Electromagnetism
under 100 pages

Max Yan

Department of Applied Physics
School of Engineering Sciences
KTH - Royal Institute of Technology, Sweden

November 20, 2022

The compendium is primarily used for course *SK1118 - Electromagnetism and Waves* at KTH. The course is attended by students from two programs, *TCOMK "Kandidatprogram, informations- och kommunikationsteknik"* and *CINTE "Civilingenjörsutbildning i informationsteknik"*. It is also partly used for second-cycle course *SK2402 - Fundamentals of Photonics* at KTH.

$$\begin{aligned}
\nabla \cdot \mathbf{D} &= \rho, \\
\nabla \cdot \mathbf{B} &= 0, \\
\nabla \times \mathbf{E} &= -\frac{\partial \mathbf{B}}{\partial t}, \\
\nabla \times \mathbf{H} &= \mathbf{J} + \frac{\partial \mathbf{D}}{\partial t}.
\end{aligned}$$

$$\begin{aligned}
\oint_S \mathbf{D} \cdot d\mathbf{s} &= Q, \\
\oint_S \mathbf{B} \cdot \ d\mathbf{s} &= 0, \\
\oint_C \mathbf{E} \cdot \ d\mathbf{l} &= -\int_S \frac{\partial \mathbf{B}}{\partial t} \cdot d\mathbf{s}, \\
\oint_C \mathbf{H} \cdot \ d\mathbf{l} &= I + \int_S \frac{\partial \mathbf{D}}{\partial t} \cdot \ d\mathbf{s}.
\end{aligned}$$

Constants and Units of Physical Quantities

Constant	**Value**	**Remark**
Speed of light (free space) c	$\sim 3 \times 10^8$ m/s	Universal constant
Permeability (free space) μ_0	$4\pi \times 10^{-7}$ H/m	Universal constant
Permittivity (free space) ϵ_0	$\sim \frac{1}{36\pi} \times 10^{-9}$ F/m	Derived from $c = \sqrt{1/\epsilon_0\mu_0}$
Impedance (free space) Z_0	$\sim 120\pi$ or 377 Ω	$Z_o = \sqrt{\mu_0/\epsilon_0}$
Electron charge $-e$	-1.602×10^{-19} C	
Electron mass (rest) m_e	9.107×10^{-31} kg	

Quantity	**Unit**	**Remark [Svenska]**
Length	m (meter)	SI base unit [Längd]
Mass	kg (kilogram)	SI base unit [Vikt]
Time	s (second)	SI base unit [Tid]
Current	A (ampere)	SI base unit [Ström]
Admittance Y	S (siemens)	[Admittans]
Angular frequency ω	rad/s	[Vinkelfrekvens]
Capacitance F	F (farad)	C/V [Kapacitans]
Charge Q or q	C (coulomb)	A·s [Laddning]
Conductance G	S (siemens, $\equiv 1/\Omega$)	[Konduktans]
Conductivity σ	S/m	[Konduktivitet]
Current density (volume) $\mathbf{J}$	A/m^2	[Strömdensitet]
Electric field intensity $\mathbf{E}$	V/m	[Elektriskt fält(styrka)]
Electric potential V	V (volt)	[Elektrisk spänning]
Electric susceptibility χ_e	–	[Elektrisk susceptibilitet]
Electromotiv force $\mathscr{E}$	V	[Elektromotorisk spänning]
Energy (work) W	J (joule)	[Arbete]
Electric flux density $\mathbf{D}$	C/m^2	aka “Displacement” [E. flödstäthet]
Force $\mathbf{F}$	N (newton)	[Kraft]
Frequency f	Hz (hertz)	[Frekvens]
Impedance Z	Ω (ohm)	[Impedans]
Inductance L	H (henry)	[Induktans]
Magnetic dipole moment $\mathbf{m}$	A·m^2	[M. dipolmoment]
Magnetic field intensity $\mathbf{H}$	A/m	[Magnetfält]
Magnetic flux Φ	Wb (weber)	[Magnetiskt flöde]
Magnetic flux density $\mathbf{B}$	T (tesla)	[Magnetiskt flödestäthet]
Magnetic susceptibility χ_m	–	[Magnetisk susceptibilitet]
Magnetization $\mathbf{M}$	A/m	[Magnetisering]
Permeability μ	H/m	[Permeabilitet]
Permittivity ϵ	F/m	[Permittivitet]
Phase ϕ	rad (radian)	[Fas]
Polarization $\mathbf{P}$	C/m^2	[Elektrisk polarisation]
Power P	W (watt)	[Effekt]
Poynting vector $\mathscr{P}$	W/m^2	[Poyntings vector]
Resistance R	Ω	[Resistans]
Voltage V	V	[Spänning]
Wavelength λ	m	[Våglängd]
Wave vector $\mathbf{k}$	rad/m	[Vågvektor]
Wave number $k = \lvert\mathbf{k}\rvert$	rad/m	[Vågtal]

Contents

Chapter 1

Vectors Analysis and Other Fundamentals

1.1 What is vector?

A *vector* is a value carrying both magnitude and direction, like fluid velocity. In comparison, a *scalar* has only magnitude, like temperature. A *scalar field* is distribution of a scalar quantity over space and time, like temperature in a room. A *vector field* is distribution of a vector quantity over space and time, like fluid velocity in a pipe. Graphically, a vector is represented as an arrow placed in 3D space, which is mathematically denoted as a bold alphabet, e.g. $\mathbf{A}$, to differetiate from a scalar A. In handwriting, it is more often to write as $\vec{A}$ or sometimes $\bar{A}$. Besides representing a physical quantity, a vector can also merely represent how one point is positioned with respect to another posint in 3D space. If the reference point is the origin, we call such a vector *position vector* of a point.

A vector is expressed generally as

$$\mathbf{A} = A\hat{\boldsymbol{a}}, \tag{1.1}$$

where A is the magnitude of $\mathbf{A}$ and $\hat{\boldsymbol{a}}$ is a unit vector (vector with magnitude 1) representing the direction of $\mathbf{A}$. Magnitude A is also written as $|\mathbf{A}|$.

In a more concrete form, a vector can be written in three scalar numbers, plus coordinate information. For example in Cartesian coordinate system, one writes

$$\mathbf{A} = A_x\hat{\boldsymbol{x}} + A_y\hat{\boldsymbol{y}} + A_z\hat{\boldsymbol{z}} \tag{1.2}$$

where $\hat{\boldsymbol{x}}$, $\hat{\boldsymbol{y}}$, and $\hat{\boldsymbol{z}}$ are unit vectors along three coordinates. A_x, A_y and A_z are lengthes of projections of the vector along three axial directions. The magnitude and direction of the vector can be calculated based on the three scalars.

Example: Position vector

Calculate position vector of a point P_1 with Cartesian coordinate $(2, 4, 5)$.

Solution: $\mathbf{P}_1 = 2\hat{\boldsymbol{x}} + 4\hat{\boldsymbol{y}} + 5\hat{\boldsymbol{z}}$.

If one knows position vectors of spatial points P_1 (as $\mathbf{P}_1$) and P_2 (as $\mathbf{P}_2$), the vector pointing from P_1 to P_2, commonly referred to as *displacement vector*, is

$$\mathbf{P}_{12} = \mathbf{P}_2 - \mathbf{P}_1. \tag{1.3}$$

Example: Displacement vector

In Cartesian coordinate, calculate vector pointing from point P_1 at $(2, 4, 5)$ to point P_2 at $(3, 3, 1)$.

Solution: Position vector of P_1: $\mathbf{P}_1 = 2\hat{\boldsymbol{x}} + 4\hat{\boldsymbol{y}} + 5\hat{\boldsymbol{z}}$.
Position vector of P_2: $\mathbf{P}_2 = 3\hat{\boldsymbol{x}} + 3\hat{\boldsymbol{y}} + 1\hat{\boldsymbol{z}}$.
Vector from P_1 to P_2: $\mathbf{P}_{12} = \mathbf{P}_2 - \mathbf{P}_1 = (3-2)\hat{\boldsymbol{x}} + (3-4)\hat{\boldsymbol{y}} + (1-5)\hat{\boldsymbol{z}} = \hat{\boldsymbol{x}} - \hat{\boldsymbol{y}} - 4\hat{\boldsymbol{z}}$.

Note that coordinate is really what we impose on a physical space. One can translate, rotate, and shrink a coordinate system, or even transform one coordinate system (e.g. Cartesian) to another coordinate system (e.g. cylindrical). Upon such a coordinate transformation, the three scalar numbers and the unit vectors denoting a vector shall be changed correspondingly.

Why vector analysis? Electromagnetism is about finding *vector electric and magnetic fields* in space and time. The underlying laws governing the physics of electromagnetism (i.e. Maxwell's equations) are summarized very compactly using vector operations, without dependence on a specific coordinate system. In fact, in year 1861, Maxwell's equations were written in 20 equations with variables expressed in Cartesian coordinate system. The equations would change forms when another coordinate system is used. The current vector Maxwell's equations were formulated by Oliver Heaviside in year 1884. The formulation is based on *vector operations*. Knowledge on such operations is vital for understanding of this subject.

1.2 Vector operations

Vector addition:

$$\mathbf{C} = \mathbf{A} + \mathbf{B} \tag{1.4}$$

Vector addition is commutative. Geometrically, **C** can be obtained from **A** and **B** through the parallelogram rule. Arithmetically, addition is done by adding corresponding scalar coefficients. In Cartesian coordinate, one has

$$\mathbf{C} = (A_x + B_x)\hat{\boldsymbol{x}} + (A_y + B_y)\hat{\boldsymbol{y}} + (A_z + B_z)\hat{\boldsymbol{z}}. \tag{1.5}$$

Vector subtraction:

$$\mathbf{C} = \mathbf{A} - \mathbf{B} = \mathbf{A} + (-\mathbf{B}) \tag{1.6}$$

This can be related to the example on "Displacement vector" above: treating **A** and **B** respectively as position vectors of points A and B, **C** is the vector pointing from B to A. Subtraction can be treated as addition (through vector inversion).

Dot product:

$$\mathbf{A} \cdot \mathbf{B} \equiv AB\cos\theta \tag{1.7}$$

where θ is the angle between vectors **A** and **B**. Dot product is commutative. Geometrically, dot product calculates scalar product of A and projection of **B** on the direction of **A**. Besides geometrical interpretation, one can also use arithmetic operation based on scalar coefficients of the vectors to calculate dot product. In Cartesian coordinate, it takes the form

$$\mathbf{A} \cdot \mathbf{B} = A_xB_x + A_yB_y + A_zB_z. \tag{1.8}$$

Sometimes, "dot product" is also referred to as "scalar product".

Cross product:

$$\mathbf{A} \times \mathbf{B} \equiv AB\sin\theta\ \hat{\boldsymbol{n}} \tag{1.9}$$

where θ is the angle between vectors **A** and **B**, and $\hat{\boldsymbol{n}}$ is unit vector perpendicular to the plane defined by **A** and **B**. The specific direction of $\hat{\boldsymbol{n}}$ is decided by the *right-hand rule*[1]. Geometrically, cross product calculates area of a parallelogram formed by **A** and **B**, with result presented as a vector in the third direction.

Instead of calculating area, cross product of two vectors can be calculated through arithmetic operation based on their scalar coefficients. In Cartesian coordinate, it takes the form

$$\mathbf{A}\times\mathbf{B} = (A_yB_z - A_zB_y)\hat{\boldsymbol{x}} + (A_zB_x - A_xB_z)\hat{\boldsymbol{y}} + (A_xB_y - A_yB_x)\hat{\boldsymbol{z}}, \tag{1.10}$$

or in a more compact form

$$\mathbf{A}\times\mathbf{B} = \begin{vmatrix} \hat{\boldsymbol{x}} & \hat{\boldsymbol{y}} & \hat{\boldsymbol{z}} \\ A_x & A_y & A_z \\ B_x & B_y & B_z \end{vmatrix}. \tag{1.11}$$

Cross product is non-commutative. $\mathbf{A}\times\mathbf{B} = -\mathbf{B}\times\mathbf{A}$, following the right-hand rule.

1.3 Coordinate systems

In Eq. 1.2 we expressed a vector in Cartesian coordinate system. As mentioned, an identical vector takes a different form in some other coordinate system. The most common coordinate systems, apart from Cartesian, are cylindrical and spherical coordinate systems. These three belong to *orthogonal* coordinate systems, whose base vectors (or basis vectors) form an orthogonal triplet. Choice of a coordinate system depends on problem under consideration, and is often connected to symmetry of the problem. A formal procedure called *coordinate transformation* can be used to transform a vector from one coordinate to another.

One important aspect of a coordinate system is its *metric coefficients* — how much length changes are incurred by one unit coordinate increment in three axial directions. Cartesian coordinates are based on lengths, with metric coefficient for each coordinate being 1. If we use h_1, h_2, and h_3 to denote metric coefficients in each coordinate direction, we simply have $h_1 = h_2 = h_3 = 1$. For this reason, metric coefficients do not appear in the formulas for Cartesian coordinate. This is not true for cylindrical and spherical coordinates.

1.3.1 Cartesian coordinate

In Cartesian coordinate system, the three base vectors are $\hat{\boldsymbol{x}}$, $\hat{\boldsymbol{y}}$, and $\hat{\boldsymbol{z}}$. Usually one chooses a right-handed system, which has the following cyclic properties

$$\hat{\boldsymbol{x}}\times\hat{\boldsymbol{y}} = \hat{\boldsymbol{z}}, \quad \hat{\boldsymbol{y}}\times\hat{\boldsymbol{z}} = \hat{\boldsymbol{x}}, \quad \hat{\boldsymbol{z}}\times\hat{\boldsymbol{x}} = \hat{\boldsymbol{y}}. \tag{1.12}$$

A vector differential length is expressed in terms of the base vectors as

$$d\mathbf{l} = dx\hat{\boldsymbol{x}} + dy\hat{\boldsymbol{y}} + dz\hat{\boldsymbol{z}}. \tag{1.13}$$

A differential volume is expressed as

$$dv = dxdydz. \tag{1.14}$$

[1] Right-hand rule: one extends right-hand fingers out from **A** and close in against **B**; the thumb then points to $\hat{\boldsymbol{n}}$.

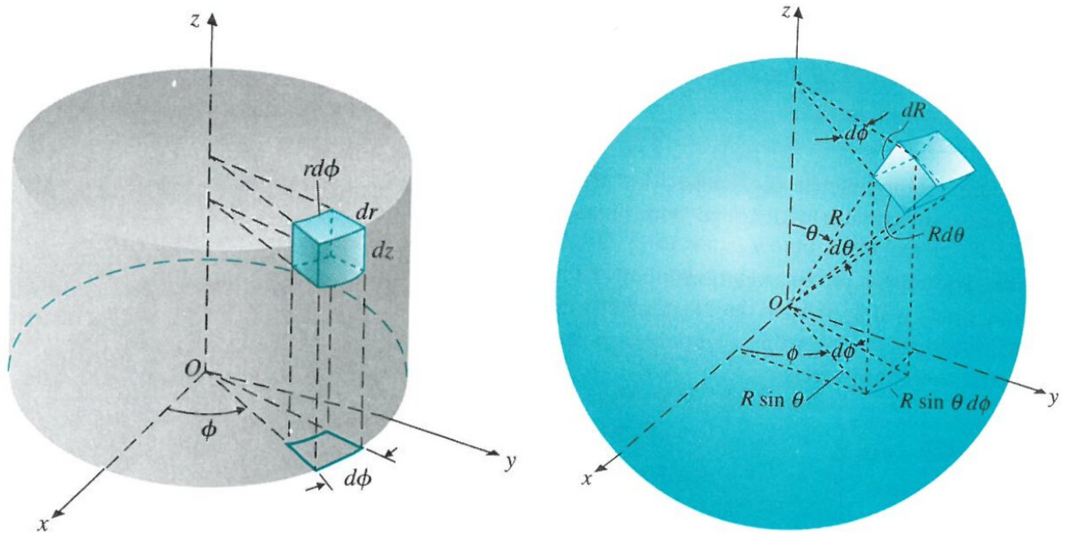

Figure 1.1: Differential volume element in cylindrical and spherical coordinate systems.

1.3.2 Cylindrical coordinate

Cylindrical coordinate system has base vectors $\hat{\boldsymbol{r}}$ (along radial direction), $\hat{\boldsymbol{\phi}}$ (azimuthal direction), and $\hat{\boldsymbol{z}}$. The their relations are

$$\hat{\boldsymbol{r}} \times \hat{\boldsymbol{\phi}} = \hat{\boldsymbol{z}}, \quad \hat{\boldsymbol{\phi}} \times \hat{\boldsymbol{z}} = \hat{\boldsymbol{r}}, \quad \hat{\boldsymbol{z}} \times \hat{\boldsymbol{r}} = \hat{\boldsymbol{\phi}}. \tag{1.15}$$

A vector in cylindrical coordinate system is expressed as

$$\mathbf{A} = A_r \hat{\boldsymbol{r}} + A_\phi \hat{\boldsymbol{\phi}} + A_z \hat{\boldsymbol{z}}. \tag{1.16}$$

r takes value ranging from 0 to ∞, ϕ from 0 to 2π, and z from $-\infty$ to $+\infty$.

In cylindrical coordinate, the azimuthal coordinate ϕ is an angle. One obtains a differential length in that direction by multiplying the angle with an appropriate metric coefficient, in this case r. The corresponding metric coefficients are

$$h_1 = 1, \quad h_2 = r, \quad h_3 = 1. \tag{1.17}$$

A differential length is

$$d\mathbf{l} = dr\hat{\boldsymbol{r}} + rd\phi\hat{\boldsymbol{\phi}} + dz\hat{\boldsymbol{z}}. \tag{1.18}$$

A differential volume is (refer to Fig. 1.1, left panel)

$$dv = rdrd\phi dz. \tag{1.19}$$

Notice how h_2 is used in the above two expressions.

1.3.3 Spherical coordinate

Cylindrical coordinate system has base vectors $\hat{\boldsymbol{r}}$ (along radial direction), $\hat{\boldsymbol{\theta}}$ (angular direction with respect to polar axis), and $\hat{\boldsymbol{\phi}}$ (azimuthal direction). Their relations are

$$\hat{\boldsymbol{r}} \times \hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\phi}}, \quad \hat{\boldsymbol{\theta}} \times \hat{\boldsymbol{\phi}} = \hat{\boldsymbol{r}}, \quad \hat{\boldsymbol{\phi}} \times \hat{\boldsymbol{r}} = \hat{\boldsymbol{\theta}}. \tag{1.20}$$

A vector in sphereical coordinate system is expressed as

$$\mathbf{A} = A_r \hat{\boldsymbol{r}} + A_\theta \hat{\boldsymbol{\theta}} + A_\phi \hat{\boldsymbol{\phi}}. \tag{1.21}$$

r takes value ranging from 0 to ∞, θ from 0 to π with zero pointing at north pole, and ϕ from 0 to 2π.

Two of the coordinates are now angles. The metric coefficients in r, θ, and ϕ directions are respectively

$$h_1 = 1, \quad h_2 = r, \quad h_3 = r\sin\theta. \tag{1.22}$$

A differential length is

$$d\mathbf{l} = dr\hat{\boldsymbol{r}} + rd\theta\hat{\boldsymbol{\theta}} + r\sin\theta d\phi\hat{\boldsymbol{\phi}}. \tag{1.23}$$

A differential volume is (refer to Fig. 1.1, right panel)

$$dv = r^2\sin\theta drd\theta d\phi. \tag{1.24}$$

1.4 Integration of vector field

Line integral (through dot product):

$$\int_C \mathbf{A}\cdot d\mathbf{l}, \quad \text{or} \quad \oint_C \mathbf{A}\cdot d\mathbf{l} \text{ (if line is closed).} \tag{1.25}$$

Subscript C stands for "contour". Note bold "**l**" in vector "d**l**". Line integral of a vector field computes integration of the vector field's magnitude projected *along* the line direction. Integration results in a scalar value.

Line integral of vector field, through dot product

In cylindrical coordinate, calculate line integral *through dot product* of vector field $\mathbf{F} = 1\hat{\boldsymbol{r}}$ along a circle with radius 1 in $r\phi$ plane centered at origin.

Solution: $\oint_C \mathbf{F}\cdot d\mathbf{l} = \oint_C 1\hat{\boldsymbol{r}}\cdot 1d\phi\hat{\boldsymbol{\phi}} = 0$. (since $\hat{\boldsymbol{r}}$ and $\hat{\boldsymbol{\phi}}$ perpendicular to each other)

A vector field can also be integrated along a line without considering line direction, as $\int_C \mathbf{A}dl$. Since $\mathbf{A}$ can be decomposed into three axial components, the line integral can be interpreted as three scalar line integrals. Final result is a vector. In Cartesian coordinate, it is calculated as

$$\int_C \mathbf{A}dl = \int_C A_x dl\hat{\boldsymbol{x}} + \int_C A_y dl\hat{\boldsymbol{y}} + \int_C A_z dl\hat{\boldsymbol{z}}. \tag{1.26}$$

Example: Line integral of vector field, without dot product

In cylindrical coordinate, calculate line integral of vector field $\mathbf{F} = 1\hat{\boldsymbol{r}}$ along a circle with radius 1 in $r\phi$ plane centered at origin.

Solution: $\oint_C \mathbf{F}dl = \oint_C 1\hat{\boldsymbol{r}}\ 1d\phi = \oint_C (1\cos\phi\hat{\boldsymbol{x}} + 1\sin\phi\hat{\boldsymbol{y}})\ 1d\phi = \oint_C \cos\phi d\phi\hat{\boldsymbol{x}} + \oint_C \sin\phi d\phi\hat{\boldsymbol{y}} = [\sin\phi]_{\phi=0}^{2\pi}\hat{\boldsymbol{x}} + [-\cos\phi]_{\phi=0}^{2\pi}\hat{\boldsymbol{y}} = 0$. (Result should be a vector, but now with zero magnitude.)

