue of the annihilation operator $\hat{a}$ is a complex number which is a function of time t. It is to be noted that in the coherent state in (39.3.19), 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 $\tilde{\alpha}$!

39.4 More on the Creation and Annihilation Operator

As seen in the photon-number states, the oscillation of the pendulum does not emerge in the quantum solutions to Schrödinger equation. Hence it is prudent to see if this physical phenomenon emerge 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. Since

$$\hat{a}^\dagger = \frac{1}{\sqrt{2}}\left(-\frac{d}{d\xi}+\xi\right) \tag{39.4.1}$$

$$\hat{a} = \frac{1}{\sqrt{2}}\left(\frac{d}{d\xi}+\xi\right) \tag{39.4.2}$$

We can relate $\hat{a}^\dagger$ and $\hat{a}$, which are non-hermitian, to the momentum operator $\hat{\pi}$ and position operator $\hat{\xi}$ previously defined which are hermitian. Then

$$\hat{a}^\dagger = \frac{1}{\sqrt{2}}\left(-i\hat{\pi}+\hat{\xi}\right) \tag{39.4.3}$$

$$\hat{a} = \frac{1}{\sqrt{2}}\left(i\hat{\pi}+\hat{\xi}\right) \tag{39.4.4}$$

[7]Note that $|\alpha,t\rangle$ is a shorthand for $f_\alpha(\xi,t)$.

From the above, by subtracting and adding the two equations, we arrive at

$$\hat{\xi} = \frac{1}{\sqrt{2}}\left(\hat{a}^\dagger + \hat{a}\right) = \xi\hat{I} \tag{39.4.5}$$

$$\hat{\pi} = \frac{i}{\sqrt{2}}\left(\hat{a}^\dagger - \hat{a}\right) = -i\frac{d}{d\xi} \tag{39.4.6}$$

Notice that both $\hat{\xi}$ and $\hat{\pi}$ are Hermitian operators in the above, with real expectation values in accordance to (39.2.1). With this, the average or expectation value of the position of the pendulum in normalized coordinate, ξ, can be found by taking expectation with respect to the coherent state, or

$$\langle\alpha|\hat{\xi}|\alpha\rangle = \frac{1}{\sqrt{2}}\langle\alpha|\hat{a}^\dagger + \hat{a}|\alpha\rangle \tag{39.4.7}$$

Since by taking the complex conjugate transpose of (39.3.1)[8]

$$\langle\alpha|\hat{a}^\dagger = \langle\alpha|\alpha^* \tag{39.4.8}$$

and (39.4.7) becomes

$$\bar{\xi} = \langle\xi\rangle = \langle\alpha|\hat{\xi}|\alpha\rangle = \frac{1}{\sqrt{2}}\left(\alpha^* + \alpha\right)\langle\alpha|\alpha\rangle = \sqrt{2}\Re e\left(\alpha\right) \neq 0 \tag{39.4.9}$$

Repeating the exercise for time-dependent case, when we let $\alpha \to \tilde{\alpha}(t) = \alpha e^{-i\omega_0 t}$, then, letting $\alpha = |\alpha|e^{-i\psi}$ yields

$$\bar{\xi}(t) = \langle\xi(t)\rangle = \langle\tilde{\alpha}(t)|\hat{\xi}|\tilde{\alpha}(t)\rangle = \frac{1}{\sqrt{2}}\left(\tilde{\alpha}^*(t) + \tilde{\alpha}(t)\right)\langle\tilde{\alpha}(t)|\tilde{\alpha}(t)\rangle = \sqrt{2}\Re e\left(\tilde{\alpha}(t)\right) \neq 0 \tag{39.4.10}$$

where for the last equality, we have made used of that $\langle\tilde{\alpha}(t)|\tilde{\alpha}(t)\rangle = 1$. Then

$$\bar{\xi}(t) = \langle\xi(t)\rangle = \sqrt{2}|\alpha|\cos(\omega_0 t + \psi) \tag{39.4.11}$$

In the above, we use ξ to denote the random variable. So $\langle\xi(t)\rangle$ refers to the average of the random variable ξ, or $\bar{\xi}(t)$ that is a function of time.

By the same token,

$$\bar{\pi} = \langle\pi\rangle = \langle\alpha|\hat{\pi}|\alpha\rangle = \frac{i}{\sqrt{2}}\left(\alpha^* - \alpha\right)\langle\alpha|\alpha\rangle = \sqrt{2}\Im m\left(\alpha\right) \neq 0 \tag{39.4.12}$$

For the time-dependent case, we let $\alpha \to \tilde{\alpha}(t) = \alpha e^{-i\omega_0 t}$,

$$\bar{\pi}(t) = \langle\pi(t)\rangle = -\sqrt{2}|\alpha|\sin(\omega_0 t + \psi) \tag{39.4.13}$$

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 have

[8]Dirac notation is homomorphic with matrix algebra notation. $(\overline{\mathbf{a}}\cdot\mathbf{x})^\dagger = \mathbf{x}^\dagger\cdot(\overline{\mathbf{a}})^\dagger$.

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.

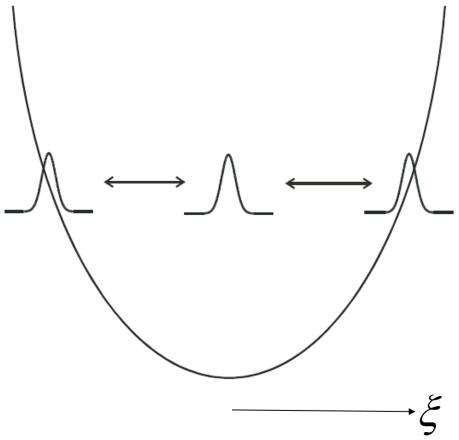


Fig. 3.4. A coherent-state wave function moves through the harmonic oscillator potential, between the classical turning points, without dispersion.

Figure 39.1: The time evolution of the coherent state. It follows the motion of a classical pendulum or harmonic oscillator (courtesy of Gerry and Knight [271]).

.

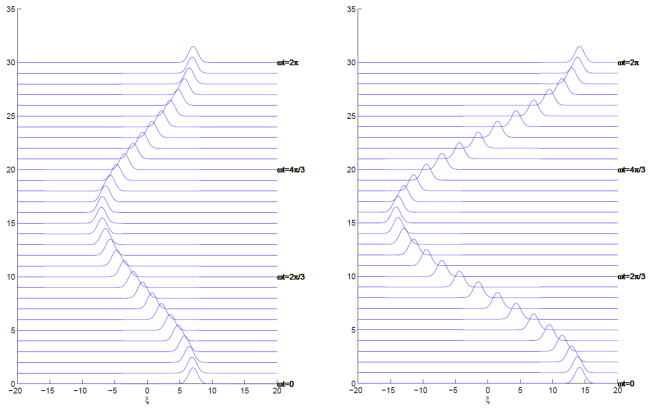


Figure 39.2: The time evolution of the coherent state for different α's. The left figure is for $\alpha = 5$ while the right figure is for $\alpha = 10$. Recall that $N = |\alpha|^2$.

.