Surface integral:

$$\int_S \mathbf{A}\cdot d\mathbf{s}, \quad \text{or} \quad \oint_S \mathbf{A}\cdot d\mathbf{s} \text{ (if surface is closed).} \tag{1.27}$$

Subscript S stands for "surface". Surface integral of a vector field computes the total field flux passing through an area. Direction of differential surface element $d\mathbf{s}$ is decided by the right-hand rule, by curving fingers around four sides of the area in a counter-clockwise direction. When performing the right-hand rule, one shall first choose a fixed viewing perspective to the surface. If the surface is a closed surface, one usually views the surface from outside, hence $\hat{\boldsymbol{n}}$ pointing outwards from the surface.

Example: Flux through an area

In cylindrical coordinate, calculate flux of a vector field $\mathbf{F} = 1\hat{\boldsymbol{r}}$ passing through a circular area with radius 1 in $r\phi$ plane centered at origin.

Solution: $\int_S \mathbf{F} \cdot d\mathbf{s} = 0$ (field direction $\hat{\boldsymbol{r}}$ is everywhere perpendicular to surface normal $\hat{\boldsymbol{z}}$).

A vector field can also be integrated with respect to a surface without considering the surface direction/orientation, as

$$\int_S \mathbf{A} ds. \tag{1.28}$$

It again can be interpreted as three scalar surface integrals. Final result is a vector.

Example: Surface integral of vector field, without dot product

In cylindrical coordinate, calculate surface integral of vector field $\mathbf{F} = 1\hat{\boldsymbol{z}}$ over a circular area with radius 1 in $r\phi$ plane centered at origin.

Solution: $\int_S \mathbf{F} ds = \int_r \int_\phi 1\hat{\boldsymbol{z}}\ r d\phi dr = \int_r r dr \int_\phi d\phi\ \hat{\boldsymbol{z}} = \left[\frac{1}{2}r^2\right]_0^1 [\phi]_0^{2\pi} = \pi\hat{\boldsymbol{z}}$.
Extension: How about field is $\mathbf{F} = 1\hat{\boldsymbol{r}}$?

Volume integral:

$$\int_V \mathbf{A} dv. \tag{1.29}$$

Subscript V stands for "volume". Note that a differential volume element dv is a scalar. Volume integral of a vector field effectively corresponds to three volume integrals of scalar fields. The final result is a vector.

1.5 The *del* operator, and its operation on fields

In physical problems, one often needs to manipulate a field to get some derived knowledge of the field. There are three primary operations - *gradient*, *divergence*, and *curl*. In electromagnetism, these operations establish relationships among sources (charge, current), fields (electric and magnetic), and potentials (electric potential or voltage, magnetic potential). It would be cumbersome to express field operations using three coordinate components of a vector field. We introduce an operator – the *del* operator denoted by ∇ (nabla), which simplifies expressions of vector analyses. The operator appears in different forms in different coordinate systems. In Cartesian coordinate, ∇ is defined as

$$\nabla \equiv \frac{\partial}{\partial x}\hat{\boldsymbol{x}} + \frac{\partial}{\partial y}\hat{\boldsymbol{y}} + \frac{\partial}{\partial z}\hat{\boldsymbol{z}}. \tag{1.30}$$

The operator takes the form of a vector. Its physical meaning is best conveyed when it operates on a field.

1.5.1 Gradient

Gradient of a scalar field (∇ operation on a scalar field) is expressed as

$$\nabla A \equiv \frac{\partial A}{\partial x}\hat{\boldsymbol{x}} + \frac{\partial A}{\partial y}\hat{\boldsymbol{y}} + \frac{\partial A}{\partial z}\hat{\boldsymbol{z}}. \tag{1.31}$$

The operation on a scalar field F (a map of scalar values) results in a vector field that describes how fast the field varies at any coordinate position (a map of arrows). See Fig. 1.2 for illustration (two-dimensional).

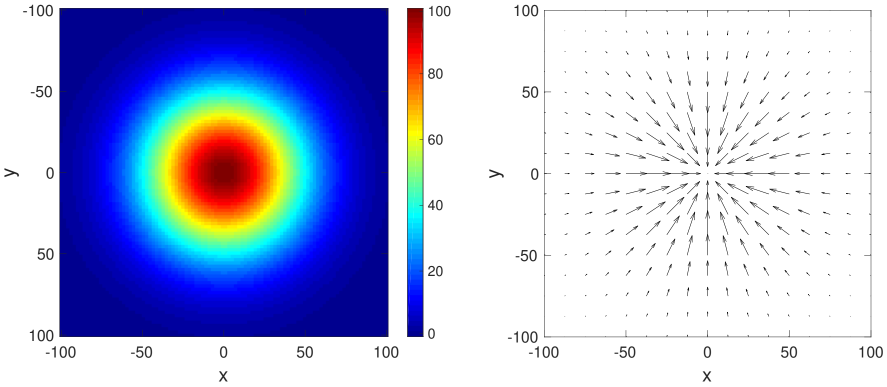

Figure 1.2: Left: A 2D scalar field $F(x, y)$ (e.g. height map of a mountain). Right: gradient of the scalar field ∇F, which is a vector field (map of *slope*).

1.5.2 Divergence

Divergence of a vector field (∇ operation on a vector field through dot product) is

$$\nabla \cdot \mathbf{A} \equiv \left(\frac{\partial}{\partial x}\hat{\boldsymbol{x}} + \frac{\partial}{\partial y}\hat{\boldsymbol{y}} + \frac{\partial}{\partial z}\hat{\boldsymbol{z}}\right) \cdot (A_x\hat{\boldsymbol{x}} + A_y\hat{\boldsymbol{y}} + A_z\hat{\boldsymbol{z}}) \tag{1.32}$$

$$= \frac{\partial A_x}{\partial x} + \frac{\partial A_y}{\partial y} + \frac{\partial A_z}{\partial z}. \tag{1.33}$$

By definition, divergence of a vector field derives net outward flux of the vector field per unit volume at each spatial point. One can correlate outward field flux at a certain location to presence of some sources creating the field (e.g. electric field owing to presence of electrical charges).

Once one has the above understanding, *divergence theorem* comes naturally, which is

$$\int_V \nabla \cdot \mathbf{A} dv = \oint_S \mathbf{A} \cdot d\mathbf{s}. \tag{1.34}$$

It describes that the volume integral of the divergence of a vector field is equivalent to integral of outward surface-normal flux over the surface enclosing the volume. Divergence theorem converts a volume integral to a surface integral or vice versa.

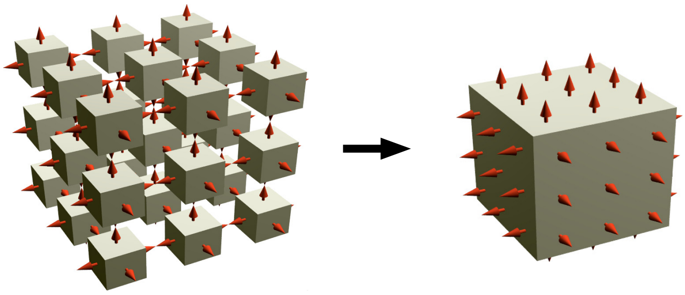

Figure 1.3: Divergence theorem - volume integral (exploded view) becomes surface integral.

1.5.3 Curl

Curl of a vector field (∇ operation on a vector field through cross product) is

$$
\begin{aligned}
\nabla \times \mathbf{A} &\equiv \left(\frac{\partial}{\partial x}\hat{\boldsymbol{x}} + \frac{\partial}{\partial y}\hat{\boldsymbol{y}} + \frac{\partial}{\partial z}\hat{\boldsymbol{z}}\right) \times (A_x\hat{\boldsymbol{x}} + A_y\hat{\boldsymbol{y}} + A_z\hat{\boldsymbol{z}}) && (1.35)\\
&= \begin{vmatrix} \hat{\boldsymbol{x}} & \hat{\boldsymbol{y}} & \hat{\boldsymbol{z}} \\ \frac{\partial}{\partial x} & \frac{\partial}{\partial y} & \frac{\partial}{\partial z} \\ A_x & A_y & A_z \end{vmatrix} && (1.36)\\
&= \left(\frac{\partial A_z}{\partial y} - \frac{\partial A_y}{\partial z}\right)\hat{\boldsymbol{x}} + \left(\frac{\partial A_x}{\partial z} - \frac{\partial A_z}{\partial x}\right)\hat{\boldsymbol{y}} + \left(\frac{\partial A_y}{\partial x} - \frac{\partial A_x}{\partial y}\right)\hat{\boldsymbol{z}}. && (1.37)
\end{aligned}
$$

By definition, curl of a vector field finds the maximum degree of circulation per unit area of the field, and at the same time orientation of the maximum circulation field. The orientation is denoted by the normal direction of the surface containing the circulation. We usually choose the positive surface-normal direction through the right-hand rule — right-hand fingers curl around a surface element's boundary in anti-clockwise direction, and the thumb will be pointing to the positive surface-normal direction. One can correlate circulating field to (vortex) sources that generate such a field. One concrete example is that a steady line current can generate circulating magnetic field around the line.

Stoke's theorem follows naturally from definition of curl. It reads

$$\int_S (\nabla \times \mathbf{A}) \cdot d\mathbf{S} = \oint_C \mathbf{A} \cdot d\mathbf{l}. \tag{1.38}$$

The theorem states that surface integral of curl of a vector field over any continuous open surface is equal to integral of the field along the outer contour of the surface. Stoke's theorem converts a surface integral to a line integral or vice versa. Figure 1.4 illustrates principle of Stoke's theorem.

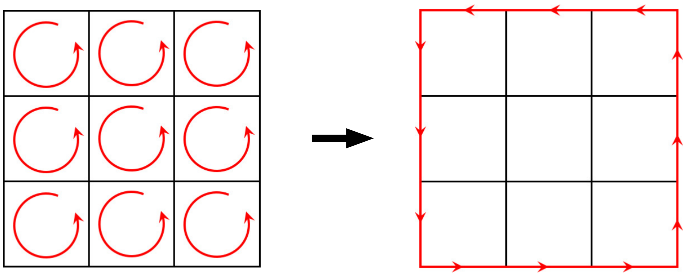

Figure 1.4: Stoke's theorem - Surface integral becomes line integral.

1.5.4 *del* operations in other coordinates

The *del* operator takes different forms in different coordinate systems. The expressions are all relatively simple for orthogonal coordinate systems. What the *del* operator does is to get spatial differentiations along three coordinate directions. One just needs to update the differential length unit when taking differentiations through considering the metric coefficients (h_1, h_2, h_3). If one expresses a general orthogonal coordinate system as (u_1, u_2, u_3) with unit vectors along coordinate axes $\hat{\boldsymbol{a}}_1$, $\hat{\boldsymbol{a}}_2$, and $\hat{\boldsymbol{a}}_3$, respectively, the general expression for ∇ is

$$\nabla \equiv \frac{\partial}{h_1 \partial u_1}\hat{\boldsymbol{a}}_1 + \frac{\partial}{h_2 \partial u_2}\hat{\boldsymbol{a}}_2 + \frac{\partial}{h_3 \partial u_3}\hat{\boldsymbol{a}}_3. \tag{1.39}$$

Divergence operation is generally defined as

$$\nabla \cdot \mathbf{A} = \frac{1}{h_1 h_2 h_3}\left[\frac{\partial(h_2 h_3 A_1)}{\partial u_1} + \frac{\partial(h_1 h_3 A_2)}{\partial u_2} + \frac{\partial(h_1 h_2 A_3)}{\partial u_3}\right]. \tag{1.40}$$

Curl operation is generally

$$\nabla \times \mathbf{A} = \frac{1}{h_1 h_2 h_3}\begin{vmatrix} h_1\hat{\boldsymbol{a}}_1 & h_2\hat{\boldsymbol{a}}_2 & h_3\hat{\boldsymbol{a}}_3 \\ \frac{\partial}{\partial u_1} & \frac{\partial}{\partial u_2} & \frac{\partial}{\partial u_3} \\ h_1 A_1 & h_2 A_2 & h_3 A_3 \end{vmatrix}. \tag{1.41}$$

Therefore, for a cylindrical coordinate system, we have the following for the *del* operations

$$\nabla V = \frac{\partial V}{\partial r}\hat{\boldsymbol{r}} + \frac{\partial V}{r\partial \phi}\hat{\boldsymbol{\phi}} + \frac{\partial V}{\partial z}\hat{\boldsymbol{z}}, \tag{1.42}$$

$$\nabla \cdot \mathbf{A} = \frac{1}{r}\frac{\partial(rA_r)}{\partial r} + \frac{\partial A_\phi}{r\partial\phi} + \frac{\partial A_z}{\partial z}, \tag{1.43}$$

$$\begin{aligned} \nabla \times \mathbf{A} &= \frac{1}{r}\begin{vmatrix} \hat{\boldsymbol{r}} & r\hat{\boldsymbol{\phi}} & \hat{\boldsymbol{z}} \\ \frac{\partial}{\partial r} & \frac{\partial}{\partial \phi} & \frac{\partial}{\partial z} \\ A_r & rA_\phi & A_z \end{vmatrix} \qquad (1.44) \\ &= \left(\frac{\partial A_z}{r\partial\phi} - \frac{\partial A_\phi}{\partial z}\right)\hat{\boldsymbol{r}} + \left(\frac{\partial A_r}{\partial z} - \frac{\partial A_z}{\partial r}\right)\hat{\boldsymbol{\phi}} + \frac{1}{r}\left[\frac{\partial(rA_\phi)}{\partial r} - \frac{\partial A_r}{\partial\phi}\right]\hat{\boldsymbol{z}}. \end{aligned}$$

In a spherical coordinate system (r, θ, ϕ), one has

$$\nabla V = \frac{\partial V}{\partial r}\hat{\boldsymbol{r}} + \frac{\partial V}{r\partial\theta}\hat{\boldsymbol{\theta}} + \frac{1}{r\sin\theta}\frac{\partial V}{\partial\phi}\hat{\boldsymbol{\phi}}, \tag{1.45}$$

$$\nabla \cdot \mathbf{A} = \frac{1}{r^2}\frac{\partial(r^2 A_r)}{\partial r} + \frac{1}{r\sin\theta}\frac{\partial(A_\theta\sin\theta)}{\partial\theta} + \frac{1}{r\sin\theta}\frac{\partial A_\phi}{\partial\phi}, \tag{1.46}$$

$$\begin{aligned} \nabla \times \mathbf{A} &= \frac{1}{r^2\sin\theta}\begin{vmatrix} \hat{\boldsymbol{r}} & r\hat{\boldsymbol{\theta}} & r\sin\theta\hat{\boldsymbol{\phi}} \\ \frac{\partial}{\partial r} & \frac{\partial}{\partial\theta} & \frac{\partial}{\partial\phi} \\ A_r & rA_\theta & r\sin\theta A_\phi \end{vmatrix} \qquad (1.47) \\ &= \frac{1}{r\sin\theta}\left[\frac{\partial\left(A_\phi\sin\theta\right)}{\partial\theta} - \frac{\partial A_\theta}{\partial\phi}\right]\hat{\boldsymbol{r}} \\ &+ \frac{1}{r}\left[\frac{1}{\sin\theta}\frac{\partial A_r}{\partial\phi} - \frac{\partial(rA_\phi)}{\partial r}\right]\hat{\boldsymbol{\theta}} \\ &+ \frac{1}{r}\left[\frac{\partial(rA_\theta)}{\partial r} - \frac{\partial A_r}{\partial\theta}\right]\hat{\boldsymbol{\phi}}. \end{aligned}$$

1.6 Vector identities

There are several proven formulas which one can readily use in vector analysis. Two most used ones are related to double *del* operations.

The first identity says curl of gradient of any scalar field is zero. That is,

$$\nabla \times (\nabla V) \equiv 0. \tag{1.48}$$

This formula can be proven through Stoke's theorem. The relation leads to two somewhat connected comments: (1) gradient of a scalar field is a curl-free vector field; (2) if a vector field is curl-free, it can be written as a gradient of a scalar field.

The second identity says that divergence of curl of any vector field is zero. That is

$$\nabla \cdot (\nabla \times \mathbf{A}) \equiv 0. \tag{1.49}$$

This identity can be proven by utilizing both divergence and Stoke's theorems. It implies: a divergenceless vector field can always be treated as curl of another vector field.

* * *

1.7 Wave

Waves are often associated with period motions, not only in space but also in time. Here we take one-dimensional mechanical wave as an example to illustrate how to mathematically express a wave and also to introduce a few most important wave properties. The concepts to be developed here are very similar to those associated with electromagnetic wave in Chapter 6. *Interference* of mechanical waves will also be discussed here, which is helpful for understanding similar phenomena for electromagnetic waves in Chapter 7.

1.7.1 Wave equation and general solution

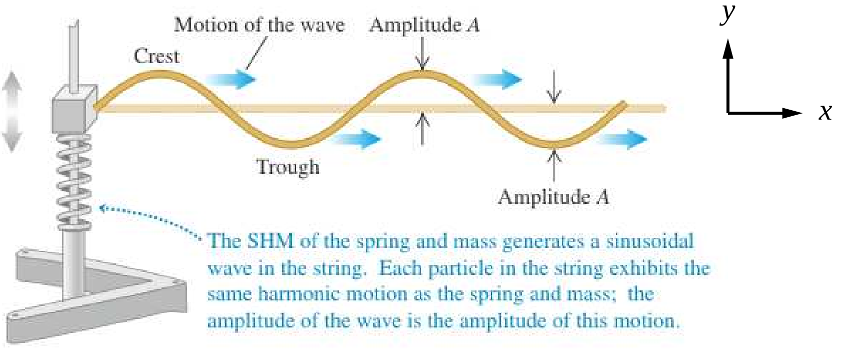
Motion of the wave
Amplitude A
Crest
Trough
Amplitude A
The SHM of the spring and mass generates a sinusoidal wave in the string. Each particle in the string exhibits the same harmonic motion as the spring and mass; the amplitude of the wave is the amplitude of this motion.
y
x

Figure 1.5: A string is excited by periodic motion at its left end to sustain a rightward-propagating wave. "SHM" stands for simple harmonic motion. [Picture taken from *University Physics with Modern Physics*, Global Edition, Pearson Education Limited, 2016].

One mechanical wave example is shown in Fig. 1.5. While the wave is moving in x direction, each particle on the string is moving in the *transverse* y direction. Such wave is called *transverse* wave. *Longitudinal* wave, such as sound wave, has particle-oscillation direction along wave-propagation direction. Importantly, one should take note that a wave does not really transport matter, but rather *disturbance*. Disturbance carries energy.

Let's say the string has mass per unit length μ and sustains a tension F. By setting up an equation of motion for an infinitely small section of the string, one can derive the following equation governing displacement y of the string (coordinate as shown in Fig. 1.5)

$$\frac{\partial^2 y}{\partial t^2} = v^2 \frac{\partial^2 y}{\partial x^2}. \tag{1.50}$$

Displacement y has dependence on both coordinate x and time t. v is a system-dependent value as $v = \sqrt{F/\mu}$, where F is tension in string and μ is string's mass per unit length. For a short while we will know physical meaning of v. Equation 1.50 is the classic *wave equation*, whose general solution[2] is

$$y(x,t) = F(-x+vt) + G(x+vt). \tag{1.51}$$

$F(-x+vt)$ and $G(x+vt)$ are functions with arbitrary profiles, traveling respectively in $+x$ and $-x$ directions with *velocity* v. Why is v wave velocity? One can easily verify in this way: at $t=0$, one has $F(-x+vt) = F(-x)$; at a later time $t = \Delta t$, one has $F(-x+vt) = F(-x+v\Delta t) = F[-(x-v\Delta t)]$. The wave profile F hence has shifted along *positive* x *direction* by $\Delta x = v\Delta t$, therefore the velocity $\Delta x/\Delta t = v$.

Most often, one is interested in only time-harmonic wave solutions with $\cos(\omega t)$ dependence, i.e. wave forms go back to their original after every period of $T = 2\pi/\omega$. ω is a positive number decided by excitation (Fig.1.5). Then, the general solution becomes

$$y(x,t) = A\cos\left[\frac{\omega}{v}(-x+vt)\right] + B\cos\left[\frac{\omega}{v}(x+vt)\right]. \tag{1.52}$$

Coefficients A and B are amplitudes of the cosine waves traveling in $+x$ and $-x$ directions, respectively. Depending on particular problem, A or B may be zero (like in Fig. 1.5, $B=0$).

Some definitions associated with such a harmonic wave are summarized as follows:

- Angular frequency: ω (unit radian/second)
- Frequency: $f = 1/T$ (unit hertz)
- Phase: $\phi = \frac{\omega}{v}(-x+vt)$ for $+x$-travelling term (unit radian)
- Amplitude: A (unit meter in this example)
- Phase velocity: v
- Period: $T = 2\pi/\omega$ (unit second)
- Wavelength: $\lambda = vT$ (unit meter)
- Wavenumber: $k = \omega/v = 2\pi/\lambda$ (unit radian/meter)

The wavenumber k characterizes how much phase variation occurs within a unit propagation length. It is also sometimes referred to as spatial frequency. Comparatively, ω, denoting how many periods occur in a unit time, can be considered as temporal frequency. As k is so often used, the general harmonic solution is usually expressed as

$$y(x,t) = A\cos(-kx+\omega t) + B\cos(kx+\omega t). \tag{1.53}$$

[2] In some references, the general solution is expressed as $y(x,t) = F(x-vt) + G(x+vt)$. This is just a matter of convention. Further elaboration will be found in the following "Phasor" section.

Example: Mechanical wave

A mechanical wave sustained on a x-directed string is expressed in terms of y displacement as $y(x,t) = 0.01\cos(-4.833x + 2070t)$ m, where x is position along the string in meter and t is time in second. Find out wave amplitude, frequency, wavelength, phase velocity, phase difference between $x = 0$ m and $x = 325$ cm at any moment, phase difference between $t = 1$ ms and $t = 5$ ms at position $x = 1$ m.

Solution: Amplitude 0.01 m; frequency 330 Hz (E note); wavelength 1.3 m; phase velocity 428 m/s; spatial phase difference $\frac{\pi}{2}$ rad; temporal phase difference 8.28 rad.

1.7.2 Interference

Two waves can be added together, creating locally stronger or weaker oscillation amplitudes. This phenomenon is called *interference*. We continue to use 1D mechanical wave as in Fig. 1.5 for an example. One can send in another $-x$-traveling wave by exciting the right end point. For a fixed point on a string (i.e. fixed coordinate x), if oscillations of two waves at that point are always *in phase*, we end up with a stronger oscillation there. This is called *constructive interference*. If completely *out of phase*, we have weakened oscillation at that point, which is called *destructive interference*. Interference effect is most drastic when two waves have the same frequency (or the same wavelength).

Mathematically, two counter-traveling waves (with the same frequency and amplitude) can be summed as

$$\begin{aligned} y(x,t) &= A_0\cos(-kx+\omega t) + A_0\cos(kx+\omega t) \\ &= 2A_0\cos(kx)\cos(\omega t) \end{aligned} \tag{1.54}$$

The superimposed wave can be considered as a wave with space-dependent amplitude $2A_0\cos(kx)$, oscillating in time with the same angular frequency ω. At certain x coordinates, oscillation amplitude reaches $2A_0$, and at some other x coordinates, it can be zero. The wave is no longer traveling: the strongly oscillating sections will always oscillate strongly (high local power), and standing-still sections will never move (zero local power). Such a wave is referred to as a *standing wave*.

One doesn't always need two excitations to get interference and standing wave. When a wave is reflected from an "obstacle", the reflected wave naturally interfere with the incoming wave, thereby forming a standing wave. For example, a guitar string is fixed on both ends. Wave excited on a string is reflected by end points. Standing waves can therefore form after interference (only at certain frequencies or tones). Figure 1.6 shows precisely shapes of such standing-wave oscillations.

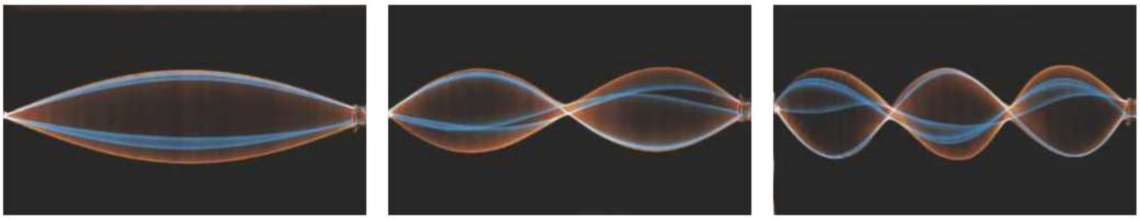

Figure 1.6: Long-time exposure of an oscillating string, with two end points almost fixed. [Picture taken from *University Physics with Modern Physics*, Global Edition, 2016].

The above discussion was limited to 1D wave. One can extend the discussion to waves in 2D or 3D space, which will be more mathematically involved. Physics however is the same - constructive and destructive interferences give rise to high and low energy-density locations. Figure 1.7 shows water wave interference pattern excited by two oscillating point sources, taken from one Youtube video.

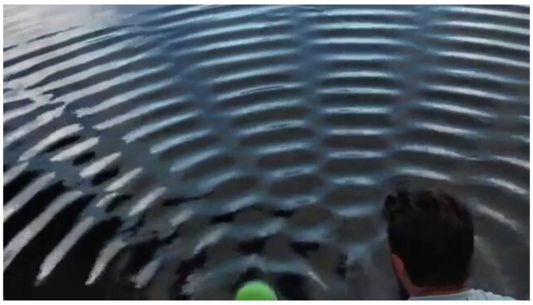

Figure 1.7: Water wave interference pattern excited by two point sources (from Youtube). For more elaborate demonstration, one may watch this particular video.

* * *

1.8 Phasor

Use of "*phasor*" can greatly simplify calculation of physical problems involving time-harmonic functions. The phasor expression for $\cos\phi$ is a complex-valued function

$$\exp(i\phi) \equiv e^{i\phi} = \cos\phi + i\sin\phi. \tag{1.55}$$

One uses phasor form for mathematical operations and takes real part of the final result as the physical solution. It can be easily verified that complex conjugate of a phasor $A = \exp(i\phi)$ is $A^* = \exp(-i\phi)$.

The corresponding phasor expression of the general solution in Eq. 1.53 is

$$y(x,t) = A\exp[i(-kx+\omega t)] + B\exp[i(kx+\omega t)]. \tag{1.56}$$

It is actually more straightforward to obtain this phasor-form solution from the wave equation 1.50, as compared to directly deriving the solution in Eq. 1.53. The reason is that arithmetics with phasors (exponential functions) is rather simple: differentiation and multiplication of phasors lead to also phasors. We show the process as follows.

Since we are interested in time-harmonic solutions, we know the phasor solutions will have the form

$$y(x,t) = y(x)\exp(i\omega t). \tag{1.57}$$

Note here we use the convention of positive time dependence $\exp(i\omega t)$ instead of negative time dependence $\exp(-i\omega t)$. By substituting Eq. 1.57 into Eq. 1.50 one has

$$-\omega^2 y = v^2\frac{d^2y}{dx^2}, \quad \rightarrow \quad \frac{d^2y}{dx^2} + k^2 y = 0. \tag{1.58}$$

Here y has only dependence on x. The general solution of this second-order differential equation is

$$y(x) = A\exp(-ikx) + B\exp(ikx). \tag{1.59}$$

Appending the time-harmonic dependence, one has

$$\begin{aligned} y(x,t) &= [A\exp(-ikx) + B\exp(ikx)]\exp(i\omega t) \\ &= A\exp(-ikx+i\omega t) + B\exp(ikx+i\omega t) \\ &= A\exp[i(-kx+\omega t)] + B\exp[i(kx+\omega t)]. \end{aligned} \tag{1.60}$$

Use of negative time dependence $\exp(-i\omega t)$ will lead to what is concerned in footnote 2.

* * *

1.9 Torque

Torque is a vector quantity that, when applied to an object, quantifies the object's tendency to rotate around a certain axis. Refer to Fig. 1.8. Torque is calculated as

$$\boldsymbol{\tau} = \mathbf{r} \times \mathbf{F}. \tag{1.61}$$

r is vector from the rotation axis to the point of force application, and **F** is the applied force.

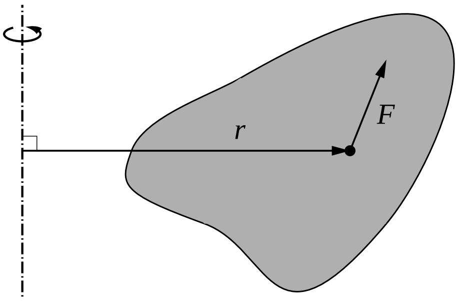

Figure 1.8: Torque due to a force on an object.

Chapter 2

Electrostatics

The key elements in electrostatics are: *charge*, *electric field*, and electrostatic *force*. Charge (scalar) produces electric field (vector); electric field exerts force, without physical contact, on other charges. It is only through existence of electrostatic force we came to know there is electric field. Electric *potential* (scalar), which is more widely known as *voltage*, is a derived property based on electric field. Electric potential can facilitate easier calculation of work done by an electric field on a moving charge. Existence of an electric field is owing to aggregation of charges of the same polarity; close packing of the same type of charges needs work input. Equivalently speaking, an electric field has an *energy*. Electric field in a conducting medium induces charge flow, or *current density* (vector). In a metal wire, charge flow has well-defined direction and a conserved amplitude. One therefore cares about just the integrated amplitude on wire cross-section, which is *current*. For a conducting object, given an applied voltage across two contacts, the ratio between voltage and current has a fixed value, which we name as *resistance*. Circuit theory is rather a subset of electrostatic theory applied to connected metal blocks, mostly wires. For non-conducting medium, an electric field displaces charges in the medium, causing material polarization (polarization field). The strength of the polarization field is related to a material constant called *permittivity*, which is tabulated through experiments. In such medium, one can sum up the applied electric field and the polarization field, and refer to the sum as *displacement field*. Displacement field is auxiliary, but can simplify many derivations. Two governing principles in electrostatics are: Gauss's law and conservative nature of electric field.

2.1 Electric charges

An electric charge can be positive or negative. The basic charges are that carried by an electron (negative) and that by a proton (positive). Electric charges exist ubiquitously but they can be hard to notice owing to perfect mixing of charges of two polarities. An object becomes charged once there is an imbalance between positive and negative charges. Two charged bodies exert to each other a pulling or pushing force, sometimes leading to a bit more scary phenomena such as sparks and lightning. Historically, the attractive force exhibited by amber for example was discovered 2000 years ago. Systematic studies came during 1600s. In early days, people thought electric charges are "fluids" inside materials. In year 1733, Frenchman "Charles François de Cisternay du Fay" concluded that a glass rod rubbed with silk is left with one type of "fluid" (positive charge, as recognized later), and amber rubbed with fur is left with "amber-type" fluid (negative charge).

The unit for charge is Coulomb (C). One electron (discovered in 1897) carries charge $-e$, where $e = 1.602176634 \times 10^{-19}$ C. One proton has charge e. Charges can't be created from nowhere, nor can they be annihilated — they are conserved.

?

Charges of two polarities are naturally in perfect balance, i.e. neutralized. Therefore, we don't usually feel charges. What are typical methods to separate charges?

2.2 Coulomb's law - field by point charge

It was known that an electric charge experiences a force when placed close to another charge. "Action at a distance" is unexplainable without introduction of a new physical quantity called "field" — it is not a substance but can carry and transfer power. The force experienced by a test charge q_t in such a field can be expressed as

$$\mathbf{F} = q_t \mathbf{E}, \tag{2.1}$$

where $\mathbf{E}$ is the invisible *electric field* (unit per definition: newton per coulomb or N/C; SI unit: volt per meter or V/m). The linear dependence on q_t is easily verifiable in experiment. Around year 1785 Frenchman Charles-Augustin de Coulomb (1736-1806) measured the force as a function of the test-charge position for a field created by a point charge q. It turns out that, in a free-space (vacuum or air), electric field generated by a point charge at origin O has the following form

$$\mathbf{E} = \frac{q}{4\pi\epsilon_0 r^2}\hat{\boldsymbol{r}} \quad \text{or} \quad \mathbf{E} = \frac{q\mathbf{r}}{4\pi\epsilon_0 r^3}, \qquad \text{(unit: volt/meter, V/m)} \tag{2.2}$$

where q is charge amount, ϵ_0 (8.854×10^{-12} F/m, or $\frac{\mathrm{C}}{\mathrm{V{\cdot}m}}$) is permittivity of free space, r is distance between observation point (or sometimes called field point, or probing point), say P, to source at origin O, and $\hat{\boldsymbol{r}}$ is a unit vector of $O \to P$ vector $\mathbf{r}$. Figure 2.1 shows graphically vector electric field around a point charge. Note that field amplitude decays radially at a rate of $1/r^2$. It is also common that one shows only field direction with streamlines, as in Fig. 2.2. Field by a single point charge is the basis for calculating field induced by many charges.

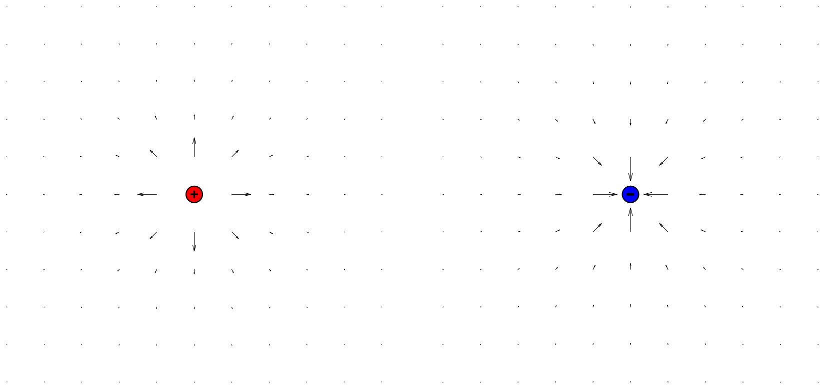

Figure 2.1: Electric field by a point charge. Left: positive charge. Right: negative charge.

Coulomb's law (Eq. 2.1 in full) says that force on a test point charge q_2 in field created by point charge q_1 is

$$\mathbf{F}_{12} = q_2\mathbf{E}_{12} = \frac{q_1 q_2}{4\pi\epsilon_0 r_{12}^2}\hat{\boldsymbol{r}}_{12}. \tag{2.3}$$

Here, $\mathbf{E}_{12}$ is electric field due to q_1 examined at q_2 position, r_{12} and $\hat{\mathbf{r}}_{12}$ are length and direction of the $q_1 \to q_2$ vector. Two charges with the same polarity repel each other, while two charges with opposite polarities attract each other.

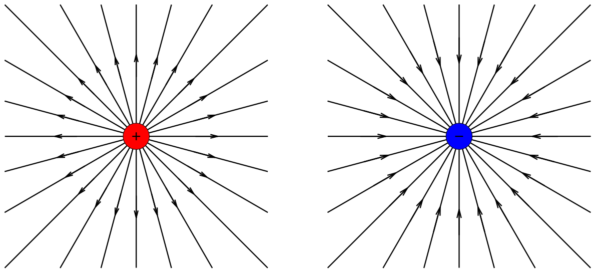

Figure 2.2: Electric field lines by a point charge.

2.3 Field by charge distribution

If there exist more than one point charge, the overall electric field is summation of fields generated by individual point charges. This is basically called *superposition principle*. Mathematically it is

$$\boxed{\mathbf{E} = \frac{1}{4\pi\epsilon_0}\sum_{k=1}^{n}\frac{q_k(\mathbf{r}-\mathbf{r}_k)}{|\mathbf{r}-\mathbf{r}_k|^3}.} \tag{2.4}$$

Here $\mathbf{r}$ and $\mathbf{r}_k$ correspond to position vectors of the observation point and the k^{th} point charge, respectively.

If the source is a continuous distribution of charges, one resorts to integration instead of summation. Charge density ρ should then be used rather than discrete charge values. In the most general case, charge density is described as charge per unit volume (ρ_v). In certain cases, charge is carried by a thin sheet or a wire, ρ can then be described as charge per unit area (ρ_s) or length (ρ_l), respectively. Electric field is calculated as

$$\boxed{\mathbf{E} = \frac{1}{4\pi\epsilon_0}\int_V \frac{\rho_v \mathbf{r}_{\text{SP}}}{r_{\text{SP}}^3}dv, \quad \text{or} \quad \frac{1}{4\pi\epsilon_0}\int_S \frac{\rho_s \mathbf{r}_{\text{SP}}}{r_{\text{SP}}^3}ds, \quad \text{or} \quad \frac{1}{4\pi\epsilon_0}\int_C \frac{\rho_l \mathbf{r}_{\text{SP}}}{r_{\text{SP}}^3}dl,} \tag{2.5}$$

where $\mathbf{r}_{\text{SP}}$ is a vector pointing from source position to observation point.

2.4 Common field profiles

A single point charge is also referred to as electric *monopole*, whose electric field was presented in Section 2.2. Another basic field profile is that induced by two charges of the same quantity but with opposite polarities, *i.e.* an electric *dipole*. According Eq. 2.4 (i.e. adding two monopole fields), one obtains dipole field as shown in Fig. 2.3. Notice that the electric field lines originate from the positive charge and terminate at the negative charge. This is a general property of electric field lines. Positive charges serve as sources for electric field, whilst negative ones serve as sinks. “Sink” can in fact be interpreted as source, except generating field lines with reversed directions.

An electric dipole possesses an *electric dipole moment* (vector), defined as

$$\boxed{\mathbf{p} = q\mathbf{d}, \qquad (\text{unit: C}\cdot\text{m})} \tag{2.6}$$

where q is charge quantity carried by each point charge, and $\mathbf{d}$ is vector pointing from the negative charge to the positive charge. Under an external $\mathbf{E}$ field, an electric dipole tends to rotate with torque

$$\boxed{\boldsymbol{\tau} = \mathbf{p}\times\mathbf{E}.} \tag{2.7}$$

The stable position after rotation is that the dipole moment aligns with $\mathbf{E}$.

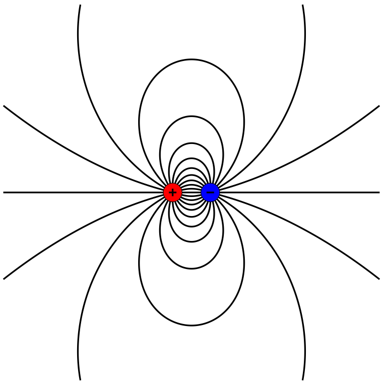

Figure 2.3: Electric field lines by an electric dipole.

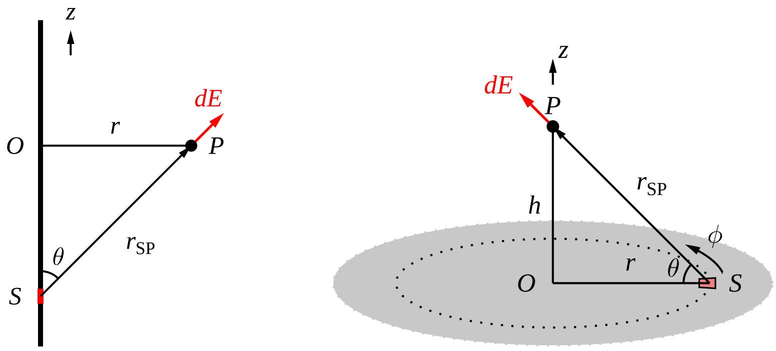

Figure 2.4: Left: geometric parameters for calculating electric field by a line charge. Right: geometric parameters for calculating electric field by a plane charge.

Electric field by a line charge

Derive **E** field by an infinite line charge with charge density ρ_l (Fig. 2.4, left panel).

Solution: One can first argue from symmetry perspective to simply calculation. Refer to the figure. First, cylindrical symmetry suggests that field vector at any point should lie on rz plane defined by the point and the charged line. Second, at any probing point, electric-field contributions from two semi-infinite line charges at two sides will cancel out in their z components. Conclusion: electric field has only r component. Contribution by an infinitesimal line section is

$$d\mathbf{E} = \frac{\rho_l dz}{4\pi\epsilon_0 r_{\mathrm{SP}}^2}\hat{\boldsymbol{r}}_{\mathrm{SP}}.$$

From the symmetry argument, we count only the r-component

$$dE_r = |d\mathbf{E}|\sin\theta = \frac{\rho_l dz}{4\pi\epsilon_0 r_{\mathrm{SP}}^2}\sin\theta.$$

Take integration. One has

$$\begin{aligned}
E_r &= \int_C dE_r = \int_C \frac{\rho_l dz}{4\pi\epsilon_0 r_{\mathrm{SP}}^2}\sin\theta = \frac{\rho_l}{4\pi\epsilon_0}\int_C \frac{1}{r_{\mathrm{SP}}^2}\sin\theta dz \\
&= \frac{\rho_l}{4\pi\epsilon_0}\int_{\theta=0}^{\pi} \frac{\sin^2\theta}{r^2}\sin\theta \frac{r}{\sin^2\theta}d\theta \qquad \leftarrow \left[\frac{-z}{r} = \cot\theta \;\rightarrow\; dz = \frac{r}{\sin^2\theta}d\theta\right] \\
&= \frac{\rho_l}{4\pi r\epsilon_0}\int_{\theta=0}^{\pi}\sin\theta d\theta = \frac{\rho_l}{4\pi r\epsilon_0}\left[-\cos\theta\right]_0^{\pi} = \frac{\rho_l}{2\pi r\epsilon_0}.
\end{aligned}$$

For an infinite line charge, electric field amplitude decays radially at a rate of $1/r$.

Electric field by a plane charge

Derive **E** field by an infinite plane charge with uniform surface charge density ρ_s (Fig. 2.4, right panel).

Solution: The plane can be viewed of circular shape with infinite radius. A simple symmetry consideration leads to that, at any probing point, electric field will be along the direction normal to the charge plane. The plane-parallel field component is canceled out owing to rotational symmetry. Refer to the figure. For space above the charge plane, contribution by an infinitesimal area section at S is

$$d\mathbf{E} = \frac{\rho_s ds}{4\pi\epsilon_0 r_{\mathrm{SP}}^2}\hat{\boldsymbol{r}}_{\mathrm{SP}}.$$

The z-component is

$$dE_z = |d\mathbf{E}|\sin\theta = \frac{\rho_s ds}{4\pi\epsilon_0 r_{\mathrm{SP}}^2}\sin\theta.$$

Integrating over the charge area (equivalently over angles θ and ϕ), one has

$$\begin{aligned}
E_z = \int_S dE_z &= \int_S \frac{\rho_s ds}{4\pi\epsilon_0 r_{\mathrm{SP}}^2}\sin\theta = \frac{\rho_s}{4\pi\epsilon_0}\int_S \frac{1}{r_{\mathrm{SP}}^2}\sin\theta ds \\
&= \frac{\rho_s}{4\pi\epsilon_0}\int_S \frac{1}{r_{\mathrm{SP}}^2}\sin\theta(r d\phi dr) \qquad \leftarrow \left[r = h\frac{\cos\theta}{\sin\theta}; dr = -h\frac{d\theta}{\sin^2\theta}\right] \\
&= \frac{\rho_s}{4\pi\epsilon_0}\int_{\theta=\frac{\pi}{2}}^{0}\int_{\phi=0}^{2\pi}\frac{\sin^2\theta}{h^2}\sin\theta\left[-h^2\frac{\cos\theta}{\sin^3\theta}d\theta d\phi\right] \\
&= \frac{\rho_s}{4\pi\epsilon_0}\int_{\theta=\frac{\pi}{2}}^{0}\int_{\phi=0}^{2\pi}(-\cos\theta)d\phi d\theta = \frac{\rho_s}{4\pi\epsilon_0}\int_{\theta=\frac{\pi}{2}}^{0}(-\cos\theta)d\theta\int_{\phi=0}^{2\pi}d\phi \\
&= \frac{\rho_s}{4\pi\epsilon_0}[-\sin\theta]_{\frac{\pi}{2}}^{0}[\phi]_0^{2\pi} = \frac{\rho_s}{4\pi\epsilon_0}\cdot 1\cdot 2\pi = \frac{\rho_s}{2\epsilon_0}.
\end{aligned}$$

Electric field is uniform in space! For space below the charge plane, one has the same field amplitude but with direction reversed.