39.4.1 Connecting Quantum Pendulum to Electromagnetic Oscillator[9]

We see that the electromagnetic oscillator in a cavity is similar or homomorphic to a pendulum. The classical Hamiltonian is

$$H = T + V = \frac{p^2}{2m} + \frac{1}{2}m\omega_0^2 x^2 = \frac{1}{2}\left[P^2(t) + Q^2(t)\right] = E \tag{39.4.14}$$

where E is the total energy of the system. In the above, ω_0 is the resonant frequency of the classical pendulum. To make the cavity mode homomorphic to a pendulum, we have to replace ω_0 with ω_l, the resonant frequency of the cavity mode. In the above, P is a normalized momentum and Q is a normalized coordinate, and their squares have the unit of energy. 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 Schrödinger equation becomes

$$\hat{H}|\psi, t\rangle = \hbar\omega_l\left(\hat{a}^\dagger\hat{a} + \frac{1}{2}\right)|\psi, t\rangle = i\hbar\partial_t|\psi, t\rangle \tag{39.4.15}$$

Our next task is to connect the electromagnetic oscillator to this pendulum. In general, the total energy or the Hamiltonian of an electromagnetic system is

$$H = \frac{1}{2}\int\limits_V d\mathbf{r}\left[\varepsilon\mathbf{E}^2(\mathbf{r}, t) + \frac{1}{\mu}\mathbf{B}^2(\mathbf{r}, t)\right]. \tag{39.4.16}$$

It is customary to write this Hamiltonian in terms of scalar and vector potentials. For simplicity, we use a 1D cavity, and let $\mathbf{A} = \hat{x}A_x$, $\nabla \cdot \mathbf{A} = 0$ so that $\partial_x A_x = 0$, and letting $\Phi = 0$. Then $\mathbf{B} = \nabla \times \mathbf{A}$ and $\mathbf{E} = -\dot{\mathbf{A}}$, and the classical Hamiltonian from (39.4.16) for a Maxwellian system becomes

$$H = \frac{1}{2}\int\limits_V d\mathbf{r}\left[\varepsilon\dot{\mathbf{A}}^2(\mathbf{r}, t) + \frac{1}{\mu}\left(\nabla \times \mathbf{A}(\mathbf{r}, t)\right)^2\right]. \tag{39.4.17}$$

For the 1D case, the above implies that $B_y = \partial_z A_x$, and $E_x = -\partial_t A_x = -\dot{A}_x$. Hence, we let

$$A_x = A_0(t)\sin(k_l z) \tag{39.4.18}$$

$$E_x = -\dot{A}_0(t)\sin(k_l z) = E_0(t)\sin(k_l z) \tag{39.4.19}$$

$$B_y = k_l A_0(t)\cos(k_l z). \tag{39.4.20}$$

where $E_0(t) = -\dot{A}_0(t)$. After integrating over the volume such that $\int_V d\mathbf{r} = \mathcal{A}\int_0^L dz$, the Hamiltonian (39.4.17) then becomes

$$H = \frac{V_0\varepsilon}{4}\left(\dot{A}_0(t)\right)^2 + \frac{V_0}{4\mu}k_l^2 A_0^2(t). \tag{39.4.21}$$

[9]May be skipped on first reading.

where $V_0 = \mathcal{AL}$, is the mode volume. The form of (39.4.21) now resembles the pendulum Hamiltonian. We can think of $A_0(t)$ as being related to the displacement of the pendulum. Hence, the second term resembles the potential energy. The first term has the time derivative of $A_0(t)$, and hence, can be connected to the kinetic energy of the system. Therefore, we can rewrite the Hamiltonian as

$$H = \frac{1}{2}\left[P^2(t) + Q^2(t)\right] \tag{39.4.22}$$

where

$$P(t) = \sqrt{\frac{V_0\varepsilon}{2}}\dot{A}_0(t) = -\sqrt{\frac{V_0\varepsilon}{2}}E_0(t), \qquad Q(t) = \sqrt{\frac{V_0}{2\mu}}k_l A_0(t) \tag{39.4.23}$$

By elevating P and Q to be quantum operators,

$$P(t) \to \hat{P} = \sqrt{\hbar\omega_l}\hat{\pi}(t), \qquad Q(t) \to \hat{Q} = \sqrt{\hbar\omega_l}\hat{\xi}(t) \tag{39.4.24}$$

so that the quantum Hamiltonian now is

$$\hat{H} = \frac{1}{2}\left[\hat{P}^2 + \hat{Q}^2\right] = \frac{1}{2}\hbar\omega_l\left(\hat{\pi}^2 + \hat{\xi}^2\right) \tag{39.4.25}$$

similar to (39.1.3) as before except now that the resonant frequency of this mode is ω_l instead of ω_0 because these are the cavity modes, each of which is homomorphic to a quantum pendulum of frequency ω_l. An equation of motion for the state of the quantum system can be associated with the quantum Hamiltonian just as in the quantum pendulum case.

We have shown previously that

$$\hat{a}^\dagger + \hat{a} = \sqrt{2}\hat{\xi} \tag{39.4.26}$$

$$\hat{a}^\dagger - \hat{a} = -\sqrt{2}i\hat{\pi} \tag{39.4.27}$$

Then we can let

$$\hat{P} = -\sqrt{\frac{V_0\varepsilon}{2}}\hat{E}_0 = \sqrt{\hbar\omega_l}\hat{\pi} \tag{39.4.28}$$

Finally, we arrive at

$$\hat{E}_0 = -\sqrt{\frac{2\hbar\omega_l}{\varepsilon V_0}}\hat{\pi} = \frac{1}{i}\sqrt{\frac{\hbar\omega_l}{\varepsilon V_0}}\left(\hat{a}^\dagger - \hat{a}\right) \tag{39.4.29}$$

Now that E_0 has been elevated to be a quantum operator $\hat{E}_0$, from (39.4.19), we can put in the space dependence in accordance to (39.4.19) to get

$$\hat{E}_x(z) = \hat{E}_0 \sin(k_l z) \tag{39.4.30}$$

Consequently, the fully quantized field is a field operator given by

$$\hat{E}_x(z) = \frac{1}{i}\sqrt{\frac{\hbar\omega_l}{\varepsilon V_0}}\left(\hat{a}^\dagger - \hat{a}\right)\sin(k_l z) \tag{39.4.31}$$

Notice that in the above, $\hat{E}_0$, and $\hat{E}_x(z)$ are all Hermitian operators and they correspond to quantum observables that have randomness associated with them but with real mean values. Also, the operators are independent of time because they are in the Schrödinger picture.

Heisenberg Picture and 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 an operator, and the mean value of the observable, which is measurable, is given by

$$\bar{O}(t) = \langle\psi(t)|\hat{O}_s|\psi(t)\rangle. \tag{39.4.32}$$

Alternatively, one can rewrite the above using (39.3.12) as

$$\bar{O}(t) = \langle\psi(0)|e^{i\hat{H}t/\hbar}\hat{O}_s e^{-i\hat{H}t/\hbar}|\psi(0)\rangle. \tag{39.4.33}$$

In other words, the above can be rewritten as

$$\bar{O}(t) = \langle\psi(0)|\hat{O}_h(t)|\psi(0)\rangle. \tag{39.4.34}$$

where

$$\hat{O}_h(t) = e^{i\hat{H}t/\hbar}\hat{O}_s e^{-i\hat{H}t/\hbar}. \tag{39.4.35}$$

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. The second way, called the Heisenberg picture, lets the operator be time dependent, but the state vector 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 $e^{-i\hat{H}t/\hbar}$.

In the Schrödinger picture, to get time dependence fields, one has to take the expectation value of the operators with respect to time-varying quantum state vector like the time-varying coherent state.

To let $\hat{E}_x$ have any meaning, it should act on a quantum state. For example,

$$|\psi_E\rangle = \hat{E}_x|\psi\rangle \tag{39.4.36}$$

Notice that thus far, all the operators derived are independent of time. To derive time dependence of these operators, one needs to find their expectation value with respect to time-dependent state vectors.[10]

To illustrate this, we can take expectation value of the quantum operator $\hat{E}_x(z)$ with

[10]This is known as the Schrödinger picture.

respect to a time dependent state vector, like the time-dependent coherent state, Thus

$$\langle E_x(z,t)\rangle = \langle \alpha,t|\hat{E}_x(z)|\alpha,t\rangle = \frac{1}{i}\sqrt{\frac{\hbar\omega_l}{\varepsilon V_0}}\langle \alpha,t|\hat{a}^\dagger - \hat{a}|\alpha,t\rangle$$
$$= \frac{1}{i}\sqrt{\frac{\hbar\omega_l}{\varepsilon V_0}}\left(\tilde{\alpha}^*(t) - \tilde{\alpha}(t)\right)\langle \alpha,t|\alpha,t\rangle = -2\sqrt{\frac{\hbar\omega_l}{\varepsilon V_0}}\Im m(\tilde{\alpha}) \tag{39.4.37}$$

From (39.3.19), we can show that $\langle \alpha,t|\alpha,t\rangle = 1$. Using the time-dependent $\tilde{\alpha}(t) = \alpha e^{-i\omega_l t} = |\alpha|e^{-i(\omega_l t+\psi)}$ in the above, similar to (39.4.13), we have

$$\langle E_x(z,t)\rangle = 2\sqrt{\frac{\hbar\omega_l}{\varepsilon V_0}}|\alpha|\sin(\omega_l t+\psi) \tag{39.4.38}$$

where $\tilde{\alpha}(t) = \alpha e^{-i\omega_l t}$. The expectation value of the operator with respect to a time-varying quantum state, in fact, gives rise to a time-varying quantity. The above, which is the average of a random field, resembles a classical field. But since it is rooted in a random variable, it has a standard deviation in addition to having a real-value mean.

We can also show that

$$\hat{B}_y(z) = k_l\hat{A}_0\cos(k_l z) = \sqrt{\frac{2\mu\hbar\omega_l}{V_0}}\hat{\xi} = \sqrt{\frac{\mu\hbar\omega_l}{V_0}}(\hat{a}^\dagger + \hat{a}) \tag{39.4.39}$$

Again, these are time-independent operators in the Schrödinger picture. To get time-dependent quantities, we have to take the expectation value of the above operator with respect to to a time-varying quantum state.

39.5 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 peek into this theory. It takes much longer to learn this subject well: after all, it is the by product of 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 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 book [39]. 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 nano-meter lengthscales to galactic lengthscales, 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.

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] H. Hertz, "On the finite velocity of propagation of electromagnetic actions," *Electric Waves*, vol. 110, 1888.

[19] 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.

[20] 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.

[21] A. Pais, "Einstein and the quantum theory," *Reviews of Modern Physics*, vol. 51, no. 4, p. 863, 1979.

[22] M. Planck, "On the law of distribution of energy in the normal spectrum," *Annalen der physik*, vol. 4, no. 553, p. 1, 1901.

[23] 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.

[24] 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.

[25] 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.

[26] D. J. Griffiths and D. F. Schroeter, *Introduction to quantum mechanics*. Cambridge University Press, 2018.

[27] 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.

[28] C. Pickover, *Archimedes to Hawking: Laws of science and the great minds behind them.* Oxford University Press, 2008.

[29] R. Resnick, J. Walker, and D. Halliday, *Fundamentals of physics.* John Wiley, 1988.

[30] J. L. De Lagrange, "Recherches d'arithmétique," *Nouveaux Mémoires de l'Académie de Berlin*, 1773.

[31] 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.

[32] J. A. Kong, *Electromagnetic Wave Theory.* EMW Publishing, 2008, also 1985.

[33] H. M. Schey, *Div, grad, curl, and all that: an informal text on vector calculus.* WW Norton New York, 2005.

[34] 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.

[35] W. C. Chew, *Waves and fields in inhomogeneous media.* IEEE Press, 1995, also 1990.

[36] V. J. Katz, "The history of Stokes' theorem," *Mathematics Magazine*, vol. 52, no. 3, pp. 146–156, 1979.

[37] J. C. Maxwell, *A Treatise on Electricity and magnetism.* Dover New York, 1954, first published in 1873, vol. 1 and 2.

[38] A. D. Yaghjian, "Reflections on Maxwell's treatise," *Progress In Electromagnetics Research*, vol. 149, pp. 217–249, 2014.

[39] P. J. Nahin, *Oliver Heaviside: sage in solitude.* IEEE Press New York, 1987.

[40] J. C. Maxwell, "Poems of James Clerk Maxwell," https://mypoeticside.com/poets/james-clerk-maxwell-poems.

[41] W. K. Panofsky and M. Phillips, *Classical electricity and magnetism.* Courier Corporation, 2005.

[42] T. Lancaster and S. J. Blundell, *Quantum field theory for the gifted amateur.* OUP Oxford, 2014.

[43] J. D. Jackson, *Classical Electrodynamics.* John Wiley & Sons, 1962.

[44] W. C. Chew, "Fields and waves: Lecture notes for ECE 350 at UIUC," *https://engineering.purdue.edu/wcchew/ece350.html*, 1990.

[45] 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.

[46] J. M. Crowley, *Fundamentals of applied electrostatics*. Krieger Publishing Company, 1986.

[47] C. Balanis, *Advanced Engineering Electromagnetics*. Hoboken, NJ, USA: Wiley, 2012.

[48] J. D. Jackson, *Classical electrodynamics*. John Wiley & Sons, 1999.

[49] R. Courant and D. Hilbert, *Methods of Mathematical Physics, Volumes 1 and 2*. Interscience Publ., 1962.

[50] R. F. Harrington, *Time-harmonic electromagnetic fields*. McGraw-Hill, 1961.

[51] 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.

[52] E. Kudeki and D. C. Munson, *Analog Signals and Systems*. Upper Saddle River, NJ, USA: Pearson Prentice Hall, 2009.

[53] A. V. Oppenheim and R. W. Schafer, *Discrete-time signal processing*. Pearson Education, 2014.

[54] E. C. Jordan and K. G. Balmain, *Electromagnetic waves and radiating systems*. Prentice-Hall, 1968.

[55] G. Agarwal, D. Pattanayak, and E. Wolf, "Electromagnetic fields in spatially dispersive media," *Physical Review B*, vol. 10, no. 4, p. 1447, 1974.

[56] S. L. Chuang, *Physics of photonic devices*. John Wiley & Sons, 2012, vol. 80.

[57] B. E. Saleh and M. C. Teich, *Fundamentals of photonics*. John Wiley & Sons, 2019.

[58] M. Born and E. Wolf, *Principles of optics: electromagnetic theory of propagation, interference and diffraction of light*. Elsevier, 2013, also 1959 to 1986.

[59] R. W. Boyd, *Nonlinear optics*. Elsevier, 2003.

[60] Y.-R. Shen, *The principles of nonlinear optics*. New York, Wiley-Interscience, 1984.