Alternative method:
One starts with alternative expression for field by a differential area element as

$$d\mathbf{E} = \frac{\rho_s ds\ \mathbf{r}_{\mathrm{SP}}}{4\pi\epsilon_0 r_{\mathrm{SP}}^3} = \frac{\rho_s ds(h\hat{\boldsymbol{z}} - r\hat{\boldsymbol{r}})}{4\pi\epsilon_0(h^2+r^2)^{\frac{3}{2}}}. \qquad \leftarrow [\mathbf{r}_{\mathrm{SP}} = h\hat{\boldsymbol{z}} - r\hat{\boldsymbol{r}}]$$

From symmetry, one can just integrate z-component. Note also $ds = rdrd\phi$ Hence

$$\begin{aligned}
E_z = \int_S dE_z &= \int_S \frac{\rho_s hr}{4\pi\epsilon_0(h^2+r^2)^{\frac{3}{2}}}d\phi\ dr \\
&= \frac{\rho_s h}{4\pi\epsilon_0}\int_{\phi=0}^{2\pi}d\phi\int_{r=0}^{\infty}\frac{r}{(h^2+r^2)^{\frac{3}{2}}}dr \\
&= \frac{\rho_s h}{4\pi\epsilon_0}2\pi\int_{u=h^2}^{\infty}\frac{1}{u^{\frac{3}{2}}}du \qquad \leftarrow \left[\text{let } u = h^2+r^2,\ \text{hence } rdr = \frac{1}{2}du\right] \\
&= -\frac{\rho_s h}{2\epsilon_0}\left.\frac{1}{u^{\frac{1}{2}}}\right|_{u=h^2}^{\infty} = \frac{\rho_s}{2\epsilon_0}.
\end{aligned}$$

Electric field by two parallel plane charges

Derive $\mathbf{E}$ field by two infinite parallel planes, one with a positive surface charge density ρ_s and the other $-\rho_s$ (an ideal two-plate capacitor).

Solution: With result in the previous example in mind, the positively charged plane has uniform $\mathbf{E}$ lines directed outwards, and the negatively charged plane has uniform $\mathbf{E}$ lines directed inwards. Superposition of the two fields results in cancellation of fields at spaces outside the two parallel plates and doubling of fields in space enclosed by the two plates. If the bottom plate is positively charged, one has field between two plates: $\mathbf{E} = \frac{\rho_s}{\epsilon_0}\hat{\boldsymbol{z}}$.

2.5 Gauss's law

Carl Friedrich Gauss (1777-1855), a German mathematician, formulated Gauss's law in 1813. It states that the total *electric flux* out of a volume is proportional to total charge Q contained in the volume. It is straightforwardly expressed in integral form as

$$\boxed{\oint_S \mathbf{E}\cdot d\mathbf{s} = \frac{Q}{\epsilon_0}. \quad \text{(Gauss's law, electric, free space)}} \tag{2.8}$$

Here the integral takes place on surface enclosing the volume. Note that the integration surface is a fictitious surface, not necessarily a physical one. The integration surface is referred to as *Gaussian surface.*

The LHS of Eq. 2.8 can be converted into a volume integral by using divergence theorem, as $\oint_S \mathbf{E}\cdot d\mathbf{s} = \int_V \nabla\cdot\mathbf{E}dv$. The RHS of Eq. 2.8 can also be written in volume integral as $\frac{Q}{\epsilon_0} = \int_V \frac{\rho_v}{\epsilon_0}dv$. By equating the kernels, we have Gauss's law in differential form

$$\nabla\cdot\mathbf{E} = \frac{\rho_v}{\epsilon_0}. \tag{2.9}$$

Gauss's law connects electric field and charges seamlessly through very compact formulation. It serves as one fundamental equation in the later on Maxwell's equations. The integral form, Eq. 2.8 can be used to gain knowledge of electric field around a charge distribution; the differential form, Eq. 2.9 can be used to derive charge distribution through a given electric field. In the previous section, Coulomb's law (Eq. 2.5) are used for calculating fields based on an electric charge distribution, which is often laborious. When symmetry exists for a charge distribution, one can utilize the integral form of Guass's law Eq. 2.8 to quickly obtain corresponding electric field distributions.

Electric field by point charge, through Gauss's law

Derive electric field around a point charge q.

Solution: One can use a spherical Gaussian surface whose center coincides with the charge point. Through symmetry argument, one knows electric field on any such spherical surface should have the same amplitude, and directs along r direction. For a spherical Gaussian surface with a radius R, from Gauss's law, one has

$$\oint_S \mathbf{E}\cdot d\mathbf{s} = \oint_S E_r ds = E_r\cdot(4\pi r^2) = \frac{q}{\epsilon_0}.$$

It follows that $E_r = \frac{q}{4\pi\epsilon_0 r^2}$, i.e. Eq. 2.2 re-derived.

Electric field by line charge, through Guass's law

Derive $\mathbf{E}$ field of an infinitely long wire with line charge density ρ_l.

Solution: See lecture slides.

Electric field by plane charge, through Guass's law

Derive $\mathbf{E}$ field of an infinitely plane charge with surface charge density ρ_s.

Solution: See lecture slides.

2.6 Electric potential

In gravitation, an object upheld in air tends to drop — the object has a gravitational potential energy. Likewise, a charge placed in an electric field tends to move — the charge has an *electric potential energy.* We define electric potential energy U of a charge q as $U = qV$, where V is called electric potential (unit: volt, with symbol V).

When a test charge q moves *without acceleration* from P_1 to P_2 in field $\mathbf{E}$, work done by an external force against electric force shall be equal to change of potential energy, i.e.

$$-\int_{P_1}^{P_2} q\mathbf{E}\cdot d\mathbf{l} = U_2 - U_1 = (V_2 - V_1)q. \tag{2.10}$$

This leads to

$$V_2 - V_1 = -\int_{P_1}^{P_2} \mathbf{E}\cdot d\mathbf{l}. \quad \text{(unit: volt, V)} \tag{2.11}$$

This difference in electric potential between two points is what we commonly know as *voltage*. Usually, we use a point at infinite away as a common reference ($V_\infty = 0\ V$) such that one can calculate absolute potential value at a point. Setting $P_1 = \infty$ in Eq. 2.11, one has potential at P_2 (now referred to as a general point P) as

$$\boxed{V = -\int_{\infty}^{P} \mathbf{E}\cdot d\mathbf{l}.} \tag{2.12}$$

It follows that electric potential at distance R from a point charge is

$$\boxed{V = -\int_{\infty}^{R} \mathbf{E}\cdot(dr\hat{\mathbf{r}}) = -\int_{\infty}^{R} \frac{q}{4\pi\epsilon_0 r^2}\hat{\mathbf{r}}\cdot(dr\hat{\mathbf{r}}) = \frac{q}{4\pi\epsilon_0 R}.} \tag{2.13}$$

By superposition principle, one can calculate electric potential owing to many point charges as

$$\boxed{V = \frac{1}{4\pi\epsilon_0}\sum_{k=1}^{n} \frac{q_k}{|\mathbf{r}-\mathbf{r}_k|}.} \tag{2.14}$$

Here $\mathbf{r}$ and $\mathbf{r}_k$ correspond to position vectors of the observation point and the k^{th} point charge, respectively. If the charges are a continuous volume distribution, then one has

$$V = \frac{1}{4\pi\epsilon_0}\int_C \frac{\rho_l}{r_{\text{SP}}}dl, \quad \text{or} \quad \frac{1}{4\pi\epsilon_0}\int_S \frac{\rho_s}{r_{\text{SP}}}ds, \quad \text{or} \quad \frac{1}{4\pi\epsilon_0}\int_V \frac{\rho_v}{r_{\text{SP}}}dv. \tag{2.15}$$

r_{SP} is length from source element to observation point.

Static electric field is conservative. It means if a test charge moves in an electric field along a closed loop, the work done by the field on the charge is zero. This “conservative” property is mathematically written as

$$\oint_C \mathbf{E} \cdot d\mathbf{l} = 0. \quad \text{(static } \mathbf{E} \text{ is conservative)} \tag{2.16}$$

By applying Stoke’s theorem, one can convert Eq. 2.16 into a differential form

$$\nabla \times \mathbf{E} = 0. \tag{2.17}$$

A conservative field has a direct implication on work done on a test charge q: when the charge moves from point P_1 to another point P_2, the work done by the field is constant regarless of the path undertaken by the charge. We have experienced the same for a mass body in gravitational field. By recalling the vector identity Eq. 1.48, Eq. 2.17 suggests that $\mathbf{E}$ can be expressed as gradient of a scalar field, which is actually electric potential V discussed above. We therefore have

$$\boxed{\mathbf{E} = -\nabla V.} \tag{2.18}$$

A negative sign is added for conforming sign convention – electric field points from high to low potential. A direct consequence is that equipotential surfaces are always normal to electric field lines. Equipotential surfaces together with electric field lines associated with a point charge as well as those for a two-parallel-plate system are shown in Fig. 2.5.

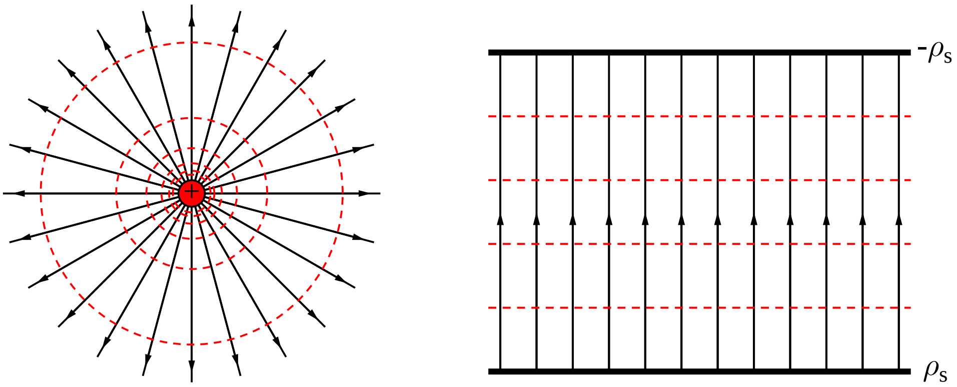

Figure 2.5: Left: Equipotential surfaces (red dashed lines) of a point charge. Right: Equipotential surfaces of parallel planes with opposite charges. Black arrows: $\mathbf{E}$ field.

In general, knowledge about potential can greatly simplify calculation of work done by electric field for a charge moving in an electric field. Once V field is defined, a charge q placed in an electric field has a potential energy $U = qV$. When the charge slowly moves from point P_1 to P_2. Change of potential energy is $\Delta U = U_2 - U_1 = q(V_2 - V_1) = q\Delta V$. The work done by the field on the charge, and that by the external force on the charge, are respectively

$$\boxed{W_{\mathrm{E}} = -\Delta U = -q\Delta V, \quad \text{and} \quad W_{\mathrm{F}} = \Delta U = q\Delta V.} \tag{2.19}$$

Furthermore, potential difference or voltage, which is readily measurable, becomes an important parameter in electric-circuit theory.

?

In electrostatics, a conducting body is always at the same potential. Why?

2.7 Behavior of materials in electric field, D field

So far, our primary focus has been on electric field in *free space*. When an external electric field is imposed on materials, charges within the materials will react and settle into appropriate states, inducing a secondary field. This process will induce change of field both inside and outside material bodies.

2.7.1 Conductors

A special class of materials are electric conductors that have charges freely running in their bodies. Let's focus on an important category of such materials called metals, where free electrons are charge carriers. In normal situation, free electrons in a metal are running around randomly, but in average they appear evenly distributed in the body and balance out the positive charges (nuclei). That is, charge density everywhere is $\rho = 0$. If a metal is placed in an electric field, free electrons (need not to be all of them) will be dragged towards one side and leave the other side positively charged. The excess surface charges (ρ_s) on two sides build up an internal electric field that is opposite to the external field. This process continues until the two fields cancel each other. Therefore, under electrostatic scenario, one has zero net electric field and zero charge density inside a metal body, i.e.

$$\boxed{\mathbf{E} = 0, \quad \rho = 0. \qquad \text{(inside metal)}} \tag{2.20}$$

If electric field is zero everywhere within metal, a metal body is at the same electric potential.

Let's further focus on the excessive charges on metal surface. The excessive surface charges will not be stable if electric field at the surface (inside metal) has a tangential component. Therefore, in a static scenario, there is no tangential electric field *just inside* a metal surface. How about *just outside*? This can be found out by carrying out a line integral of electric field along a fictitious rectangular loop of negligible height h sitting on a metal surface. The schematic is shown in Fig. 2.6 (left panel). The fictitious loop can always be chosen to have a small enough width w to ensure that field will not vary much over the integration path. Static $\mathbf{E}$ is conservative; therefore $\oint_C \mathbf{E} \cdot d\mathbf{l} = 0$. Line integral over vertical line sectors has no net contribution; line integral over the horizontal line sector inside metal is zero owing to absence of field there; hence, we conclude that just outside a metal surface

$$\boxed{E_t = 0. \qquad \text{(outside metal)}} \tag{2.21}$$

This is true even when there exist induced charges at metal surface.

The surface-normal component can be obtained by setting up a fictitious closed cylindrical surface with an infinitesimal height across the metal boundary. See Fig. 2.6 (right panel). According to Gauss's law (Eq. 2.8), surface integral over the closed surface shall be equal to the total charge enclosed by the cylinder. Contribution from the side and the bottom (in metal) can be neglected. One is left with only surface integral on the top surface (in free space). The fictitious cylinder can be chosen small enough such that both field and surface charge do not vary much over the cylinder's cross-section. Therefore, one has

$$\int_S E_n ds = \frac{1}{\epsilon_0} \int_S \rho_s ds. \tag{2.22}$$

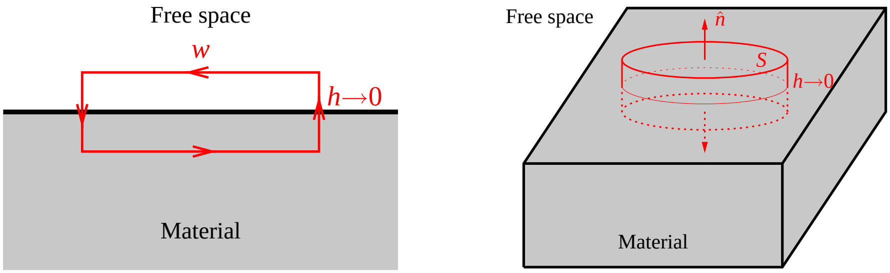

Figure 2.6: Left: Line integral of $\mathbf{E}$ along a fictitious loop on a material interface, to find E_t. Right: Area integral of $\mathbf{E}$ on a fictitious closed cylindrical surface, to find E_n.

Hence, the surface-normal component of electric field just outside a metal surface is

$$\boxed{E_n = \frac{\rho_s}{\epsilon_0}. \qquad \text{(outside metal)}} \tag{2.23}$$

We began our discussion with a neutral metal body placed in an external $\mathbf{E}$ field. The above analyses in fact are still valid for metals loaded with excessive charges. Therefore, the above conclusions can help to understand the electric field of a charged metal object.

Electric field by charged metal sphere

A solid metal sphere with radius R_o carries a charge Q. Derive $\mathbf{E}$.

Solution: The excessive charge carriers tend to push themselves furthest apart, finally settling *evenly* on the sphere's surface. From Eqs. 2.21 and 2.23, we know there exists only radially directed $\mathbf{E}$ field. As a matter of fact, $\mathbf{E}$ outside the sphere is identical to that generated by a point source Q placed at the sphere's center. Therefore in the sphere's coordinate, $\mathbf{E} = \frac{Q}{4\pi\epsilon_0 R^2}\,\hat{\boldsymbol{r}}$ for $R > R_o$ and $\mathbf{E} = 0$ inside the sphere.

Electric field by charged hollow metal sphere

A *hollow* metal sphere with inner and outer radii R_i and R_o carries a charge Q. Derive $\mathbf{E}$.

Solution: In the sphere's coordinate, $\mathbf{E} = \frac{Q}{4\pi\epsilon_0 R^2}\,\hat{\boldsymbol{r}}$ for $R > R_o$, and $\mathbf{E} = 0$ for $R_i < R < R_o$ as well as $R < R_i$.

Neutral hollow metal sphere in electric field

A *hollow* metal sphere with inner and outer radii R_i and R_o carrying no excessive charge is placed in a uniform field $\mathbf{E} = E_0\hat{\boldsymbol{z}}$. Derive $\mathbf{E}$ inside the hollow region.

Solution: Zero.

Charge balance between two metal spheres

A solid metal sphere of radius 4 cm is loaded with 1 nC charge. Another solid metal sphere of radius 1 cm is placed 1 m away. Two spheres are connected with a straight, fine metal wire of negligible radius (which holds no charge). Notice that the distance between two spheres is relatively large compared to the sphere sizes.

One can assume that charges on each sphere is uniformly distributed. Calculate amount of charge on each sphere at equilibrium.

Solution: Refer to Problem 1 in exam 2019.

2.7.2 Dielectrics

Dielectric materials, such as oxides and polymers etc., have no freely moving charges. However, under an external electric field, their bound charges can be displaced slightly, creating dipoles with dipole moments more or less aligned with the applied field direction. One simple example is that electron cloud of an atom can be pulled aside against its immobile nuclei, resulting in a dipole. In addition, many dielectrics have *polar* molecules (e.g. H_2O) due to asymmetric placement of constituent atoms. Each molecule is in fact a dipole, but no net electric field is generated owing to random orientation of molecules. Under excitation by an external field, the dipoles will undergo rotation and therefore become aligned with the field. A dielectric material becomes *polarized* by an external electric field.

Polarized atoms or molecules have their own electric field (Fig. 2.3), leading to a change of total electric field in dielectric media. The polarization-induced field is due to a spatial distribution of dipoles, referred to as *polarization field* **P**. **P** at certain position is defined as vector sum of dipole moments in a unit volume at that position. Instead of calculating electric field, one introduces an auxiliary (not physical) quantity called *displacement field* **D** which lumps both the external electric field and the polarization field, as

$$\mathbf{D} = \epsilon_0 \mathbf{E} + \mathbf{P}. \quad \text{(unit: coulomb per square meter, C/m}^2\text{)} \tag{2.24}$$

In this way one manages (two critical steps skipped) to keep the form of Gauss's law intact as

$$\nabla \cdot \mathbf{D} = \rho. \tag{2.25}$$

This above can be considered as Gauss's law in a more general form, valid for both dielectric media and free space (by setting $\mathbf{P} = 0$). Displacement field is auxiliary in the sense that it is, unlike electric field, not a measurable quantity. In integral form, the generalized Gauss's law reads

$$\oint_S \mathbf{D} \cdot d\mathbf{s} = Q. \quad \text{(Gauss's law, electric)} \tag{2.26}$$

Instead of mentioning **P** for describing a material's response, one can go a step further by noting that **P** for most dielectrics is spatially uniform and proportional to external field **E**, as

$$\mathbf{P} = \epsilon_0 \chi_e \mathbf{E}. \tag{2.27}$$

Here χ_e is called *electric susceptibility*. Therefore, instead of Eq. 2.24, one has

$$\mathbf{D} = \epsilon_0 (1 + \chi_e) \mathbf{E} = \epsilon_0 \epsilon_r \mathbf{E} = \epsilon \mathbf{E}. \tag{2.28}$$

Here, a new material parameter called *relative permittivity* $\epsilon_r = 1 + \chi_e$ is defined to describe material response. Sometimes, one describes a material's response by just permittivity $\epsilon = \epsilon_0 \epsilon_r$. Relative permittivty (like electric susceptibility) has no unit, but ϵ_0 and total permittivity ϵ have (F/m).

2.7.3 Dielectric strength

A dielectric material can not stand for any value of electric field. When external field is high enough, a dielectric medium can be damaged (bound charges are completely torn apart). This process is usually called *dielectric breakdown* or *ionization*. The critical electric field that a dielectric can withstand is called *dielectric strength* of the medium. For air, it is 3 kV/mm, which decreases somewhat when humidity arises.

2.8 Boundary conditions

In many situation, one needs to know boundary conditions regarding how electric field crosses a material interface in order to solve field over a large domain. This can be achieved through utilizing conservative property of **E** field as well as the generalized Gauss's law.

Refer to the left panel in Fig. 2.6. The upper half plane can be replaced with one dielectric medium with permittivity ϵ_1 while the medium below has ϵ_2. We similarly make an integral of electric field around the line loop across the interface, which should be zero according to conservative nature of **E** field. Integrations over the vertical line sections are trivial and canceling each other. The remaining integration over the horizontal sections are: $E_{t1}w - E_{t2}w = 0$. Hence, boundary condition for tangential field component is

$$\boxed{E_{t1} = E_{t2}.} \tag{2.29}$$

Refer to the right panel in Fig. 2.6; the upper free space can be treated as a dielectric medium with permittivity ϵ_1 while the medium below has ϵ_2. Gauss's law says that surface integral of **D** field over the cylindrical surface should be equal to free charges bounded by the surface. Surface integral over the vertical surface is trivial and tends to cancel itself. The remaining integration over the top and bottom surfaces are: $D_{n1}S - D_{n2}S = \rho_s S$. Hence, boundary condition for normal component of displacement field is

$$\boxed{D_{n1} - D_{n2} = \rho_s.} \tag{2.30}$$

In case that a surface charge is absent, one has

$$D_{n1} - D_{n2} = 0, \quad \text{or} \ \ \epsilon_1 E_{n1} = \epsilon_2 E_{n2}. \tag{2.31}$$

2.9 Capacitor

We know that a single conductor body is at the same potential. From Eq. 2.13, electric potential is proportional to the amount of charges on the body. We use *capacitance* to denote the ratio between the amount of charges on a conductor body and its electric potential. That is

$$C = \frac{Q}{V}. \tag{2.32}$$

For capacitance of a single object, V is object's potential with respect to potential at infinity. Capacitance has unit coulomb per volt (C/V), or equivalently farad (F).