[61] N. Bloembergen, *Nonlinear optics*. World Scientific, 1996.

[62] P. C. Krause, O. Wasynczuk, and S. D. Sudhoff, *Analysis of electric machinery*. McGraw-Hill New York, 1986.

[63] A. E. Fitzgerald, C. Kingsley, S. D. Umans, and B. James, *Electric machinery*. McGraw-Hill New York, 2003, vol. 5.

[64] M. A. Brown and R. C. Semelka, *MRI.: Basic Principles and Applications*. John Wiley & Sons, 2011.

[65] C. A. Balanis, *Advanced engineering electromagnetics*. John Wiley & Sons, 1999, also 1989.

[66] Wikipedia, "Lorentz force," https://en.wikipedia.org/wiki/Lorentz_force/, accessed: 2019-09-06.

[67] R. O. Dendy, *Plasma physics: an introductory course*. Cambridge University Press, 1995.

[68] Wikipedia, "Kennelly-Heaviside Layer," https://en.wikipedia.org/wiki/Kennelly-Heaviside_layer.

[69] ——, "Spectral line shape," https://en.wikipedia.org/wiki/Spectral_line_shape.

[70] 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.

[71] D. A. Miller, *Quantum Mechanics for Scientists and Engineers*. Cambridge, UK: Cambridge University Press, 2008.

[72] W. C. Chew, "Quantum mechanics made simple: Lecture notes for ECE 487 at UIUC," *http://wcchew.ece.illinois.edu/chew/course/QMAll20161206.pdf*, 2016.

[73] B. G. Streetman and S. Banerjee, *Solid state electronic devices*. Prentice hall Englewood Cliffs, NJ, 1995.

[74] 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.

[75] K. G. Budden, *Radio waves in the ionosphere*. Cambridge University Press, 2009.

[76] R. Fitzpatrick, *Plasma physics: an introduction*. CRC Press, 2014.

[77] G. Strang, *Introduction to linear algebra*. Wellesley-Cambridge Press Wellesley, MA, 1993, vol. 3.

[78] 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.

[79] J. Kraus, *Electromagnetics*. McGraw-Hill, 1984, also 1953, 1973, 1981.

[80] Wikipedia, "Circular polarization," https://en.wikipedia.org/wiki/Circular_polarization.

[81] Q. Zhan, "Cylindrical vector beams: from mathematical concepts to applications," *Advances in Optics and Photonics*, vol. 1, no. 1, pp. 1–57, 2009.

[82] H. Haus, *Electromagnetic Noise and Quantum Optical Measurements*, ser. Advanced Texts in Physics. Springer Berlin Heidelberg, 2000.

[83] 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.

[84] L. Brillouin, *Wave propagation and group velocity*. Academic Press, 1960.

[85] R. Plonsey and R. E. Collin, *Principles and applications of electromagnetic fields*. McGraw-Hill, 1961.

[86] M. N. Sadiku, *Elements of electromagnetics*. Oxford University Press, 2014.

[87] A. Wadhwa, A. L. Dal, and N. Malhotra, "Transmission media," https://www.slideshare.net/abhishekwadhwa786/transmission-media-9416228.

[88] P. H. Smith, "Transmission line calculator," *Electronics*, vol. 12, no. 1, pp. 29–31, 1939.

[89] F. B. Hildebrand, *Advanced calculus for applications*. Prentice-Hall, 1962.

[90] J. Schutt-Aine, "Experiment02-coaxial transmission line measurement using slotted line," http://emlab.uiuc.edu/ece451/ECE451Lab02.pdf.

[91] 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.

[92] R. E. Collin, *Field theory of guided waves*. McGraw-Hill, 1960.

[93] 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.

[94] Wikipedia, "Snell's law," https://en.wikipedia.org/wiki/Snell's_law.

[95] G. Tyras, *Radiation and propagation of electromagnetic waves*. Academic Press, 1969.

[96] L. Brekhovskikh, *Waves in layered media*. Academic Press, 1980.

[97] Scholarpedia, "Goos-hanchen effect," http://www.scholarpedia.org/article/Goos-Hanchen_effect.

[98] 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.

[99] E. Glytsis, "Slab waveguide fundamentals," http://users.ntua.gr/eglytsis/IO/Slab_Waveguides_p.pdf, 2018.

[100] Wikipedia, "Optical fiber," https://en.wikipedia.org/wiki/Optical_fiber.

[101] Atlantic Cable, "1869 indo-european cable," https://atlantic-cable.com/Cables/1869IndoEur/index.htm.

[102] Wikipedia, "Submarine communications cable," https://en.wikipedia.org/wiki/Submarine_communications_cable.

[103] 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.

[104] Wikipedia, "Brewster's angle," https://en.wikipedia.org/wiki/Brewster's_angle.

[105] H. Raether, "Surface plasmons on smooth surfaces," in *Surface plasmons on smooth and rough surfaces and on gratings*. Springer, 1988, pp. 4–39.

[106] 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.

[107] Wikipedia, "Homomorphic Encryption," https://en.wikipedia.org/wiki/Homomorphic_encryption.

[108] ——, "Surface plasmon," https://en.wikipedia.org/wiki/Surface_plasmon.

[109] A. Sommerfeld, *Über die Ausbreitung der Wellen in der drahtlosen Telegraphie*. Verlag der Königlich Bayerischen Akademie der Wissenschaften, 1909.

[110] W. Chew, *Waves and Fields in Inhomogeneous Media, 378±381*. Van Nostrand, 1990.

[111] Wikimedia, "Gaussian wave packet," https://commons.wikimedia.org/wiki/File:Gaussian_wave_packet.svg.

[112] Wikipedia, "Charles K. Kao," https://en.wikipedia.org/wiki/Charles_K._Kao.

[113] H. B. Callen and T. A. Welton, "Irreversibility and generalized noise," *Physical Review*, vol. 83, no. 1, p. 34, 1951.

[114] R. Kubo, "The fluctuation-dissipation theorem," *Reports on progress in physics*, vol. 29, no. 1, p. 255, 1966.

[115] 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.

[116] W. C. Chew, *Waves and Fields in Inhomogeneous Media*. IEEE Press, 1996.

[117] M. Abramowitz and I. A. Stegun, *Handbook of mathematical functions: with formulas, graphs, and mathematical tables*. Courier Corporation, 1965, vol. 55.

[118] ——, "Handbook of mathematical functions: with formulas, graphs, and mathematical tables," http://people.math.sfu.ca/~cbm/aands/index.htm.

[119] 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.

[120] Wikipedia, "Very Large Array," https://en.wikipedia.org/wiki/Very_Large_Array.

[121] C. A. Balanis and E. Holzman, "Circular waveguides," *Encyclopedia of RF and Microwave Engineering*, 2005.

[122] M. Al-Hakkak and Y. Lo, "Circular waveguides with anisotropic walls," *Electronics Letters*, vol. 6, no. 24, pp. 786–789, 1970.

[123] Wikipedia, "Horn Antenna," https://en.wikipedia.org/wiki/Horn_antenna.

[124] 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.

[125] R. Garg and I. Bahl, "Microstrip discontinuities," *International Journal of Electronics Theoretical and Experimental*, vol. 45, no. 1, pp. 81–87, 1978.

[126] P. Smith and E. Turner, "A bistable fabry-perot resonator," *Applied Physics Letters*, vol. 30, no. 6, pp. 280–281, 1977.

[127] A. Yariv, *Optical electronics*. Saunders College Publ., 1991.

[128] Wikipedia, "Klystron," https://en.wikipedia.org/wiki/Klystron.

[129] ——, "Magnetron," https://en.wikipedia.org/wiki/Cavity_magnetron.

[130] ——, "Absorption Wavemeter," https://en.wikipedia.org/wiki/Absorption_wavemeter.

[131] 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.

[132] A. R. Choudhuri, *The physics of fluids and plasmas: an introduction for astrophysicists*. Cambridge University Press, 1998.

[133] I. S. Gradshteyn and I. M. Ryzhik, *Table of integrals, series, and products*. Academic press, 2014.

[134] L. Nagel and D. Pederson, "Simulation program with integrated circuit emphasis," in *Midwest Symposium on Circuit Theory*, 1973.

[135] 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.

[136] Wikipedia, "Guglielmo Marconi," https://en.wikipedia.org/wiki/Guglielmo_Marconi.

[137] S. A. Schelkunoff and H. T. Friis, *Antennas: theory and practice*. Wiley New York, 1952, vol. 639.

[138] H. G. Schantz, "A brief history of UWB antennas," *IEEE Aerospace and Electronic Systems Magazine*, vol. 19, no. 4, pp. 22–26, 2004.

[139] E. Kudeki, "Fields and Waves," http://remote2.ece.illinois.edu/~erhan/FieldsWaves/ECE350lectures.html.

[140] Wikipedia, "Antenna Aperture," https://en.wikipedia.org/wiki/Antenna_aperture.

[141] L. Josefsson and P. Persson, *Conformal array antenna theory and design*. John Wiley & Sons, 2006, vol. 29.

[142] R. J. Mailloux, *Phased array antenna handbook*. Artech House, 2017.

[143] J. G. Proakis, *Digital signal processing: principles algorithms and applications*. Pearson Education India, 2001.

[144] 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.

[145] 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.

[146] Wikipedia, "Dipole Antenna," https://en.wikipedia.org/wiki/Dipole_antenna.

[147] ——, "Twin-Lead," https://en.wikipedia.org/wiki/Twin-lead.

[148] 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.

[149] Wikipedia, "Yagi-Uda Antenna," https://en.wikipedia.org/wiki/Yagi-Uda_antenna.

[150] 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.

[151] Antenna-theory.com, "Slot Antenna," http://www.antenna-theory.com/antennas/aperture/slot.php.

[152] A. D. Olver and P. J. Clarricoats, *Microwave horns and feeds*. IET, 1994, vol. 39.

[153] B. Thomas, "Design of corrugated conical horns," *IEEE Transactions on Antennas and Propagation*, vol. 26, no. 2, pp. 367–372, 1978.

[154] P. J. B. Clarricoats and A. D. Olver, *Corrugated horns for microwave antennas*. IET, 1984, no. 18.

[155] P. Gibson, "The Vivaldi aerial," in *1979 9th European Microwave Conference*. IEEE, 1979, pp. 101–105.

[156] Wikipedia, "Vivaldi Antenna," https://en.wikipedia.org/wiki/Vivaldi_antenna.

[157] ——, "Cassegrain Antenna," https://en.wikipedia.org/wiki/Cassegrain_antenna.

[158] ——, "Cassegrain Reflector," https://en.wikipedia.org/wiki/Cassegrain_reflector.

[159] W. A. Imbriale, S. S. Gao, and L. Boccia, *Space antenna handbook*. John Wiley & Sons, 2012.

[160] 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.

[161] D.-C. Chang and M.-C. Huang, "Microstrip reflectarray antenna with offset feed," *Electronics Letters*, vol. 28, no. 16, pp. 1489–1491, 1992.

[162] 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.

[163] 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.

[164] D. De Schweinitz and T. L. Frey Jr, "Artificial dielectric lens antenna," Nov. 13 2001, US Patent 6,317,092.

[165] K.-L. Wong, "Planar antennas for wireless communications," *Microwave Journal*, vol. 46, no. 10, pp. 144–145, 2003.

[166] H. Nakano, M. Yamazaki, and J. Yamauchi, "Electromagnetically coupled curl antenna," *Electronics Letters*, vol. 33, no. 12, pp. 1003–1004, 1997.

[167] 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.

[168] K. Luk, C. Mak, Y. Chow, and K. Lee, "Broadband microstrip patch antenna," *Electronics letters*, vol. 34, no. 15, pp. 1442–1443, 1998.

[169] M. Bolic, D. Simplot-Ryl, and I. Stojmenovic, *RFID systems: research trends and challenges*. John Wiley & Sons, 2010.

[170] 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.

[171] 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.

[172] 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.

[173] C.-T. Chen, *Linear system theory and design*. Oxford University Press, Inc., 1998.

[174] S. H. Schot, "Eighty years of Sommerfeld's radiation condition," *Historia mathematica*, vol. 19, no. 4, pp. 385–401, 1992.

[175] D. Hinton, *Analects*. Counterpoint, 2014.

[176] H. Bible, "New Revised Standard Version (NRSV)," *Grand Rapids MI: Zondervan*, 1989.

[177] V. Rumsey, "Reaction concept in electromagnetic theory," *Physical Review*, vol. 94, no. 6, p. 1483, 1954.

[178] R. E. Collin, *Foundations for microwave engineering*. John Wiley & Sons, 2007, also 1966.

[179] 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.

[180] A. Ruehli, G. Antonini, and L. Jiang, *Circuit oriented electromagnetic modeling using the PEEC techniques*. John Wiley & Sons, 2017.

[181] A. Ishimaru, *Electromagnetic wave propagation, radiation, and scattering from fundamentals to applications*. Wiley Online Library, 2017, also 1991.

[182] 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.

[183] Wikipedia, "Christiaan Huygens," https://en.wikipedia.org/wiki/Christiaan_Huygens.

[184] ——, "George Green (mathematician)," https://en.wikipedia.org/wiki/George_Green_(mathematician).

[185] C.-T. Tai, *Dyadic Green's Functions in Electromagnetic Theory*. PA: International Textbook, Scranton, 1971.

[186] ——, *Dyadic Green functions in electromagnetic theory*. Institute of Electrical & Electronics Engineers (IEEE), 1994.

[187] W. Franz, "Zur formulierung des huygensschen prinzips," *Zeitschrift für Naturforschung A*, vol. 3, no. 8-11, pp. 500–506, 1948.

[188] Wikipedia, "Faraday Cage," https://en.wikipedia.org/wiki/Faraday_cage.

[189] J. A. Stratton, *Electromagnetic Theory*. McGraw-Hill Book Company, Inc., 1941.

[190] W. Meissner and R. Ochsenfeld, "Ein neuer effekt bei eintritt der supraleitfähigkeit," *Naturwissenschaften*, vol. 21, no. 44, pp. 787–788, 1933.

[191] Wikipedia, "Superconductivity," https://en.wikipedia.org/wiki/Superconductivity.

[192] 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.

[193] Wikipedia, "Snell's law," https://en.wikipedia.org/wiki/Snell's_law.

[194] 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.

[195] W. J. Smith, *Modern optical engineering.* McGraw-Hill New York, 1966, vol. 3.

[196] 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.

[197] 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.