The term "capacitor" is used specially for designating an important device consisting two charged planar metal plates placed relatively close to each other. The classic configuration is two parallel metal plates: one charged with Q and the other $-Q$. Capacitance of the double-plate structure is defined similarly with Eq. 2.32, except that V is potential difference between the two plates, and Q is magnitude of charge on one plate.

For a parallel-plate capacitor with plate area S, separation d, and dielectric filling ϵ, one can further deduce the relation between Q and V through calculating electric field between two plates. Assume the charges are uniformly distributed with surface density ρ_s. Within the plates, the electric field is uniform, everywhere-normal to the plates, with value (Section 2.4, third example)

$$E = \frac{\rho_s}{\epsilon} = \frac{Q}{\epsilon S}. \tag{2.33}$$

The potential difference, or voltage, between the two plates is

$$V = Ed = \frac{Qd}{\epsilon S}. \tag{2.34}$$

Therefore, capacitance of a two-plate capacitor according to Eq. 2.32 is

$$\boxed{C = \frac{\epsilon S}{d}. \quad \text{(parallel-plate)}} \tag{2.35}$$

2.10 Electrostatic energy

A static electric field is generated by an aggregate of charges of the same polarity. The field contains energy. One way to understand it is by imagining that each charge in the aggregate has to be moved from infinity to its place against electric force due to other charges in the aggregate. The total energy contained by the system is the total external work done in moving charges to their respective places. It turns out that the energy can be directly related to electric field intensity, as

$$W_e = \frac{1}{2}\int_V \mathbf{E}\cdot\mathbf{D}dv = \frac{1}{2}\int_V \epsilon E^2 dv, \tag{2.36}$$

which is valid for linear dielectric materials.

A parallel-plate capacitor has approximately uniform electric field, and its stored electrostatic energy is then

$$\boxed{W_e = \frac{1}{2}\int_V \epsilon\left(\frac{V}{d}\right)^2 dv = \frac{1}{2}\epsilon\left(\frac{V}{d}\right)^2 (Sd) = \frac{1}{2}CV^2.} \tag{2.37}$$

Summary of equations

Two fundamental equations governing electrostatics are

$$\oint_S \mathbf{D}\cdot d\mathbf{s} = Q, \text{ or } \nabla\cdot\mathbf{D} = \rho; \qquad \text{(Gauss's law, electric)}$$

$$\oint_C \mathbf{E}\cdot d\mathbf{l} = 0, \text{ or } \nabla\times\mathbf{E} = 0. \qquad \text{(static } \mathbf{E} \text{ is conservative)}$$

Exercises

1. A point charge $q = 1$ nC is placed at origin in free space. Calculate its electric field at point with Cartesian coordinates $x = 10$ cm, $y = 10$ cm, and $z = 0$ cm, or $(10, 10, 0)$ cm in short. Give both magnitude and direction.

2. A point charge $q = -1$ nC is located at Cartesian coordinates (0, 2, 0) cm in free space. Calculate its electric field at $(10, 10, -1)$ cm. Give both magnitude and direction.

3. A planar line loop of square shape with side length $a = 10$ cm is placed on xy plane in free space, centered at origin. The line loop carries a total charge of $Q = 1$ nC, which is uniformly distributed along the loop. Calculate electric field by the charged line loop at the point with Cartesian coordinates $(0, 0, 10)$ cm.

4. A point charge $q_1 = 1$ nC is placed at origin in free space. Another point charge $q_2 = 0.1$ nC is at a point with Cartesian coordinates $(10, 10, 0)$ cm. Calculate electric force experienced by q_2 due to electric field of q_1. Calculate also work done by electric field of q_1 on q_2 when q_2 moves from $(10, 10, 0)$ to $(20, 20, 0)$ cm in a straight line.

5. A hollow spherical conductor is centered at origin in free space. It has inner and outer radii at $r_1 = 10$ cm and $r_2 = 20$ cm, respectively, and carries an excessive charge of $Q = 1$ nC. Use Gauss's law to calculate electric field at the point with Cartesian coordinates $(40, 40, 0)$ cm. How about electric field at $(5, 5, 0)$ cm (and by which principle)?

6. Same as the previous problem. Calculate electric potential at $(40, 40, 0)$, $(15, 15, 0)$, and $(5, 5, 0)$ cm, respectively.

7. A static electric field has the following spatial dependence in Cartesian coordinates: $\mathbf{E} = 2y\hat{\boldsymbol{x}} + 2x\hat{\boldsymbol{y}} + 0\hat{\boldsymbol{z}}$ V/m. SI units are assumed. Calculate the electric potential difference between the point at $(4, 6, 0)$ m and origin. (Try to solve with two approaches.)

8. A uniformly distributed circular line charge with a radius $R = 10$ cm is placed on xy plane in free space, centered at origin. Line charge density is $\rho_l = 0.1$ μC/cm. Another point charge $q = -2$ nC moves *slowly* by an external force from point P_1 with Cartesian coordinates $(0, 0, 10)$ cm to point P_2 at $(0, 0, 5)$ cm. Calculate work done by the external force on q.

9. A hollow dielectric sphere is centered at origin in free space. It has inner and outer radii at $r_1 = 10$ cm and $r_2 = 20$ cm, respectively. The dielectric material has a relative permittivity of $\epsilon_r = 10$. If a point charge of $Q = 1$ nC is placed at the center of the hollow sphere, calculate electric fields at points with Cartesian coordinates $(40, 40, 0)$ cm, $(15, 15, 0)$ cm, and $(5, 5, 0)$ cm, respectively.

10. Consider a charge-free interface between air (approximated by vacuum) and a dielectric medium with relative permittivity $\epsilon_r = 2.5$. The interface plane can be treated as a planar surface in xy plane placed at $z = 0$. The electric field in air at a point *just outside*, or with Cartesian coordinates $(0, 0, 0^+)$, is $\mathbf{E}_1 = 10\hat{\boldsymbol{x}} - 8\hat{\boldsymbol{y}} + 6\hat{\boldsymbol{z}}$ V/m. Calculate electric field *just inside* the dielectric medium, at $(0, 0, 0^-)$.

11. A coaxial cable has a center metal wire with radius $r_1 = 0.5$ mm and a thin cylindrical metallic shell with radius $r = 4$ mm. Space in between the two metal layers is filled with a dielectric material with relative permittivity $\epsilon_r = 8$. Calculate capacitance per meter length of the coaxial cable.

Chapter 3

Electric Circuit

We have so far dealt with static electric charges and their fields. One of the greatest achievements in electrical engineering is creation of electric circuits, through which electric charges are set into a stable motion by an applied voltage (electric field). In the simplest but not trivial case, the work done by an electric field on charges becomes kinetic energy of charge carriers (i.e. electrons), which through collision with lattice is then converted to heat and in turn thermal radiation. This was how we had incandescent electric lighting for the whole twentieth century. Although in an electric circuit, charges are no longer static, the formulas obtained in "electrostatics" are still valid as long as physical quantities involved *do not vary sharply* in time. We show that the basic concepts and formulas encountered in circuit theory can be derived from our knowledge on electrostatics in combination with classical dynamics.

The most primitive parameters in an electric circuit are voltage V and current I. V is due to existence of electric field. To understand current I, one has first to know what is *current density* $\mathbf{J}$. It follows that, resistence R, which is merely ratio V/I, can be quantified for well-defined conductor geometries. Circuit laws, including Joule's law and Kirchhoff's laws for both current and voltage, can be clarified from basic principles in classical dynamics and electrostatics.

3.1 Electric current density and current

Free charge carriers in a conducting medium tend to be set into motion when there exists an electric field[1]. If it is a uniform medium excited by a uniform field, all charges will settle macroscopically into a constant velocity called drift velocity[2] v; in a more general scenario, charge motions can be space-dependent. A general charge flow can be described by a vector field called *electric current density* $\mathbf{J}$. At a certain spatial point, $\mathbf{J}$ has the direction of applied electric field, and its magnitude corresponds to the amount of electric charges passing through a unit cross-sectional area per unit time. While a conducting medium can consist multiple types of charge carriers (*e.g.* in plasma), here we focus on metals which have just free electrons responsible for charge flow. Under an electric field, free electrons in a metal move along (actually opposite to) the field direction, resulting in an electric current density $\mathbf{J}$. If volume density of free electrons is N, and each electron carries charge q and has an average speed v, $\mathbf{J}$ has magnitude

$$J = Nqv. \tag{3.1}$$

[1] Random thermal motion will not lead to a macroscopic charge flow.

[2] We use v to denote speed. Note that v is sometimes used to represent voltage. Also, do not mix with dv in volume integration.

N and v depend on exact metal type; v is also, of course, decided by excitation field. The direction of $\mathbf{J}$ is decided by v direction, modified by sign of q. Usually, for a certain metal in a steady state, v is proportional to imposed electric field magnitude; therefore, $\mathbf{J}$ is proportional to applied electric field $\mathbf{E}$, i.e.

$$\mathbf{J} = \sigma\mathbf{E}. \tag{3.2}$$

Here, σ is called *conductivity*, which summarizes material properties including N and lattice size, etc. Value of σ for common conductors are well documented in handbooks.

The above equations are valid at any volume position. An electric circuit, however, restricts $\mathbf{E}$ and $\mathbf{J}$ into singly-directed quantities, *i.e.* along metal wire. Therefore, one can discard their vector nature and care only about their magnitudes. Charge distribution across a thin metal wire's cross-section is unimportant, whilst how fast charges flow through wire is. Hence, one has definition of *electric current* as

$$I = \frac{dQ}{dt}, \tag{3.3}$$

where Q is amount of charge passing through a wire's cross-section.

Generally, for any virtual surface S (with differential element $d\mathbf{s}$) in a conducting medium where a $\mathbf{J}$ field exists, one calculates current passing through the surface as

$$I = \int_S \mathbf{J} \cdot d\mathbf{s}. \tag{3.4}$$

3.2 Ohm's law, resistance

If a voltage V is applied on two ends of a metal wire with cross-sectional area S, one has an electric field E and thereof current J in the wire. Scalar quantities are used since their directions are unambiguously defined as the wire direction. In electric circuit, current is $I = JS$. Equation 3.2 can be written as

$$\frac{I}{S} = \sigma\frac{V}{l}. \tag{3.5}$$

One has voltage-current relationship

$$V = \left(\frac{l}{\sigma S}\right) I, \tag{3.6}$$

which is the classic *Ohm's law* in circuit theory

$$\boxed{V = RI.} \tag{3.7}$$

Resistance R is calculated from the conductor's material and geometric parameters as

$$\boxed{R = \frac{l}{\sigma S}. \quad \text{(uniform conductor)}} \tag{3.8}$$

3.3 Joule's law

Here we show that under an applied voltage V, electric power dissipated in a piece of uniform metal wire (length l, cross-section S, resistance R) is $P = VI = I^2R$, *i.e.* Joule's law.

While drifting due to applied electric field, free electrons will experience collisions with stationary ions. After collision, the free electrons will pick up some random moving directions — drift velocity is temporarily reset to zero, and then increases quadratically with respect to traveling distance until next collision. To simplify analysis, we ignore the reset-acceleration processes, and assume the electrons immediately gain a constant average drift velocity v after each collision. v is defined as the ratio between average collision distance δl and average collision time δt, or $v = \delta l/\delta t$. In this picture, kinetic energy is lost suddenly and then re-picked up immediately on each collision. Note that random electron motion due to thermal drift can be ignored. Therefore, charge velocity and hence current density are in the same direction as the applied electric field. Hence, scalar quantities can be used in the following analysis.

Energy dissipated in each collision for charge q is equal to work done by the electric field on the charge as

$$\delta w = (qE)\ \delta l, \tag{3.9}$$

corresponding to a power

$$p = \frac{\delta w}{\delta t} = qEv. \tag{3.10}$$

For a small volume dv with number of charges per unit volume as N, the power dissipated is

$$dP = NqEv\ dv. \tag{3.11}$$

Since $J = Nqv$, the above equation becomes

$$dP = EJ\ dv. \tag{3.12}$$

The total power dissipated in a volume is

$$\boxed{P = \int_V EJ\ dv.} \tag{3.13}$$

For an elongated conductor with cross-sectional area S and length l,

$$\boxed{P = \int_V EJ\ dv = \int_C E\ dl \int_S J\ ds = VI = I^2R.} \tag{3.14}$$

3.4 Kirchhoff's current law

In a general 3D volume, $\mathbf{J}$ is a position-dependent vector field depicting flow of charges. The principle of *conservation of charges* dictates that, during any time interval, charges flowing out of a volume is equal to change of charges within the volume. Mathematically, the equation reads

$$\boxed{\oint_S \mathbf{J} \cdot d\mathbf{s} = -\frac{dQ}{dt}.} \tag{3.15}$$

The LHS can be converted to a volume integral through divergence theorem, and the RHT side is in fact also a volume integral in terms of charge density ρ. Therefore

$$\int_V \nabla \cdot \mathbf{J}\ dv = -\int_V \frac{\partial \rho}{\partial t}\ dv. \tag{3.16}$$

By equating the kernels, we have *equation of continuity*

$$\nabla \cdot \mathbf{J} = -\frac{\partial \rho}{\partial t}. \tag{3.17}$$

In a steady state, Q and ρ do not vary with time. Hence Eqs. 3.17 and 3.15 become respectively

$$\nabla \cdot \mathbf{J} = 0, \quad \text{and} \quad \oint_S \mathbf{J} \cdot d\mathbf{s} = 0. \tag{3.18}$$

The latter in Eq. 3.18 is in fact *Kirchhoff's current law* in circuit theory, which says sum of currents into (or out of) a node is zero. That is

$$\boxed{\sum_j I_j = 0.} \tag{3.19}$$

3.5 Kirchhoff's voltage law

Kirchhoff's voltage law states that sum of voltages around a closed circuit is equal to zero. It is a direct result from conservation nature of (static) electric field, and is mathematically equivalent to Eq. 2.16. Kirchhoff's voltage law is expressed as

$$\sum_j V_j = 0. \tag{3.20}$$

3.6 Circuit with capacitor

3.6.1 DC bias

Refer to the simple electric circuit with a resistor and capacitor connected in series in Fig. 3.1. Such a circuit is referred to as a RC circuit. We would like to find out current I running in the circuit as a function of time, after the circuit is connected. Under DC bias, based on Kirchhoff's voltage law, one has

$$RI + V_c = V, \tag{3.21}$$

or

$$RI + \frac{Q}{C} = V. \tag{3.22}$$

Take derivative against time on both sides. One has

$$R\frac{dI}{dt} + \frac{1}{C}\frac{dQ}{dt} = \frac{dV}{dt}. \tag{3.23}$$

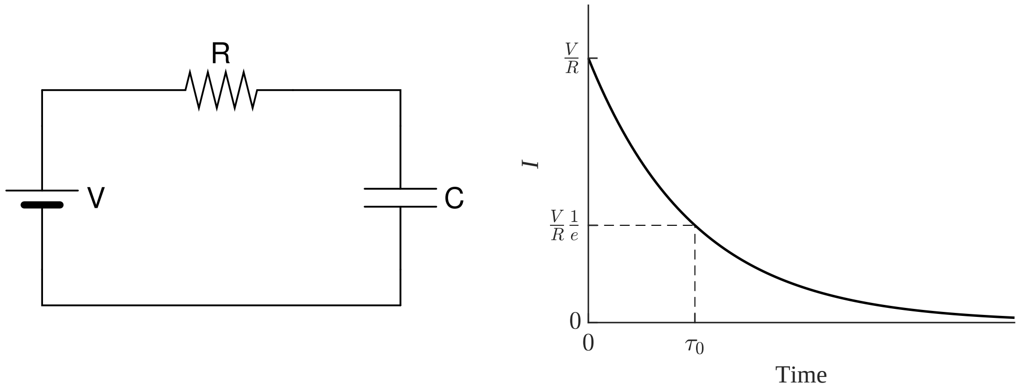

Figure 3.1: Left: Simple RC electric circuit. Right: Current as a function of time.

For DC, $dV/dt = 0$. And time variation of charge dQ/dt is simply current I. Hence,

$$I = \frac{dQ}{dt}. \tag{3.24}$$

Equation 3.23 becomes

$$\frac{dI}{dt} + \frac{1}{RC} I = 0. \tag{3.25}$$

This first-order differential equation can be solved as

$$I(t) = \frac{V}{R} \exp\left(-\frac{t}{RC}\right). \tag{3.26}$$

As shown in the right panel in Fig. 3.1, current drops from V/R to 0, with a *time constant* of $\tau_0 = RC$ of the RC-circuit. Charge and voltage on the capacitor are

$$Q(t) = VC\left[1 - \exp\left(-\frac{t}{\tau_0}\right)\right], \tag{3.27}$$

$$V_c(t) = V\left[1 - \exp\left(-\frac{t}{\tau_0}\right)\right]. \tag{3.28}$$

3.6.2 AC bias

If source voltage has sinusoidal time dependence, one can resort to frequency-domain analysis through the so-called *phasor expressions.* A harmonic source voltage and current are written as complex phasors as

$$V = V_0 \exp(i\omega t), \quad \text{and } I = I_0 \exp(i\omega t), \tag{3.29}$$

where ω is angular frequency and i is the imaginary unit ($i^2 = -1$). From Eq. 3.23, one obtains

$$R\frac{dI}{dt} + \frac{1}{C} I = \frac{dV}{dt} \tag{3.30}$$

Substituting Eq. 3.29, one has

$$R(i\omega)I_0 + \frac{1}{C} I_0 = i\omega V_0. \tag{3.31}$$

A common $\exp(i\omega t)$ factor has been eliminated for all terms. One has straightforwardly relation between V_0 and I_0

$$V_0 = \left(R + \frac{1}{i\omega C}\right) I_0. \tag{3.32}$$

We define *impedance* of a capacitor as

$$Z_c = \frac{1}{i\omega C}, \tag{3.33}$$

which has the same unit as resistance but with a purely imaginary value.

Voltage across the capacitor is

$$V_c = V - RI = V_0 \exp(i\omega t)\left(1 - \frac{i\omega RC}{1 + i\omega RC}\right). \tag{3.34}$$

Voltage on the capacitor goes to zero as $\omega \to \infty$. The circuit is a low-pass filter.

3.6.3 Cascaded capacitors

Cascaded capacitors can be treated effectively as a single capacitor. One uses individual impedances, as if they are resistances for resistors, to calculate the overall effective impedance. In the case of serial connection, the total impedance is calculated as

$$Z = Z_1 + Z_2 + ... + Z_n. \tag{3.35}$$

If all capacitors are of parallel-plate type with the same geometry, Z corresponds to impedance of an effective capacitor with plate distance increased by n times, compared to an individual capacitor in the series. Total capacitance is reduced by n times.

In the case of parallel connection, one has

$$\frac{1}{Z} = \frac{1}{Z_1} + \frac{1}{Z_2} + ... + \frac{1}{Z_n}. \tag{3.36}$$

Again, if all capacitors are of parallel-plate type with the same geometry, the overall effective capacitor has a plate area n times as large as that of an individual capacitor. Total capacitance is increased by n times.

Exercises

1. A planar sheet of infinite size is placed on xy plane. The sheet carries a uniform charge density $\rho_s = 1$ nC/cm^2. If the sheet moves towards $+x$ direction with a speed of 10 m/s. Calculate effective surface current density on the sheet from a standing-still observer.

2. A circular line loop carries a uniform charge of 20 μC. The loop has radius 15 cm and is placed on xy plane with its center at origin. If the loop is set to rotation around z axis with a speed of 50 rounds per second (clockwise as observed from $z = +\infty$), calculate the effective current in the loop for a stationary observer.

3. A point charge of 20 μC is placed at 15 cm from origin. If the point charge is set to rotation around z axis with a speed of 50 rounds per second (clockwise as observed from $z = +\infty$), calculate the effective current due to the moving charge for a stationary observer.

Chapter 4

Magnetostatics

When we talk about "action at a distance", magnets probably leave us a deeper impression. Brio's toy trains never fail to amuse toddlers, simply because (I guess) the force they feel has a different character than gravity. For thousands of years, people thought this magic force was only associated with certain stones or metals until Hans Christian Ørsted[1] observed in 1820 about magnetic field around a current-carrying metal wire. Since then, we came to know that magnetism is caused by charges in motion, or electric current.

4.1 Magnetic field and force

4.1.1 Magnetic field

Magnetic field, denoted by symbol **B**, has unit of tesla (T). A magnetic field is generated by moving charges, and the field exerts force on other moving charges. In principle, one can formulate magnetic field and magnetic force using charges and their velocities. However, it is sometimes convenient to use the high-level, more-measurable quantity — current I. This is especially true since charge flow is most often guided in thin metal wires, not unbounded in 3D space. A *static* magnetic field is generated by a *constant* current.

The shape of magnetic field generated by a small straight section of current I placed at origin O is depicted by the Biot-Savart's law[2], which is equivalent to Coulomb's law in electrostatics. It says

$$d\mathbf{B} = \frac{\mu_0 I}{4\pi}\frac{d\mathbf{l}\times\hat{\boldsymbol{r}}}{r^2}, \quad \text{or} \quad = \frac{\mu_0 I}{4\pi}\frac{d\mathbf{l}\times\mathbf{r}}{r^3}. \quad \text{(unit: tesla, T)} \tag{4.1}$$

μ_0 is a fundamental constant called free-space permeability ($\mu_0 = 4\pi \times 10^{-7}$ henry/meter, H/m). The current source has magnitude I and direction encoded in $d\mathbf{l}$. $\mathbf{r}$ is position vector with length r and unit vector direction $\hat{\boldsymbol{r}}$; if the current section is not placed at origin, one should use displacement vector $\mathbf{r}_{\text{SP}}$, vector from source point S to observation point P. Note that current always goes in a closed loop; therefore, the total magnetic field at a point usually requires one carries out a closed-loop line integral of the expression above.

Question: According to Eq. 4.1, one can picture magnetic field generated by a small line section of current $I\ d\mathbf{l}$. Describe main properties of the field.

[1] Hans Christian Ørsted (1777-1851): Danish physicist who discovered that there exists circulating magnetic field around a current-carrying metal wire.

[2] Formulated by Frenchmen Jean-Baptiste Biot and Félix Savart in 1820.

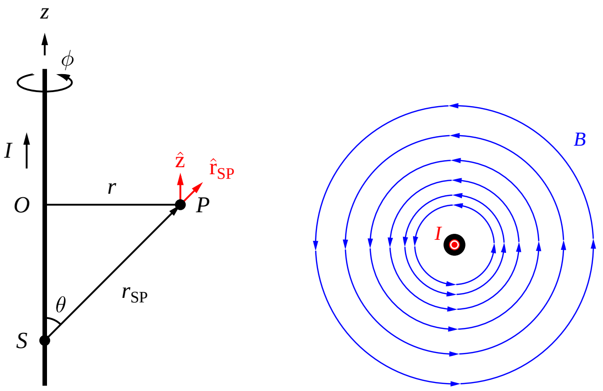

Figure 4.1: Left: Schematic for calculation of magnetic field produced by an infinite line current. Right: **B** field generated (view from $+z$ side).

Magnetic field by line current

Based on Biot-Savart's law, derive magnetic field generated by a straight, infinitely long metal wire carrying a current I (Fig. 4.1, left panel).

Solution: Use cylindrical coordinate, and place the wire on the z axis with current running in $+z$ direction. Based on symmetry, we can argue that the field generated will be independent of azimuthal angle ϕ, or the z coordinate. Therefore, we only pay attention to the field's dependence on r coordinate. We assume the observation point P is leveled at $z = 0$.
Field at P by an elemental current section at S is

$$d\mathbf{B} = \frac{\mu_0 I dz}{4\pi}\frac{\hat{\boldsymbol{z}}\times\hat{\boldsymbol{r}}_{\mathrm{SP}}}{r_{\mathrm{SP}}^2} = \frac{\mu_0 I dz}{4\pi}\frac{1}{r_{\mathrm{SP}}^2}\sin\theta\hat{\boldsymbol{\phi}}.$$

Total field at P is

$$\begin{aligned}\mathbf{B} &= \int_{z=-\infty}^{\infty} d\mathbf{B} = \int_z \frac{\mu_0 I dz}{4\pi}\frac{1}{r_{\mathrm{SP}}^2}\sin\theta\hat{\boldsymbol{\phi}} = \frac{\mu_0 I}{4\pi}\hat{\boldsymbol{\phi}}\int_z \frac{1}{r_{\mathrm{SP}}^2}\sin\theta dz \\ &= \frac{\mu_0 I}{4\pi}\hat{\boldsymbol{\phi}}\int_\theta \frac{\sin^2\theta}{r^2}\sin\theta\frac{r}{\sin^2\theta}d\theta \quad \leftarrow \left[\frac{-z}{r} = \cot\theta \;\rightarrow\; dz = \frac{r}{\sin^2\theta}d\theta\right] \\ &= \frac{\mu_0 I}{4\pi}\hat{\boldsymbol{\phi}}\int_\theta \frac{\sin\theta}{r}d\theta = \frac{\mu_0 I}{4\pi r}\hat{\boldsymbol{\phi}}[-\cos\theta]_0^\pi = \frac{\mu_0 I}{2\pi r}\hat{\boldsymbol{\phi}}.\end{aligned}$$

The magnetic field is rotating around the line current (Fig. 4.1, right panel), with magnitude decaying according to radial distance as $1/r$.

Biot-Savart's law dictates that the direction of magnetic field generated by a line current is related to the current flow direction through the *right-hand rule*: with right-hand thumb pointing towards the current, the rest fingers naturally curl along the magnetic field direction.

In the above example, we didn't consider complete current loop. It is a valid approximation if our interested field region is relatively close to the straight line current compared to the rest of current circuit. If complete current distribution is considered, calculation of **B** field can be tedious. However, for problems possessing certain symmetry, solution via Biot Savart's law can be manageable.

Magnetic field by circular current loop

In cylindrical coordinate, a circular line current loop with radius a and current $I\hat{\boldsymbol{\phi}}$ is placed on $r\phi$ plane with center at origin. Based on Biot-Savart's law, derive magnetic field along z axis. [Exam 2019]

Solution: For a small line segment on the current loop $d\mathbf{l}$ one has

$$d\mathbf{l} = ad\phi\hat{\boldsymbol{\phi}}, \quad \text{and} \quad \mathbf{r}_{\text{SP}} = z\hat{\mathbf{z}} - a\hat{\mathbf{r}}.$$

The total magnetic flux density at P is

$$\begin{aligned}
\mathbf{B} &= \oint_C \frac{\mu_0 I}{4\pi}\frac{d\mathbf{l}\times\mathbf{r}_{\text{SP}}}{r_{\text{SP}}^3} = \int_\phi \frac{\mu_0 I}{4\pi}\frac{ad\phi\hat{\boldsymbol{\phi}}\times(z\hat{\mathbf{z}}-a\hat{\mathbf{r}})}{(z^2+a^2)^{\frac{3}{2}}} \\
&= \int_\phi \frac{\mu_0 I}{4\pi}\frac{(az\hat{\mathbf{r}}+a^2\hat{\mathbf{z}})}{(z^2+a^2)^{\frac{3}{2}}}d\phi \qquad \leftarrow [\hat{\mathbf{r}}\text{ comp. cancels out with integration}] \\
&= \int_\phi \frac{\mu_0 I}{4\pi}\frac{a^2}{(z^2+a^2)^{\frac{3}{2}}}d\phi\ \hat{\mathbf{z}} = \frac{\mu_0 I a^2}{2(z^2+a^2)^{\frac{3}{2}}}\ \hat{\mathbf{z}}
\end{aligned}$$

4.1.2 Magnetic force

Magnetic force experienced by a charge q moving in velocity $\mathbf{v}$ is

$$\boxed{\mathbf{F}_m = q\mathbf{v}\times\mathbf{B}.} \tag{4.2}$$

The force is always perpendicular to velocity. Magnetic force does no work to a charged particle. If one injects a charged particle into a magnetic field with an initial velocity perpendicular to the field direction, trajectory of the particle will form a circle. The radius of circle has been widely used for determining charge and mass of unknown particles.

It is more often in electrical engineering to have a stream of moving charges or current. A small section of conductor $d\mathbf{l}$ carrying a current I in a magnetic field $\mathbf{B}$ experiences a force, governed by

$$\boxed{d\mathbf{F}_m = Id\mathbf{l}\times\mathbf{B}.} \tag{4.3}$$

On a whole conductor loop, the net force will be integration of $d\mathbf{F}_m$ along the loop path.

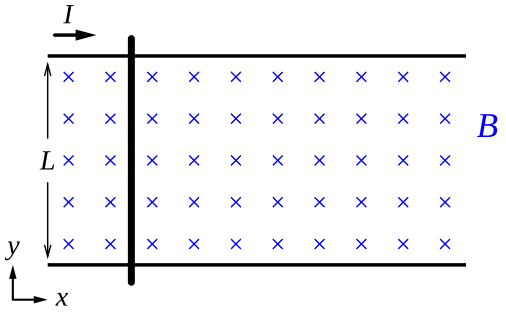

Figure 4.2: Magnetic force on moving conductor (with electric contact) in a magnetic field.

Magnetic force on conductor carrying current

Refer to Fig. 4.2. A straight conductor is placed on top of two parallel metal rails separated by distance L. Between the two rails, there exists a uniform magnetic field with magnitude B, directed into paper. Calculate magnetic force on the vertical conductor when a current I flows as shown.

Solution: Through Eq. 4.3, the total force is

$$\mathbf{F}_m = \int_C I d\mathbf{l} \times \mathbf{B} = \int_C I(-dl\hat{\boldsymbol{y}}) \times (-B\hat{\boldsymbol{z}}) = \int_C BIdl\hat{\boldsymbol{x}} = BIL\hat{\boldsymbol{x}}.$$

4.2 Magnetic dipole

As mentioned, magnetic monopole doesn't exist. The most primitive magnetic field source is a *magnetic dipole* — in the form of a small current loop. The second example in the previous section revealed partially magnetic field of a magnetic dipole. The full field of a magnetic dipole can be qualitatively pictured by applying right-hand rule along the current loop. A sketch of the full magnetic field lines is shown in Fig. 4.3 (right panel). It is compared by the electric-field lines of an electric dipole (left). In fact, the two vector fields share exactly the same spatial dependence at distance much larger than the geometric size of the dipoles. Some features include: the fields have rotation symmetry around their axis (independent of ϕ); their fields have the maximum amplitude at equator (when $\theta = 90°$); and their fields have zero amplitude along polar directions (when $\theta = 0°$ or $180°$).

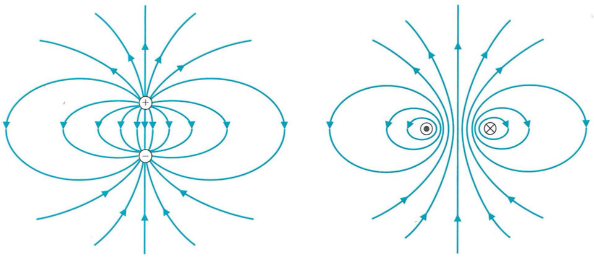

Figure 4.3: Left: Electric-field lines of an electric dipole. Right: Magnetic-field lines of a magnetic dipole.

For a *planar* current loop, regardless of its shape, we define its *magnetic dipole moment* as product of current and area of current loop. For a circular current loop with radius a, one has

$$\boxed{\mathbf{m} = I(\pi a^2)\hat{\boldsymbol{n}} \quad (\text{unit: A}\cdot\text{m}^2).} \tag{4.4}$$

$\hat{\boldsymbol{n}}$ is surface normal unit vector, determined from current direction by right-hand rule.

4.3 Magnetic torque

Torque due to magnetic force deserves a separate section, since the mechanism underpins operation of all electric motors. If field $\mathbf{B}$ is uniform, the total force on a loop in the field is always zero. However, a closer look tells that a current loop tends to turn. Refer to the left panel in Fig. 4.4, a rectangular current loop carrying current I with length a and height b is placed in a uniform magnetic field $\mathbf{B}$. The left and right current sections do not experience magnetic force since they are parallel to $\mathbf{B}$. According to Eq. 4.3, the upper section experiences a magnetic force $F = BIa$, pointing out of paper; and the lower section experiences the same amount of force but in opposite direction. The loop shall undergo rotation around the indicated red axis, with a torque $T = 2F\frac{b}{2} = Fb = BIab = Bm$, where m is amplitude of the loop's magnetic moment $\mathbf{m}$. The torque is at its largest magnitude when the current loop is at the illustrated position, or when *loop plane is parallel to magnetic field*. Imagine after rotating 90 degrees, we arrive at the middle panel in Fig. 4.4. It is difficult to see the current loop; therefore, we switch our perspective by observing the loop from below. What we see then is as shown in the right panel of Fig. 4.4. Biot-Savart's law tells that now all line sections experience a force. Forces on the vertical sections F' vary in magnitude during the rotation, but they always cancel each other and do not contribute to rotation. Forces on the horizontal sections F have constant amplitudes during rotation; at this particular position they don't result in a torque as they pass through the central axis. This minimum torque occurs when *loop plane is normal to magnetic field*, or *the current loop's own magnetic field is aligned with external magnetic field*.

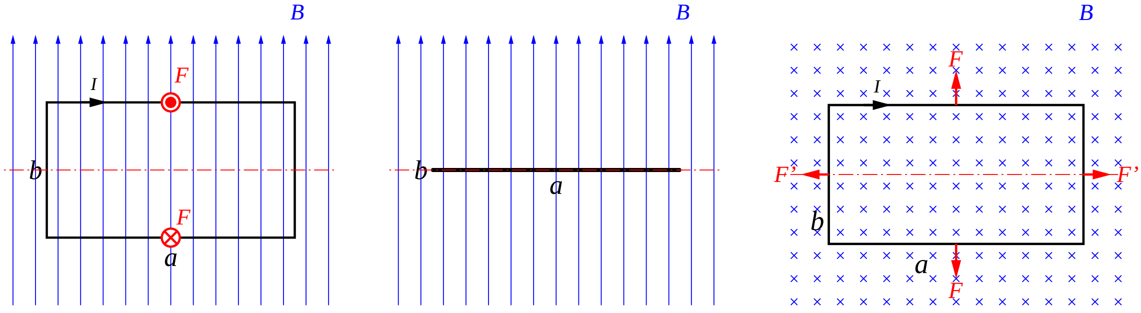

Figure 4.4: Left: Force and torque on a current loop in magnetic field. Middle: Rotation after 90 degrees. Right: Same as the middle panel but viewed from below.

A more rigorous analysis can lead to the following general formula for magnetic torque on a planar current loop of any shape in a uniform magnetic field,

$$\boxed{\mathbf{T} = \mathbf{m} \times \mathbf{B}.} \tag{4.5}$$

The torque is affected by angle between the magnetic field and the surface normal of loop plane. In a stable condition (minimum torque), a magnetic dipole aligns its dipole moment with external magnetic field.

Magnetic torque on circular current loop

Calculate torque on circular current loop by magnetic field $\mathbf{B}$ (Fig. 4.5).

Solution: A quick examination reveals that the upper half circle and the bottom half circle give the same amount of torque around the rotation axis. We focus on the top half. For a differential current section, the magnetic force at point P (Eq. 4.3) is

$$d\mathbf{F} = Id\mathbf{l} \times \mathbf{B} = I(ad\phi)(-\hat{\boldsymbol{\phi}}) \times \mathbf{B} = Iad\phi B \sin\phi\ \hat{\boldsymbol{z}}.$$

Differential torque at P is

$$d\mathbf{T} = \mathbf{r}_{\text{O'P}} \times d\mathbf{F} = (a \sin\phi\ \hat{\boldsymbol{y}}) \times (Iad\phi B \sin\phi\ \hat{\boldsymbol{z}}) = Ia^2 B \sin^2\phi d\phi\ \hat{\boldsymbol{x}}.$$

The total torque is

$$\begin{aligned}\mathbf{T} = \int d\mathbf{T} &= 2 \cdot \int_{\phi=0}^{\pi} Ia^2 B \sin^2\phi d\phi \hat{\boldsymbol{x}} \\ &= 2Ia^2 B \int_0^{\pi} \sin^2\phi d\phi\ \hat{\boldsymbol{x}} = I(\pi a^2)B\ \hat{\boldsymbol{x}}.\end{aligned}$$

One can arrive at the solution much more quickly through Eq. 4.5.

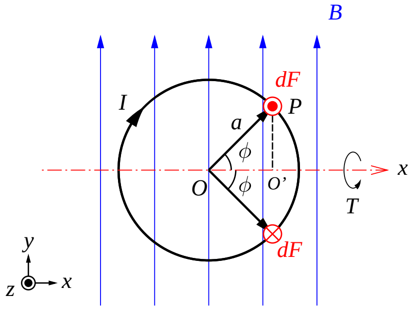

Figure 4.5: Magnetic torque on a circular current loop.

4.4 Laws for static magnetic field

Behavior of static magnetic field $\mathbf{B}$ is governed by two fundamental equations, which conform to our observations so far. First, there exist no isolated magnetic monopoles (or magnetic charges). In Gauss's formulation, it is expressed as

$$\oint_S \mathbf{B} \cdot d\mathbf{s} = 0. \quad (\text{Gauss}'\text{s law, magnetic field}) \tag{4.6}$$

Magnetic flux coming out of a closed surface is zero. Through divergence theorm, it can be written in differential form

$$\nabla \cdot \mathbf{B} = 0. \tag{4.7}$$

This equation is also referred to as *Gauss's law for magnetic field*. The absence of magnetic monopole is also evidenced by the fact that the magnetic field lines close on themselves, whereas electric field lines originate from positive charges (monopoles) and close on negative charges (monopoles).

Another equation governing magnetic field is due to Andrè-Marie Ampère[3]. It relates vector line integral of magnetic field along a closed loop (can be a fictitious line loop, also known as *Amperian loop*) to total current passing through the loop. In free space, it is expressed as

$$\oint_C \mathbf{B} \cdot d\mathbf{l} = \mu_0 I. \quad (\text{Ampère's law in free space, magnetostatics}) \tag{4.8}$$

According to Stoke's theorem, LHS of Eq. 4.8 can be converted to a surface integral, as $\oint_C \mathbf{B} \cdot d\mathbf{l} = \int_S \nabla \times \mathbf{B} \cdot d\mathbf{s}$. RHS of Eq. 4.8 can also be written in integral form as $\mu_0 I = \int_S \mu_0 \mathbf{J} \cdot d\mathbf{s}$. By equating the kernels, one can put Eq. 4.8 in differential form as

$$\nabla \times \mathbf{B} = \mu_0 \mathbf{J}. \tag{4.9}$$

[3] Andrè-Marie Ampère (1775-1836): French physicist who contributed to relation between electricity and magnetism.

Equation 4.9 relates curl of magnetic field to current density at a spatial point. In a general scenario, current does not have to be uni-directional or of uniform density. Eq. 4.8 (or equivalently Eq. 4.9) is called *Ampère's law*, valid so far for magnetostatics.

We focus on applications of integral forms of the two fundamental relations, *i.e.* Eqs. 4.6 and 4.8 in this text. It is worth pointing out (again) that the integration path or surface in the two relations are arbitrary. Below we show how *Ampère's law* can be used to simplify calculation magnetic field around a straight line current.

Magnetic field by line current, through Ampère's law

Based on Ampère's law, derive magnetic field generated by a straight, infinitely long metal wire carrying a current I.

Solution: Symmetry reasoning says that the generated magnetic field should be directed along ϕ, invariant along z, and constant at a fixed radial r position. Therefore we use a circular integration path in $r\phi$ plane, with the line current at its center. There, we should have a uniform $\mathbf{B} = B\hat{\boldsymbol{\phi}}$. Based Eq. 4.8, integration along the circular path is

$$\oint_C \mathbf{B}\cdot d\mathbf{l} = \oint_C B\hat{\boldsymbol{\phi}}\cdot rd\phi\hat{\boldsymbol{\phi}} = Br\oint_\phi d\phi = Br[\phi]_0^{2\pi} = B\ 2\pi r.$$

The quantity should be equal to $\mu_0 I$. Therefore $B = \frac{\mu_0 I}{2\pi r}$, in ϕ direction.

Magnetic field by planar current, through Ampère's law

Derive magnetic field generated by an infinite, thin current sheet placed on xy plane with a uniform surface current density J_s flowing along $+x$ direction.

Solution: Refer to lecture notes. Importantly, owing to symmetry, field is uniform in $+z$ half space with direction pointing to $-y$, and is uniform of the same magnitude but directed towards $+y$ direction in $-z$ half space.

4.5 Magnetization, H field

We now know all atoms are made of charged particles. Among them, electrons are continuously circulating around heavier nuclei, giving rise to orbiting currents and hence magnetic dipole moments. Of less significant effect, the electrons, while orbiting, are spinning around their own axes, also leading to magnetic dipole moments. Without an external magnetic field, these magnetic dipole moments are oriented in random directions, resulting no net magnetic field. Under an external magnetic field, the magnetic dipoles tend to align their dipole moments towards the field direction, hence a net dipole moment appears. The material is being *magnetized*. We use *magnetization* $\mathbf{M}$ to denote volume density of resulted magnetic dipole moments. Mathematically, $\mathbf{M}$ is vector summation of magnetic dipole moments of atoms in each unit volume; practically, $\mathbf{M}$ is an un-measurable quantity. What we do know is that the resulted $\mathbf{M}$ will lead to a secondary internal magnetic field $\mathbf{B}_i$, whose strength is proportional to $\mathbf{M}$. We simply say $\mathbf{B}_i = \mu_0\mathbf{M}$. Taking curl of this equation, one has

$$\nabla\times\frac{\mathbf{B}_i}{\mu_0} = \nabla\times\mathbf{M}. \tag{4.10}$$

Side note: It can be proven that RHS is simply effective current generated in the medium, *i.e.* $\nabla \times \mathbf{M} = \mathbf{J}_m$.

On the other hand, from Eq. 4.9, one has

$$\nabla \times \frac{\mathbf{B}_e}{\mu_0} = \mathbf{J}, \tag{4.11}$$

where $\mathbf{B}_e$ and $\mathbf{J}$ are external magnetic field and free current density which generates $\mathbf{B}_e$, respectively. By combining the two above equations, we have

$$\frac{1}{\mu_0}\nabla \times (\mathbf{B}_e + \mathbf{B}_i) = \mathbf{J} + \nabla \times \mathbf{M}. \tag{4.12}$$

We denote the sum $(\mathbf{B}_e + \mathbf{B}_i)$ as the total magnetic field in the medium $\mathbf{B}$. One then has

$$\nabla \times \left(\frac{\mathbf{B}}{\mu_0} - \mathbf{M}\right) = \mathbf{J}. \tag{4.13}$$

If we define a new quantity called *magnetic field intensity* $\mathbf{H}$ as

$$\boxed{\mathbf{H} = \frac{\mathbf{B}}{\mu_0} - \mathbf{M}, \quad \text{(unit: ampere per meter, A/m)}} \tag{4.14}$$

Eq. 4.13 can be simplified as

$$\nabla \times \mathbf{H} = \mathbf{J}. \tag{4.15}$$

$\mathbf{H}$ is therefore an auxiliary field which encompasses both total magnetic field as well as magnetization of a medium. Use of $\mathbf{H}$ can simplify mathematical expression relating *total field in a material* and *free current* $\mathbf{J}$. In free space, from Eq. 4.14, one has $B = \mu_0 H$.