[198] G. A. Deschamps, "Ray techniques in electromagnetics," *Proceedings of the IEEE*, vol. 60, no. 9, pp. 1022–1035, 1972.

[199] 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.

[200] R. Kouyoumjian, "The geometrical theory of diffraction and its application," in *Numerical and Asymptotic Techniques in Electromagnetics.* Springer, 1975, pp. 165–215.

[201] 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.

[202] Wikipedia, "Fermat's principle," https://en.wikipedia.org/wiki/Fermat's_principle.

[203] 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.

[204] 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.

[205] A. Sommerfeld, *Partial differential equations in physics.* Academic Press, 1949, vol. 1.

[206] R. Haberman, *Elementary applied partial differential equations.* Prentice Hall Englewood Cliffs, NJ, 1983, vol. 987.

[207] G. A. Deschamps, "Gaussian beam as a bundle of complex rays," *Electronics letters*, vol. 7, no. 23, pp. 684–685, 1971.

[208] 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.

[209] D. L. Andrews, *Structured light and its applications: An introduction to phase-structured beams and nanoscale optical forces.* Academic Press, 2011.

[210] 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.

[211] 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.

[212] 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.

[213] G. Mie, "Beiträge zur optik trüber medien, speziell kolloidaler metallösungen," *Annalen der physik*, vol. 330, no. 3, pp. 377–445, 1908.

[214] Wikipedia, "Mie scattering," https://en.wikipedia.org/wiki/Mie_scattering.

[215] L. B. Felsen and N. Marcuvitz, *Radiation and scattering of waves*. John Wiley & Sons, 1994, also 1973, vol. 31.

[216] P. P. Ewald, "Die berechnung optischer und elektrostatischer gitterpotentiale," *Annalen der physik*, vol. 369, no. 3, pp. 253–287, 1921.

[217] E. Whitaker and G. Watson, *A Course of Modern Analysis*. Cambridge Mathematical Library, 1927.

[218] J. Kong, "Electromagnetic fields due to dipole antennas over stratified anisotropic media," *Geophysics*, vol. 37, no. 6, pp. 985–996, 1972.

[219] Wikipedia, "FLOPS," https://en.wikipedia.org/wiki/FLOPS.

[220] ——, "Computer," https://en.wikipedia.org/wiki/computer.

[221] W. C. H. McLean, *Strongly elliptic systems and boundary integral equations*. Cambridge University Press, 2000.

[222] G. C. Hsiao and W. L. Wendland, *Boundary integral equations*. Springer, 2008.

[223] P. K. Banerjee and R. Butterfield, *Boundary element methods in engineering science*. McGraw-Hill London, 1981, vol. 17.

[224] O. C. Zienkiewicz, R. L. Taylor, P. Nithiarasu, and J. Zhu, *The finite element method*. McGraw-Hill London, 1977, vol. 3.

[225] 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.

[226] 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.

[227] J.-M. Jin, *The finite element method in electromagnetics*. John Wiley & Sons, 2015.

[228] G. Strang, *Linear algebra and its applications*. Academic Press, 1976.

[229] Cramer and Gabriel, *Introduction a l'analyse des lignes courbes algebriques par Gabriel Cramer...* chez les freres Cramer & Cl. Philibert, 1750.

[230] J. A. Schouten, *Tensor analysis for physicists*. Courier Corporation, 1989.

[231] A. C. Polycarpou, "Introduction to the finite element method in electromagnetics," *Synthesis Lectures on Computational Electromagnetics*, vol. 1, no. 1, pp. 1–126, 2005.

[232] J. P. A. Bastos and N. Sadowski, *Electromagnetic modeling by finite element methods*. CRC press, 2003.

[233] Ö. Özgün and M. Kuzuoğlu, *MATLAB-based Finite Element Programming in Electromagnetic Modeling*. CRC Press, 2018.

[234] 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.

[235] 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.

[236] 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.

[237] 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.

[238] 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.

[239] A. Taflove and S. C. Hagness, *Computational electrodynamics: the finite-difference time-domain method*. Artech house, 2005, also 1995.

[240] W. Yu, R. Mittra, T. Su, Y. Liu, and X. Yang, *Parallel finite-difference time-domain method*. Artech House Norwood, 2006.

[241] D. Potter, "Computational physics," 1973.

[242] W. F. Ames, *Numerical methods for partial differential equations*. Academic press, 2014, also 1977.

[243] K. W. Morton, *Revival: Numerical Solution Of Convection-Diffusion Problems (1996)*. CRC Press, 2019.

[244] K. Aki and P. G. Richards, *Quantitative seismology*, 2002.

[245] F. B. Hildebrand, *Introduction to numerical analysis*. Courier Corporation, 1987.

[246] W. C. Chew, "Electromagnetic theory on a lattice," *Journal of Applied Physics*, vol. 75, no. 10, pp. 4843–4850, 1994.

[247] J. v. Neumann, *Mathematische Grundlagen der Quantenmechanik, Berlın.* Springer, New York, Dover Publications, 1943.

[248] R. Courant, K. Friedrichs, and H. Lewy, "Über die partiellen differenzengleichungen der mathematischen physik," *Mathematische annalen*, vol. 100, no. 1, pp. 32–74, 1928.

[249] M. Desbrun, A. N. Hirani, M. Leok, and J. E. Marsden, "Discrete exterior calculus," *arXiv preprint math/0508341*, 2005.

[250] W. C. Chew, "Vector potential electromagnetics with generalized gauge for inhomogeneous media: Formulation," *Progress In Electromagnetics Research*, vol. 149, pp. 69–84, 2014.

[251] 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.

[252] 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.

[253] 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.

[254] B. Bapat, "Newton's rings," http://www.iiserpune.ac.in/~bhasbapat/phy221_files/NewtonsRing.pdf.

[255] Wikipedia, "Double-slit experiment," https://en.wikipedia.org/wiki/Double-slit_experiment.

[256] Shmoop.Com, "Young's double-slit," https://www.shmoop.com/optics/young-double-slit.html.

[257] Wikipedia, "John Dalton," https://en.wikipedia.org/wiki/John_Dalton.

[258] ——, "Max Planck," https://en.wikipedia.org/wiki/Max_Planck.

[259] ——, "Photoelectric effect," https://en.wikipedia.org/wiki/Photoelectric_effect.

[260] ——, "Louis de Broglie," https://en.wikipedia.org/wiki/Louis_de_Broglie.

[261] ——, "Newton's laws of motion," https://en.wikipedia.org/wiki/Newton's_laws_of_motion.

[262] ——, "Quantum electrodynamics," https://en.wikipedia.org/wiki/Quantum_electrodynamics.

[263] C. Christopoulos, *The transmission-line modeling method: TLM.* IEEE New York, 1995, vol. 221.

[264] Wikipedia, "William Rowan Hamilton," https://en.wikipedia.org/wiki/William_Rowan_Hamilton.

[265] 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.

[266] Wikipedia, "Gaussian beam," https://en.wikipedia.org/wiki/Gaussian_beam.

[267] M. Kira and S. W. Koch, *Semiconductor quantum optics*. Cambridge University Press, 2011.

[268] Wikipedia, "Quantum harmonic oscillator," https://en.wikipedia.org/wiki/Quantum_harmonic_oscillator.

[269] ——, "Roy J. Glauber," https://en.wikipedia.org/wiki/Roy_J._Glauber.