One can integrate LHS of Eq. 4.15 over a surface and convert it to a line integral through Stoke's theorem as

$$\int_S (\nabla \times \mathbf{H}) \cdot d\mathbf{s} = \oint_C \mathbf{H} \cdot d\mathbf{l}. \tag{4.16}$$

Vector integral of RHS of Eq. 4.15, *i.e.* $\mathbf{J}$, on the same surface is nothing but total current I passing through the surface. Therefore, one has

$$\oint_C \mathbf{H} \cdot d\mathbf{l} = I. \quad \text{(Ampère's law, magnetostatics)} \tag{4.17}$$

Equation 4.17 is a more general form of Ampère's law, valid in both free space and in medium.

4.6 Permeability, magnetic materials

For relatively small degree of magnetization, $\mathbf{M}$ has linear dependence on magnetic field intensity $\mathbf{H}$. Hence,

$$\mathbf{M} = \chi_m \mathbf{H}. \tag{4.18}$$

We call the proportionality constant χ_m as *magnetic susceptibility*. Substitute the relation to Eq. 4.14, one has

$$\mathbf{B} = \mu_0(1 + \chi_m)\mathbf{H}. \tag{4.19}$$

μ_0 is free-space permeability. The term in parentheses represents magnetic response of a medium, which we call *relative permeability*, *i.e.* $\mu_r = (1 + \chi_m)$. The product $\mu_0\mu_r$ is simply called permeability. Hence, we have *constitutive relation* for magnetic field as

$$\mathbf{B} = \mu_0\mu_r\mathbf{H} = \mu\mathbf{H}. \tag{4.20}$$

Free space has $\mu_r = 1$.

Magnetic property of a material that undergoes linear magnetization can be captured by a fixed μ_r value (linear B-H curve). We can roughly put materials into three categories according to their μ_r values. They are

- **Diamagnetic** if $\mu_r \lesssim 1$
- **Paramagnetic** if $\mu_r \gtrsim 1$
- **Ferromagnetic** if $\mu_r \gg 1$

For both diamagnetic and paramagnetic materials, μ_r values differ from 1 by some small value on the order of 10^{-5}. Such materials are therefore hardly attracted by a magnet. Ferromagnetic materials exhibit strong responses to a magnetic field. Microscopically, a ferromagnetic material has small naturally *magnetized domains* separated by *domain walls*. Under a magnetic field, domains with magnetizations directed along the field tend to grow in size, whereas the other domains shrink. This causes an increase in B field. The B-H relation is initially linear and reversible. However at large H, the relation becomes nonlinear and reversible only through a hysteretic process (following by a lag). The lag is caused by resistance in movement of domain walls. In short, B-H relation does not follow the same curve when one increases or decreases H (magnetizing) field; rather, the magnetization process traces a loop in the BH plane (Fig. 4.6), which we refer to as a hysteresis loop.

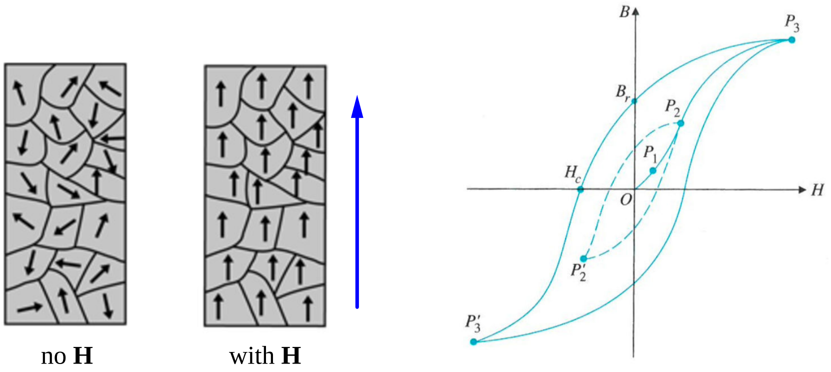

Figure 4.6: Left: Ferromagnet with and without excitation. Right: Magnetization curve - hysteresis loop.

Some comments can be drawn from a hysteresis curve, such as

- Ferromagnetic material can become permanent magnet.
- The area of a hysteresis loop corresponds to energy loss per applied magnetic field intensity cycle (*i.e.* $H_{\text{max}} \to -H_{\text{max}} \to H_{\text{max}}$). For applications where repeated change of magnetic field is used, one shall use soft ferromagnetic materials with narrow hysteresis loops; for permanent magnets, one prefer hard ferromagnetic materials with fat hysteresis loops.

A permanent magnet can be de-magnetized through heating. The critical temperature at which residual magnetization disappears is called *curie temperature*.

4.7 Boundary conditions

To solve for magnetic field distribution, one needs to know the boundary conditions when field crosses a material boundary. This can be done similarly as we did for electrostatics, but with the two fundamental relations for magnetostatics in Sections 4.4&4.5. We re-use figures in Fig. 2.6 for following clarifications, except that the "free space" is replaced by another material (material 1 with permeability μ_1 and the material below has μ_2).

Refer to Fig. 2.6 (right panel). The condition for normal component can be obtained by making surface integral of **B** on a fictitious box with infinitesimal height on a material interface. The absence of magnetic monopole requires that the normal B component should be continuous across a material interface. That is,

$$\boxed{B_{n1} = B_{n2}, \quad \text{or } \mu_1 H_{n1} = \mu_2 H_{n2}.} \tag{4.21}$$

Refer to Fig. 2.6 (left panel). The condition for tangential H field component can be obtained by making line integral of the field on a fictitious loop with infinitesimal height on a material interface. According to Ampère's law for magnetostatics (Eq. 4.17), the result should be equal to total current passing through the integration loop. One has

$$\boxed{H_{t1} - H_{t2} = J_s.} \tag{4.22}$$

J_s is surface current density directed *normal* to the integration loop. For non-conducting media, one usually has $J_s = 0$; hence the tangential H field is in general continuous across a material interface.

4.8 Magnetic energy

A static magnetic field has energy. One can imagine a magnetic field is generated by a collection of circulating currents or magnetic dipoles placed close to each other (within a magnet, for example). In corresponding ground state, the current loops are placed infinitely apart. So the energy contained by a magnetic field is the amount of energy needed to move the loops from being far away to a close adjacency (while keeping currents unchanged in all loops). In terms of field strength, magnetic energy is

$$W_m = \frac{1}{2}\int_V \mathbf{H}\cdot\mathbf{B}\, dv = \frac{1}{2}\int_V \frac{B^2}{\mu}\, dv. \tag{4.23}$$

In next chapter, it will be shown that magnetic field energy can also be defined in terms of *inductance*.

4.9 Cascaded current loops — coil

We have studied magnetic field due to small section of line current (Biot-Savart's law), a current loop (a magnetic dipole). One important device that critically contributes to our modern electrification is *current coil*, alternatively known as *solenoid*.

Shown in Fig. 4.7 (left panel), a current coil is a helical conductor wire. It can be treated equivalently as many current loops placed side by side with the same electric current running through them. Despite its relatively complex structure, one can easily calculate magnetic field inside a current coil through Ampère's law. Consider a line-integration loop (indicated by the red dashed line in right panel of Fig. 4.7) with one side running within the coil. According to Ampère's law, line integral of **B** field around the integration loop shall be equal to the total current passing through the loop. For a coil with N turns and current I, one has

$$\oint_C \mathbf{B}\cdot d\mathbf{l} = \mu_0 N I. \tag{4.24}$$

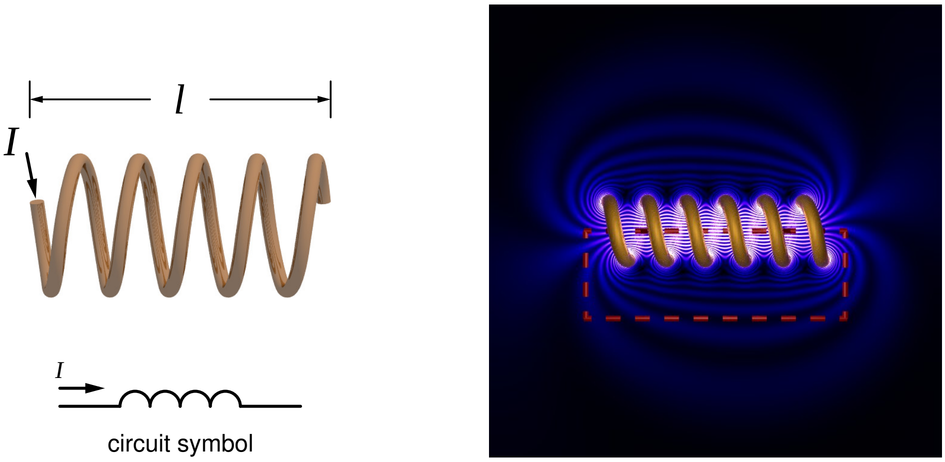

Figure 4.7: Left: A current coil and its circuit symbol. Right: Magnetic field of a coil.

As Fig. 4.7 shows, magnetic field lines spread out heavily outside the coil; in other words, field amplitude outside in general is much weaker than field inside the coil. When treating LHS of Eq. 4.24, one can approximately discard integration along line sections outside the coil. Furthermore, if one assumes **B** is evenly distributed inside the coil, one has

$$Bl = \mu_0 NI, \tag{4.25}$$

where l is the coil length. Hence

$$B = \frac{\mu_0 NI}{l}. \tag{4.26}$$

If the coil is wound around a ferromagnetic bar with permeability μ, one has

$$\boxed{B = \frac{\mu NI}{l}.} \tag{4.27}$$

Such a device is also referred to as *electromagnet.* A ferromagnetic core can greatly enhance magnetic flux density (e.g. iron has $\mu_r = 200000$). An extended ferromagnetic core can channel the flux to another coil, so as to achieve power transfer without electrical connection (transformer; see next chapter).

A current coil with a constant current generates magnetic field of very similar pattern as that by a permanent magnet. This is not a coincidence. In a permanent magnet, there are residual magnetic dipole moments pointing to its north pole. One can think of these dipoles as caused by current loops (e.g. electrons orbiting around their nuclei) with loop planes normal to the magnet-bar direction. The microscopic currents tend to cancel each other so there appear to be no net current within the magnet *except on the surface of the magnet.*

Summary of equations

Two fundamental equations governing magnetostatics are

$$\oint_S \mathbf{B} \cdot d\mathbf{s} = 0, \text{ or } \nabla \cdot \mathbf{B} = 0; \quad \text{(Gauss's law, magnetic field)} \tag{4.28}$$

$$\oint_C \mathbf{H} \cdot d\mathbf{l} = I, \text{ or } \nabla \times \mathbf{H} = \mathbf{J}. \quad \text{(Ampère's law, magnetostatics)} \tag{4.29}$$

Exercises

1. Consider a section of straight line conductor of length 10 cm aligned on x axis with its center at origin. A current of 10 A follows along $+x$ direction in the conductor. Calculate magnetic flux density $\mathbf{B}$ generated by the line current section at point with Cartesian coordinates (0,0,5) cm.

2. A square-shaped current loop with side length 10 cm is placed on xy plane, centered at origin. From $+z$ perspective, one sees a current of 10 A flowing in the loop along counterclockwise direction. Calculate magnetic flux density $\mathbf{B}$ generated by the current loop at point with Cartesian coordinates (0, 0, 5) cm.

3. A circular current loop with radius 5 cm is placed plane-parallel to xy plane with a displacement along $-z$ direction by 5 cm. Another current loop of the same size is placed also plane-parallel to xy plane but with a displacement along $+z$ by 5 cm. Both loops are centered around z axis and carry a current of 10 A in counterclockwise direction (if viewed from $z = +\infty$). Calculate $\mathbf{B}$ field at origin generated by the two current loops.

4. Refer to Exercise 2 in Chap. 3. Calculate the generated magnetic flux density $\mathbf{B}$ at the loop center.

5. A circular disk carries a uniform charge of 100 μC. The disk has a radius of 5 cm and is placed on xy plane with its center at origin. If the disk rotates about z axis at a speed of 50 rounds per second (counterclockwise observed from $z = +\infty$), calculate the generated magnetic flux density $\mathbf{B}$ at the disk center.

6. A straight wire of length 8 cm is oriented along x axis, centered at origin. Two end sections of the wire, each with length of 2 cm, is charged with a uniform line charge $\rho_l = 2$ nC/cm. If the wire is set to rotation around z axis with a speed of 100 rounds per second (counterclockwise observed from $z = +\infty$), calculate $\mathbf{B}$ field generated at origin.

7. A particle carrying a charge of $q = -1$ nC is moving with an instantaneous velocity $v = (300\hat{\boldsymbol{x}} + 500\hat{\boldsymbol{y}})$ m/s in a uniform magnetic field $\mathbf{B} = 2\hat{\boldsymbol{x}}$ mT. Calculate the magnetic force on the particle.

8. Same as the above but with a magnetic field $\mathbf{B} = (2\hat{\boldsymbol{y}} + 2\hat{\boldsymbol{z}})$ mT.

9. A charged particle is accelerated to velocity $v = 1\hat{\boldsymbol{x}}$ km/s and is then sent into a uniform magnetic field $\mathbf{B} = 5\hat{\boldsymbol{z}}$ T. The particle has mass $m = 1$ pg and charge $q = -0.2$ nC. Describe, as quantitatively as possible, the motion of the particle in the magnetic field.

10. Two straight line conductors are oriented parallel to each other with a separation distance of 10 cm. Both conductors carry a 10 A current towards the same direction. Calculate magnetic force experienced by each conductor owing to magnetic field of the other conductor. Specify their directions.

11. A thin conductor sheet of infinite size is placed on xy plane. A uniform current with surface current density $J_s = 0.1$ A/mm flows in the conductor sheet along $+x$ direction. Calculate $\mathbf{B}$ field generated by the current sheet at point with Cartesian coordinate (0, 0, 10) cm.

Chapter 5

Magnetic Induction

Magnetic field generated by electric current in one circuit or device can “induce” a current in a second circuit or device placed in the field. This action of “inductance” happens when there is a change of magnetic field of, or equivalently a change of current in the primary circuit. The primary circuit acts as an electromagnet. A permanent magnet can replace the role of the primary circuit. One would then have to move the permanent magnet in order to induce a current in the second circuit. A current in the second circuit is essentially due to existence of electric field. Hence, *time-varying magnetic field generates electric field.* Magnetic induction tells that electric field and magnetic field are not separate phenomena, but are coupled to each other.

5.1 Faraday’s discovery

Michael Faraday[1] discovered in 1831 that if one inserts or pulls a magnet in or out of a conductor loop, a current is observed in the loop (see Fig. 5.1). The magnitude of current is related to the speed of magnet movement.

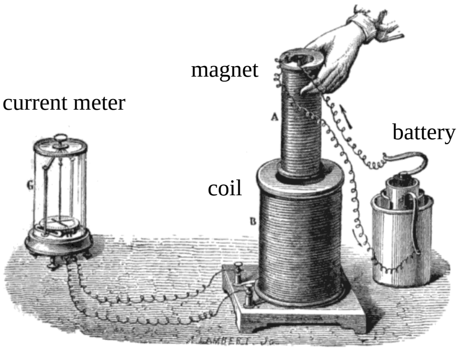

Figure 5.1: Faraday’s experiment on magnetic induction.

This phenomenon is summarized as *Faraday’s law of electromagnetic induction* — change of magnetic flux in a loop leads to an *electromotive force (EMF)* $\mathscr{E}$ (equivalent to a voltage sustained by a battery) along the loop. Furthermore, it is observed that the resulted current due to the EMF generates a magnetic flux that compensates change of the original magnetic flux (Lenz’s law). Mathematically, Faraday’s law says

$$\mathscr{E} = -\frac{d\Phi}{dt} \qquad \text{(unit: volt, V)}, \tag{5.1}$$

[1] Michael Faraday (1791-1867): English self-taught scientist who contributed critically to electromagnetic induction, and consequently electric motors.

where Φ is *total* flux passing through concerned circuit.

The magnitude of Φ essentially depends on coupling between the "inducing" magnetic field and the circuit under consideration in the field. To quantify this coupling, we create a new quantity called "inductance", discussed below.

5.2 Inductor and inductance

5.2.1 Self inductance

In electrostatics, we came across capacitor and capacitance. A capacitor (usually two metal pieces) stores charges after being applied an electric potential. Capacitance is defined as ratio between the stored charge and the potential; mediating field is electric field. In magnetostatics, one has similar concepts: *inductor* and *inductance*. An inductor (e.g. a metal loop) holds a *magnetic flux* after being applied an electric current. The mediating field is magnetic field. Magnetic flux flowing through a loop is

$$\boxed{\Phi = \int_S \mathbf{B} \cdot d\mathbf{s}. \qquad \text{(unit: weber, Wb)}} \tag{5.2}$$

In general, definition of inductance[2] L is

$$L = \frac{\Phi}{I}. \qquad \text{(unit: henry, H)} \tag{5.3}$$

I is current in an inductor circuit, and Φ is total magnetic flux passing through the circuit. Like capacitance, inductance L depends on the inductor's geometry and material. For a linear medium, L does not depends on current in the loop.

Based on the above definition, it is straightforward to find out inductance of a current coil. Given a coil with number of turns N, current I, and length l, magnetic field inside the coil is $B = \frac{\mu N I}{l}$. The total magnetic flux that the coil encloses is

$$\Phi = \int_{S_{\text{coil}}} \mathbf{B} \cdot d\mathbf{s} = \int_{S_{\text{coil}}} B ds = \int_{S_{\text{coil}}} \frac{\mu N I}{l} ds. \tag{5.4}$$

Notice that a subscript "coil" is added just as a reminder that the surface integration is not on the coil's cross-section, but on *a surface bounded by the coil's helical line.* One can well treat a N-turn coil as N separated current loops, and thereby the integration surface can be approximated by N circular surfaces bounded by each current loop. That is to say, the surface integral shall be carried on the coil's cross-section S for N times. Furthermore, direction of the differential surface element is in line with the magnetic field direction; hence the vector dot product became product of two scalars.

$$\Phi = \frac{\mu N I}{l}(S)(N) = \frac{\mu N^2 S}{l} I. \tag{5.5}$$

Hence the (self) inductance of a coil is

$$\boxed{L_{\text{coil}} = \frac{\mu N^2 S}{l}.} \tag{5.6}$$

Definition of inductance allows expression of magnetic energy stored in an inductor as

$$W_m = \frac{1}{2} L I^2. \tag{5.7}$$

[2] Symbol L is used in honour of the physicist Heinrich Lenz.

5.2.2 Mutual inductance

For a capacitor, its electric field originates from positive charges and terminates on negative charges; the electric field can be well confined and does not affect other neighbouring capacitors. In the contrary, magnetic field line has to form a loop, which is extended in space. Therefore, an inductor's magnetic field not only passes through its own circuit, but also can pass through a neighbouring inductor, e.g. another circuit. For this reason, we have "self inductance" and "mutual inductance". The term "inductance" by default refers to "self inductance".

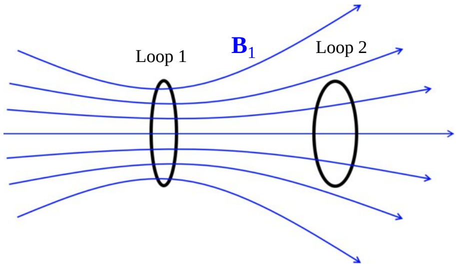

Figure 5.2: Mutual inductance.

Take two circular loops in Fig. 5.2 for example. Current in loop 1 generates a magnetic field, some of which passes through loop 2. Loop 1 has a "magnetic influence" on loop 2; or, there exists a mutual inductance between the two loops. To quantify the mutual inductance, one first calculates the amount of magnetic flux generated by loop 1 that passes through loop 2, as

$$\Phi_{12} = \int_{S_2} \mathbf{B}_1 \cdot d\mathbf{s}_2. \tag{5.8}$$

Then the mutual inductance defined as

$$\boxed{L_{12} = \frac{\Phi_{12}}{I_1},} \tag{5.9}$$

where I_1 is current in loop 1. It can be proven that there is always

$$\boxed{L_{12} = L_{21}.} \tag{5.10}$$

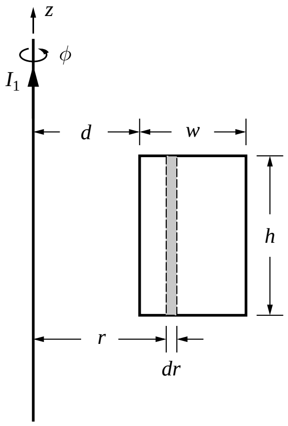

Figure 5.3: Mutual inductance between infinite line conductor and square conductor loop.

Mutual inductance

Calculate mutual inductance between an infinitely long conductor and a rectangular conductor loop placed as shown in Fig. 5.3.