[270] ——, "E.C. George Sudarshan," https://en.wikipedia.org/wiki/E._C._George_Sudarshan.

[271] C. Gerry, P. Knight, and P. L. Knight, *Introductory quantum optics*. Cambridge University Press, 2005.

Index

A source on top of a layered medium, 389
    horizontal electric dipole, 391
    vertical electric dipole, 390
Absorbing boundary conditions, 428
Advective equations, 32
AM 920 station, 74
Ampere's law, 7, 186
Andrew Lloyd Weber
    Cats, 84
Anisotropic media, 69
Anisotropic medium, 93
Antenna
    net gain, 335
    radiation power density, 335
    resonance tunneling, 304
    transmit and receive patterns, 335
Aperture antenna, 208
Array antennas, 293
    broadside array, 296
    endfire array, 297
    far-field approximation, 294, 298
    linear phase array, 295
    lobes, 296
    radiation pattern, 295
Axial ratio, 96

Back-of-cab formula, 40
Back-of-the-cab formula, 31, 187, 188
Bell's theorem, 7
Bessel equation, 199
Bessel function, 187, 199, 205
Bi-anisotropic media, 69
Biot-Savart law, 50
Born, Max, 443
Bound electron case
    Drude-Lorentz-Sommerfeld model, 85
Boundary conditions, 41
    Ampere's law, 46
    conductive media, 52
    Faraday's law, 44
    Gauss's law, 45, 48
Boundary-value problems
    BVP, 39
Branch cuts, 396
Brewster's angle, 155
Brief history
    electromagnetics, 6
    optics, 6

Cats
    Andrew Lloyd Weber, 84
Cavity resonator, 219
    cylindrical waveguide resonator, 221
    full-width half maximum bandwidth, 237
    pole location, 236
    quality factor, 233
    quarter wavelength resonator, 221
    transfer function, 236
    transmission line model, 219
Characteristic impedance, 214
Circuit theory, 257
    capacitance, 263
    energy storage method, 264
    inductor, 262
    Kirchhoff current law, 257
    Kirchhoff voltage law, 257
    resistor, 263

Circular waveguide, 187, 198, 205
  cut-off frequency, 200
  eigenmode, 199
Circularly polarized wave, 208
Coherent state, 451
Compass, 4
Complex power, 65
Computational electromagnetics, 401
Conductive media, 69
  highly conductive case, 80
  lowly conductive case, 81
Conjugate gradient method, 412
Constitutive relations, 29, 69
  anisotropic media, 71
  bi-anisotropic media, 72
  inhomogeneous media, 72
  isotropic frequency dispersive media, 69
  nonlinear media, 73
  uniaxial and biaxial media, 72
Constructive interference, 175, 176, 178
Copenhagen school
  quantum interpretation, 7
Corpuscular nature of light, 103
Coulomb gauge, 41, 247
Coulomb's Law, 8
Curl operator, definition, 22
Curl operator, physical meaning, 26

Damping or dissipation case
  Drude-Lorentz-Sommerfeld model, 85
de Broglie, Louis, 433
Derivation
  Gauss's law from Coulomb's law, 12
Diagonalization, 92
Dielectric slab, 152
Dielectric waveguide, 175
  cut-off frequency, 181, 182, 184
  TE polarization, 178
  TM polarization, 183
Differential equation, 401
Differential forms, 425
Dirac, Paul, 2
Directive gain pattern, 276, 289
  directivity, 290
Dirichlet boundary condition, 189, 194, 201
Discrete exterior calculus, 425
Dispersion relation, 80, 147, 151
  plane wave, 75
Divergence operator, physical meaning, 21
Divergence, definition of, 17
Drude-Lorentz-Sommerfeld model, 82, 169, 437
  bound electron case, 85
  broad applicability of, 86
  damping or dissipation case, 85
Duality principle, 140, 149, 162, 212, 214

Effective aperture, 336, 338
Effective mass, 88, 89
Einstein, 2
Electric current, 342
Electric dipole, 142
Electric susceptibility, 30
Electromagnetic compatibility, 349
Electromagnetic interference, 349
Electromagnetic theory
  unification, 6
Electromagnetic wave, 73
  triumph of Maxwell's equations, 30
Electromagnetics, 1
  history, 1
  importance, 1
Electrostatics, 34
Equivalence theorems, 339
  general case, 341
  inside-out case, 340
  outside-in case, 341
Euler's formula, 62
Evanescent wave, 152, 159
  plasma medium, 84
Extinction theorem, 341, 345

Far zone, 300
Faraday rotation, 91
Faraday's law, 7
Fermat's principle, 363, 392
Finite-Difference Time-Domain, 415
  grid-dispersion error, 422

stability analysis, 420
Yee algorithm, 424
Folded dipole, 305
Fourier transform technique, 64, 65
Frequency dispersive media, 69, 88
Frequency domain analysis, 61
Fresnel reflection coefficient, 145
Fresnel zone, 300
Fresnel, Austin-Jean, 145
Functional, 411
Functional space, 241
Hilbert space, 242
Fundamental mode, 192

Gauge theory, 2
Gauss's divergence theorem, 17, 19
Gauss's Law, 11
Gauss's law
differential operator form, 21
Gaussian beam, 366
Gauss's law, 7
Gedanken experiment, 339
General reflection coefficient, 122
Generalized impedance, 123
Generalized reflection coefficient, 137, 166, 175
Generalized transverse resonance condition, 175, 176, 223
Geometrical optics, 362
Goos-Hanschen shift, 152, 178
Goubau line, 305, 313
Green's function
dyadic, 347
scalar, 344
static, 36
Green's function method, 248
Green's theorem, 344
vector, 347
Green, George, 343
Group velocity, 170, 182, 192, 194
Guidance condition, 175, 177, 178, 183, 191
Guided mode, 173, 213
Gyrotropic media
Faraday rotation, 91
Gyrotropic medium, 93

Hamilton, William Rowan, 438
Hamiltonian mechanics, 438
Hankel function, 199, 205
Heaviside, 28
Heaviside layer
Kennelly-Heaviside layer, 84
Heinrich Hertz, 271
experiment, 5
Helmholtz equation, 116, 170, 187, 188
Hermite polynomial, 369
Hermite-Gaussian functions, 441
Hermitian matrix, 240
Hertzian dipole, 271, 301
far field, 276
near field, 275
point source approximation, 273
radiation resistance, 279
Hidden variable theorey, 7
Hollow waveguide, 185, 190, 211
cut-off frequency, 192
cut-off mode, 190
cut-off wavelength, 192
eigenmode, 191, 194
eigenvalue, 191
eigenvector, 191
TE mode, 186, 190, 212
TM mode, 186, 189
Homogeneous solution, 159, 173, 219, 234, 235, 318, 321–323
Hooke's law, 437
Horn antenna, 208, 307
Huygens' principle
electromagnetic waves case, 346
scalar waves case, 344
Huygens, Christiaan, 343
Hybrid modes, 210

Image theory, 351
electric charges, 352
electric dipoles, 352
magnetic charges, 353
magnetic dipoles, 353
multiple images, 355

perfect magnetic conductor surfaces, 354
Impressed current, 56–58, 293, 328, 331–334, 342
Inhomogeneous media, 69
Initial value proble, IVP, 323
Integral equation, 402
Intrinsic impedance, 77, 99

James Clerk Maxwell
displacement current, 5
Jump condition, 41, 43, 48
Jump discontinuity, 43, 45–48

Kao, Charles, 176
Kernel, 403
Kinetic inductance, 212

Laguere polynomials, 369
associated, 369
Laplace's equation, 36
Laplacian operator, 31, 35
Lateral wave, 152
Layered media, 166
LC tank, 437
Leap-frog scheme, 426
Legendre polynomial
associate, 382
Lenz's law, 262
Linear phase array, 295
Linear time-invariant system, 88
Lode stone, 4
Lorentz force law, 82, 91
Lorenz gauge, 248
Lossless conditions, 105, 106
Lossy media, 69
Love's equivalence theorem, 404

Magnetic charge, 142, 249
Magnetic current, 142, 249, 343
Magnetic dipole, 142
Magnetic monopole, 142
Magnetic source, 140
Magnetostatics, 39
Marconi
experiment, 84
Marconi, Guglielmo, 271
Matrix eigenvalue problem, 191, 240
Maxwell, 28
lifespan, 28
marriage, 28
poems, 28
Maxwell's equations
accuracy, 2
broad impact, 2
differential operator form, 26
integral form, 7
quantum regime, 2
relativistic invariance, 2
valid over vast length scale, 1
Yang-Mills theory, 2
Maxwell's theory
wave phenomena, 5
Maxwell, James Clerk, 1
Metasurfaces, 365
Microstrip patch antenna, 307
Mie scattering, 379
Mode conversion, 216
Mode matching method, 216
Mode orthogonality, 240
resonant cavity modes, 244
waveguide modes, 242
Momentum density, 103
Mono-chromatic signal, 168

Nano-fabrication, 6
Nanoparticles, 89
Natural solution, 219, 235, 321, 322
Near zone, 300
Negative resistance, 58
Neumann boundary condition, 188, 190, 191, 200
Neumann function, 199
Nonlinear media, 69
Normalized generalized impedance, 123
Null space solution, 173, 322
Numerical methods, 401

Oliver Heaviside, 84
One-way wave equations, 32
Optical fiber, 152, 210

Optical theorem, 380
Optimization, 410

Paraxial wave, 206
Paraxial wave equation, 366
Pendulum, 434
Perfect electric conductor, 351
Perfectly matched layers, 429
Permeability, 30
Permittivity, 30
Phase matching condition, 147, 150, 157, 175
Phase shift, 152
Phase velocity, 77, 88, 168, 182, 192, 210
Phasor technique, 61, 62, 79, 83
phasor technique, 65
Photo-electric effect, 6
Photoelectric effect, 433
Photon number state, 441
Pilot potential, 196, 205
Pilot potential method, 186
Pilot vector, 187
Planck radiation law, 6
Planck, Max, 433
Plane waves
    in lossy conductive media, 79
    uniform, 75
Plasma medium, 191
    cold collisionless, 83
Plasmonics, 89
Poisson's equation, 35
Polarizability, 374
Polarization, 94, 102
    CP, 94
    LHCP, 95
    LHEP, 95
    RHCP, 95
    RHEP, 95
Polarization density, 29
Power flow
    LHCP, 98
    LHEP, 98
    RHCP, 99
    RHEP, 99
Poynting's theorem
    complex, 104
    instantaneous, 56
Poynting's vector
    complex, 66, 99
    instantaneous, 58, 65
Probability distribution function, 443

Quantum electrodynamics, 434
Quantum harmonic oscillator, 440, 448
Quantum theory, 2
Quasi-optical antennas, 309
Quasi-plane wave, 366
Quasi-static electromagnetic theory, 250, 255
Quasi-TEM mode, 210

Radiation condition; Sommerfeld radiation condition, 325
Radiation fields, 284
    effective aperture, 290
    effective area, 290
    far-field approximation, 284
    locally plane wave approximation, 286
Rayleigh distance, 299
Rayleigh scattering, 371
Reaction
    Rumsey, 330
Reactive power, 67
Reciprocity theorem, 342
    conditions, 331
    mathematical derivation, 328
    two-port network, 331
Rectangular cavity, 222
Rectangular waveguide, 187, 190, 194, 195
    cut-off frequency, 194
    TE mode, 214
    TM mode, 214
Reduced wave equation, 189
Reflection coefficient, 122, 135, 136, 145, 156, 158, 162
Reflector antenna, 278
Refractive index, 177
Relativistic invariance, 2
Resonant solution, 219, 320–324

Resonant solution, homogeneous solution, 322
Resonator, 219
Retardation, 116
RFID, 312
Riemann sheets, 396
Rumsey reaction theorem, 330

Scalar potential, 187, 190
  electrodynamics, 247
  statics, 246
Schrödinger equation, 440
Separation of variables, 190, 198
Shielding, 349
  electrostatic, 349
  relaxation time, 350
Shielding by conductive media, 52
Slot antenna, 306
Snell's law, 147, 157, 364
  generalized, 365
Sommerfeld identity, 387, 392
Sommerfeld radiation condition, 325
Spatially dispersive, 69
Special relativity, 2
Spectral representations, 383
  point source, 384
Spherical Bessel equation, 382
Spherical Bessel function, 382
Spherical Hankel functions, 382
Spherical harmonic, 382
Spherical Neumann function, 382
Spin angular momentum, 102
Standing wave, 190, 195
Static electricity, 4
Static electromagnetics
  differential operator form, 34
  integral form, 8
Statics
  electric field, 8
Stationary phase method, 392
Stokes's theorem, 22
structured lights, 370
Subspace projection, 407
Surface current, 340
Surface plasmon, 158, 173
Surface plasmonic mode, 212
Surface plasmonic waveguide, 212

Tangent plane approximations, 362
TE polarization, 146
  reflection coefficient, 148
  transmission coefficient, 148
Telegrapher's equations, 118, 122, 214, 215
  frequency domain, 116
  time domain, 115
Telegraphy
  lack of understanding, 5
TEM mode, 186
Time-Harmonic fields, 62
TM mode, 194
TM polarization, 149
  reflection coefficient, 149
  Transmission coefficient, 149
Toroidal antenna, 142
Total internal reflection, 150, 178, 181
Transmission coefficient, 156, 162, 163
Transmission line, 112, 121, 214
  capacitive effects, 113
  characteristic impedance, 115, 119
  current, 113, 122
  distributed lumped element model, 113
  equivalent circuit, 135, 136, 138
  frequency analysis, 116
  inductive effects, 113
  load, 122
  lossy, 118
  matched load, 123
  parasitic circuit elements, 140
  stray/parasitic capacitances, 139
  stray/parasitic inductance, 139
  time-domain analysis, 113
  voltage, 113
Transmission line matrix, 333, 434
Transverse resonance condition, 174, 175, 220
Transverse wave number, 213
Trapped wave, 159
Traveling wave, 195

Twin-lead transmission line, 305

Uncertainty principle, 445
Uniqueness theorem, 317, 324, 339, 341–343, 348
    conditions, 320
    connection to poles of a linear system, 323
    radiation from antenna sources, 324

Vector Poisson's equation, 40
Vector potential
    electrodynamics, 247
    non-unique, 41
    statics, 246
Voltaic cell
    telegraphy, 4

Wave equations, 115
Wave packet, 169
Wave Phenomenon
    frequency domain, 73
Wave-particle duality, 431
Weyl identity, 385, 392

Yagi-Uda antenna, 306
Young, Thomas, 431