Solution: Consider straight line conductor as circuit 1 and the square conductor loop as circuit 2. It is easier to calculate magnetic flux due to the straight conductor passing through circuit 2 than the other way around. We know from Ampère's law that the magnetic field due to current in circuit 1 is $\mathbf{B} = \frac{\mu_0 I_1}{2\pi r}\hat{\boldsymbol{\phi}}$. Flux passing through circuit 2 is

$$\begin{aligned}\Phi_{12} &= \int_{S_2} \mathbf{B}_1 \cdot d\mathbf{s}_2 = \int_{r=d}^{d+w} \left(\frac{\mu_0 I_1}{2\pi r}\hat{\boldsymbol{\phi}}\right) \cdot (drh\hat{\boldsymbol{\phi}}) \\ &= \frac{\mu_0 I_1 h}{2\pi} \int_{r=d}^{d+w} \frac{1}{r} dr = \frac{\mu_0 I_1 h}{2\pi} \ln\left(1 + \frac{w}{d}\right).\end{aligned}$$

Hence, the mutual inductance is

$$L_{12} = \frac{\Phi_{12}}{I_1} = \frac{\mu_0 h}{2\pi} \ln\left(1 + \frac{w}{d}\right).$$

5.3 Faraday's law of induction

5.3.1 Induction leads to electromotive force

The electromotive force defined by Faraday's law in Eq. 5.1 can be written in terms of inductance, by use of Eq. 5.3, as

$$\boxed{\mathscr{E} = -L\frac{dI}{dt}.} \tag{5.11}$$

Here L can be self *or* mutual inductance. In the case of self inductance, an EMF (more properly voltage) is induced by the inductor's own current variation. In the case of mutual inductance, an EMF is induced by current change in some other inductor!

5.3.2 Faraday's law of induction, generalized

Faraday's law of magnetic induction, *i.e.* Eq. 5.1, applies well to a conductor loop. However, it has far more general implication. Effectively, $\mathscr{E}$ is line integral of induced **E**, and magnetic flux Φ is area integral of magnetic field **B** — *time-varying magnetic field produces electric field!* Faraday's observation is one special instance of the following relation between field quantities **E** and **B**

$$\oint_C \mathbf{E} \cdot d\mathbf{l} = -\int_S \frac{\partial \mathbf{B}}{\partial t} \cdot d\mathbf{s}. \quad \text{(Faraday's law)} \tag{5.12}$$

Through Stoke's theorem, LHS can be converted to surface integral of $\nabla \times \mathbf{E}$. Therefore, the above equation becomes

$$\int_S \nabla \times \mathbf{E} \cdot d\mathbf{s} = -\int_S \frac{\partial \mathbf{B}}{\partial t} \cdot d\mathbf{s}. \tag{5.13}$$

Although in Faraday's observation, this equation is valid for a surface bounded by a conductor loop encompassing a time-varying **B** field, a little forward-thinking leads to the belief that it should be valid on all surfaces, regardless of size, location, or orientation. Neither **E** nor **B** needs existence of a conductor loop. Taking the kernels of Eq. 5.13, one has an equation in differential form

$$\nabla \times \mathbf{E} = -\frac{\partial \mathbf{B}}{\partial t}. \quad \text{(Faraday's law)} \tag{5.14}$$

The above equation is valid at any spatial point for all material systems. Equations 5.12&5.14 are generalized Faraday's law of induction, which shall replace the fundamental relation $\oint_C \mathbf{E} \cdot d\mathbf{l} = 0$ (or differential form $\nabla \times \mathbf{E} = 0$) in electrostatics. *Electric field becomes non-conservative in a time-varying scenario.*

5.3.3 EMF due to moving conductor in magnetic field

When a metal is moving in a magnetic field, free electrons in the metal can experience a magnetic force. This force consequently drags the electrons towards one direction, resulting in a build-in electric field, hence an induced EMF in the moving metal. The electric field exerts an electric force on electrons, which is working against the magnetic force. An equilibrium state is reached when the electric force is equal to the magnetic force in magnitude. The steady-state EMF in a magnetic field **B** is

$$\boxed{\mathscr{E} = \int_C (\mathbf{v} \times \mathbf{B}) \cdot d\mathbf{l}\ ,} \tag{5.15}$$

where **v** is velocity of the conductor.

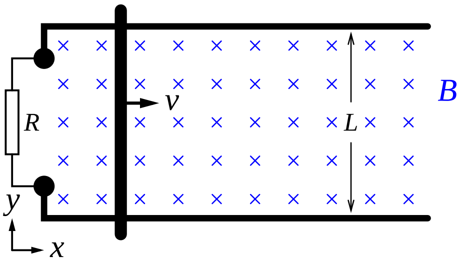

Figure 5.4: EMF induced by moving conductor in stationary magnetic field.

For a moving conductor loop, one uses loop integral. The final result is equivalent to $\mathscr{E} = -\frac{d\Phi}{dt}$, *i.e.* EMF generated in the loop is proportional to (negative) change of magnetic flux in the loop per unit time. Consequence: no *overall* EMF in the circuit if **B** is spatially uniform.

EMF due to moving conductor in magnetic field

Refer to Fig. 5.4. A metal bar is moving in a constant magnetic field with a fixed velocity. Calculate current through the resistor connected to the left terminals.

<u>Solution</u>: The voltage between the two terminals (top v.s. bottom) is

$$\mathscr{E} = \int_{\text{bottom}}^{\text{top}} (\mathbf{v} \times \mathbf{B}) \cdot d\mathbf{l} = \int_{y=0}^{L} [v\hat{\boldsymbol{x}} \times (-B\hat{\boldsymbol{z}})] \cdot dy\hat{\boldsymbol{y}} = vBL.$$

Current is then $I = \frac{\mathscr{E}}{R} = \frac{vBL}{R}$, along $-y$ direction.

5.4 Electric circuit with inductor

Figure 5.5 (left panel) shows a simple electric circuit with a voltage source V, a resistor R as well as an inductor L connected in series (a RL circuit). One can solve analytically for current, given a DC or AC voltage source.

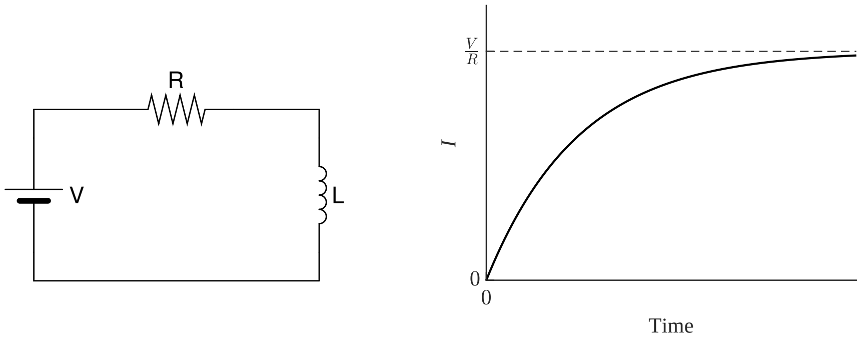

Figure 5.5: Left: Simple electric circuit with a resistor and an inductor, or RL circuit. Right: Current response.

DC case

Given a DC voltage source, the RL circuit will take some time to settle to a steady state after switch is closed. According to Kirchhoff's voltage law, sum of voltage drops across the resistor V_R and the inductor V_L shall be equal to that of the voltage source.

$$V_R + V_L = V. \tag{5.16}$$

Note that voltage "decreases" across an inductor under "rising" current in the coil (and vice versa), *i.e.* $V_L = -\mathscr{E} = L\frac{dI}{dt}$. Hence

$$RI + L\frac{dI}{dt} = V. \tag{5.17}$$

Current can be solved as

$$I(t) = \frac{V}{R}\left[1 - \exp\left(-\frac{R}{L}t\right)\right]. \tag{5.18}$$

One sees $I_{t=0} = 0$ and $I_{t=\infty} = V/R$, as illustrated in the right panel in Fig. 5.5. The voltage across the inductor is

$$V_L(t) = V - RI(t) = V\exp\left(-\frac{R}{L}t\right). \tag{5.19}$$

AC case

If the voltage is sinusoidal. One resorts to phasor expressions

$$V = V_0\exp(i\omega t), \tag{5.20}$$

$$I = I_0\exp(i\omega t). \tag{5.21}$$

Substituting them into Eq. 5.17, one has

$$RI_0 + (i\omega)LI_0 = V_0. \tag{5.22}$$

It follows that

$$I_0 = \frac{V_0}{R + i\omega L}. \tag{5.23}$$

We define impedance of the inductor as

$$Z_L = i\omega L. \tag{5.24}$$

The voltage across the inductor is

$$V_L = V - RI = V\frac{i\omega L}{R + i\omega L}. \tag{5.25}$$

One notices $V_{L,\omega=0} = 0$; or, voltage is appreciable only when frequency is high. The circuit is a high-pass filter.

Notice that impedance Z_L plays a similar role as resistance. When coil inductors are connected in series, one can lump the coils as a single inductor with an impedance $Z_L = Z_{L1} + Z_{L2} + ... + Z_{LN}$. When coils are connected in parallel, one has $\frac{1}{Z_L} = \frac{1}{Z_{L1}} + \frac{1}{Z_{L2}} + ... + \frac{1}{Z_{LN}}$. We assume here that there is no mutual inductance among the inductors.

5.5 Transformers

A ferromagnetic material with large permeability can be heavily magnetized and carries a large total magnetic field **B** in the material. A ferromagnetic bar, also referred to as "magnetic core", is often used to confine and guide magnetization and thereby magnetic field for various applications. A classic application is *transformer*, as illustrated in Fig. 5.6. Under AC bias, a transformer transfers electric energy based on "perfect" magnetic induction from one electric circuit to another through a magnetic core; in doing so, it can transform voltage as well as current. The ease of varying voltage with transformers across an electric network made AC a natural choice for distribution of electricity since early 1900s.

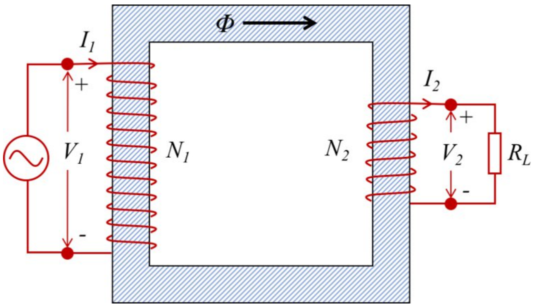

Figure 5.6: Schematic diagram for a transformer.

Refer to Fig. 5.6 and take note of the indicated directions and polarities. Magnetic field generated in the left primary coil (N_1 turns) is guided to the secondary coil (N_2 turns), without leakage in the case of $\mu_r \to \infty$ for the magnetic core. Guidance of magnetic field without leakage allows us to use total magnetic flux Φ. According to Faraday's law of induction

$$V_1 = N_1\frac{d\Phi}{dt}, \text{ and } V_2 = N_2\frac{d\Phi}{dt}. \tag{5.26}$$

Therefore, one has

$$\boxed{\frac{V_1}{V_2} = \frac{N_1}{N_2},} \tag{5.27}$$

which is well-known voltage relation across a transformer. For a perfect transformer, one assumes $\mu_r \to \infty$ in its magnetic core; in turn, one has $H = \frac{B}{\mu} \to 0$ in the magnetic core.

By applying Ampère's law, *i.e.* $\oint_C \mathbf{H} \cdot d\mathbf{l} = I$, on a closed integration path around the magnetic core, one can obtain relation between the input and output currents, as

$$\boxed{N_1 I_1 = N_2 I_2, \text{ or } \frac{I_1}{I_2} = \frac{N_2}{N_1}.} \tag{5.28}$$

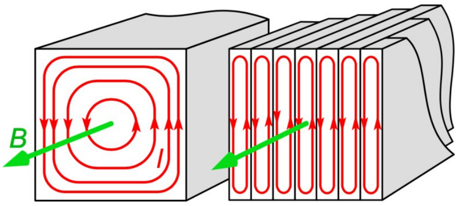

Figure 5.7: Eddy current in magnetic core and its mitigation.

Ferromagnetic core is usually made of conducting media, mostly commonly iron. As a result, an issue with transformer, which is also due to magnetic induction, is existence of Eddy current in the magnetic core, as illustrated in Fig. 5.7 (left). Eddy current can cause heating of the core material owing to Ohmic resistance, which leads to energy loss. An effective way to reduce such loss is to use a laminated ferromagnetic core — core made of a stack of ferromagnetic material pieces electrically insulated from each other, as shown in Fig. 5.7 (right). With such a design, Eddy current is confined in each individual lamination. Since magnitude of Eddy current is inversely proportional to cross-sectional area of a single lamination, one can minimize energy loss.

5.6 Summary of equations

Faraday discovered that time-varying magnetic field creates electric field

$$\oint_C \mathbf{E} \cdot d\mathbf{l} = -\int_S \frac{\partial \mathbf{B}}{\partial t} \cdot d\mathbf{s}, \quad \text{or } \nabla \times \mathbf{E} = -\frac{\partial \mathbf{B}}{\partial t}. \quad \text{(Faraday's law)}$$

Exercises

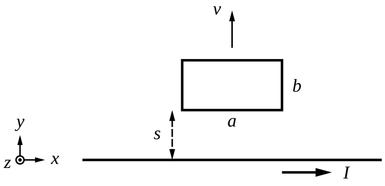

Figure 5.8: Exercise 1.

1. Refer to Fig. 5.8. A planar rectangular conductor loop is placed in plane with an infinite straight conductor in free space. A constant current $I = 10$ A flows in the line conductor towards the right. The rectangular conductor loop has width $a = 10$ cm and height $b = 5$ cm. The conductor loop is moving away from the line current with a velocity $v = 25$ cm/s. Calculate the electromotive force generated in the conductor loop at the moment when distance between its lower edge is separated from the line current by $s = 7.5$ cm, and current direction in the loop as a consequence of the generated electromotive force.

Chapter 6

Maxwell’s Equations and Wave Solutions

6.1 Maxwell’s equations

Before Faraday’s discovery in 1831, electric and magnetic phenomena were considered to have no mutual connection. Faraday showed that a varying magnetic field can create electric field. Not much happened thereafter in further development of theories governing the fields. Maxwell[1] tried in 1861 (barely 30 years old) to assemble the disparate laws into a coherent set of equations, and he discovered that the equation system can be consistent only if one accepts that, apart from current, *time-varying electric field can also induce magnetic field*. His postulate leads to a modified Ampère’s law, which is somewhat easier to understand from the differential form, as $\nabla \times \mathbf{H} = \mathbf{J} + \frac{\partial \mathbf{D}}{\partial t}$. The red term $\frac{\partial \mathbf{D}}{\partial t}$ is Maxwell’a addition, and it seemingly plays a similar role as free current density $\mathbf{J}$. For this reason, the term is sometimes referred to as *displacement current density*. The concept of displacement current is helpful for understanding e.g. the simple RC circuit in Chap. 3.1. More generally, together with Faraday’s law, it states that electric and magnetic fields are *mutually* coupled. Maxwell’s effort in unifying the laws and his contribution in adding the new term in Ampère’s law warrant that history remembers this set of equations in his name.

Maxwell’s equations in integral forms are

$$\oint_S \mathbf{D} \cdot d\mathbf{s} = Q, \quad \text{(Gauss's law, electric)} \tag{6.1}$$

$$\oint_S \mathbf{B} \cdot \ d\mathbf{s} = 0, \quad \text{(Gauss's law, magnetic)} \tag{6.2}$$

$$\oint_C \mathbf{E} \cdot \ d\mathbf{l} = -\int_S \frac{d\mathbf{B}}{dt} \cdot d\mathbf{s}, \quad \text{(Faraday's law)} \tag{6.3}$$

$$\oint_C \mathbf{H} \cdot \ d\mathbf{l} = I + \int_S \frac{\partial \mathbf{D}}{\partial t} \cdot \ d\mathbf{s}. \quad \text{(Ampère's law)} \tag{6.4}$$

[1] James Clerk Maxwell (1831–1879): Scottish scientist who unified electromagnetic equations and claimed that light is in fact electromagnetic wave.

And, in differential form, they are

$$\nabla \cdot \mathbf{D} = \rho, \quad \text{(Gauss's law, electric)} \tag{6.5}$$

$$\nabla \cdot \mathbf{B} = 0, \quad \text{(Gauss's law, magnetic)} \tag{6.6}$$

$$\nabla \times \mathbf{E} = -\frac{\partial \mathbf{B}}{\partial t}, \quad \text{(Faraday's law)} \tag{6.7}$$

$$\nabla \times \mathbf{H} = \mathbf{J} + \frac{\partial \mathbf{D}}{\partial t}. \quad \text{(Ampère's law)} \tag{6.8}$$

Maxwell's equations determine electric and magnetic fields (now electromagnetic field) as a function of space and time, from any given sources ρ and $\mathbf{J}$. One shall take note that ρ and $\mathbf{J}$ are not independent; they are connected through the principle of *charge conservation* or *equation of continuity*

$$\oint_S \mathbf{J} \cdot d\mathbf{s} = -\frac{dQ}{dt} \text{ (integral), or } \nabla \cdot \mathbf{J} = -\frac{\partial \rho}{\partial t} \text{ (differential).} \tag{6.9}$$

The principle states that total current flowing out from a closed surface is equal to time rate of charge decrease inside the enclosed volume. The addition of displacement-current term by Maxwell is precisely to make this equation satisfied, which one can verify by taking divergence of the Ampère's law (differential form).

6.2 Time-harmonic electromagnetic wave equation

Electromagnetic waves are generated by charge and current sources (ρ and $\mathbf{J}$). In many practical situations, the sources' dependences on time are of sinusoidal nature. Correspondingly, the fields generated are also sinusoidal. If not, a complicated time-dependence can be represented by multiple time-harmonic functions (detailed discussion of such decomposition falls into a specific subject called Fourier analysis). Time-harmonic sources and fields are therefore of significant interest in electrical engineering.

When analyzing time-harmonic systems, one often uses phasor notations. As an example, sinusoidal current $I(t) = I_o \cos{(\omega t + \phi)}$ is written in complex quantity during mathematical derivation as $I(t) = I_o \exp[i(\omega t + \phi)]$. I_0 is amplitude, ω is angular frequency, and ϕ is initial phase. i is imaginary unit which satisfies $i^2 = -1$. Final solutions are obtained by taking real part of the complex solutions. By using complex phasor expressions, one avoids manipulation of complicated trigonometric functions — exponential functions in phasor notation remain their forms upon differentiation or integration.

In phasor notation, a time-harmonic electric field can be expressed as $\mathbf{E}(x, y, z)\exp(i\omega t)$, and so on. Substituting the expressions into the Maxwell's equations (Eqs. 6.5-6.8) and eliminating the common $\exp(i\omega t)$ factor, one has

$$\nabla \cdot \mathbf{D} = \rho, \tag{6.10}$$

$$\nabla \cdot \mathbf{B} = 0, \tag{6.11}$$

$$\nabla \times \mathbf{E} = -i\omega \mathbf{B}, \tag{6.12}$$

$$\nabla \times \mathbf{H} = \mathbf{J} + i\omega \mathbf{D}. \tag{6.13}$$

Take note that the source and field quantities are now functions of space only, though the same symbols are used.

In a *homogeneous*, *source-free* region (ϵ and μ constant, $\rho = 0$ and $\mathbf{J} = 0$), by taking another curl on Eq. 6.12 and using the vector identity

$$\nabla \times \nabla \times \mathbf{A} = \nabla(\nabla \cdot \mathbf{A}) - \nabla^2 \mathbf{A}, \tag{6.14}$$

one can arrive at a Helmholz wave equation

$$\nabla^2 \mathbf{E} + \frac{\omega^2}{v^2}\mathbf{E} = 0, \tag{6.15}$$

where $v = \frac{1}{\sqrt{\epsilon\mu}}$. ∇^2 is another *del* operator called Laplace operator which we did not mention in Chap. 1. Mathematically, ∇^2 is "divergence of gradient", or $\nabla\cdot\nabla$. In Cartesian coordinate, it has the form

$$\nabla^2 = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2}. \tag{6.16}$$

Correspondingly, in Cartesian coordinate, Eq. 6.15 can be decomposed into three scalar wave equations.

6.3 Plane wave solution

The most primitive form of electromagnetic wave is *plane wave* — wave with planar constant-phase front. Such a wave has its electric field directed in one direction. If we assume that $\mathbf{E}$ is solely along x, we have the following scalar wave equation

$$\left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2}\right) E_x + \frac{\omega^2}{v^2} E_x = 0. \tag{6.17}$$

Further, if we consider that E_x field varies only along z direction, one has

$$\frac{\partial^2}{\partial z^2} E_x + k^2 E_x = 0, \tag{6.18}$$

where we used substitution $k = \frac{\omega}{v}$. The general solution to Eq. 6.18 is

$$\mathbf{E} = E_0^+ \exp(-ikz)\hat{\boldsymbol{x}} + E_0^- \exp(ikz)\hat{\boldsymbol{x}}, \tag{6.19}$$

where $E_0^\pm$ is electric-field amplitudes. The first term on RHS is a plane wave traveling towards $+z$ axis and the second term is traveling towards $-z$. k indicates how quickly phase varies along z; it is therefore spatial frequency of the wave. A more common name for k is *wave number*, indicating how many wavelengths in a unit propagation length (times 2π). It is easy to verify that $k = \frac{\omega}{v} = \frac{2\pi}{\lambda}$, where λ is electromagnetic wavelength in the medium. Comparatively, ω indicates how quickly phase varies in time t, and is therefore called *temporal (angular) frequency* of the wave.

Putting $\mathbf{E}$ (only the $+z$ traveling wave) into Eq. 6.12, one can compute the corresponding magnetic field as

$$\mathbf{H} = \frac{i}{\omega\mu}\frac{\partial E_x}{\partial z}\hat{\boldsymbol{y}} = \frac{1}{\sqrt{\frac{\mu}{\epsilon}}} E_0 \exp(-ikz)\hat{\boldsymbol{y}}. \tag{6.20}$$

One sees that the directions of electric field, magnetic field, and wave propagation form an orthogonal triplet. The magnetic field is in phase with the electric field and has a magnitude relative to that of the electric field by a factor called *intrinsic impedance* of the medium

$$\boxed{Z = \sqrt{\frac{\mu}{\epsilon}}.} \tag{6.21}$$

Including time-harmonic dependence, one has $+z$-propagating plane-wave solutions as

$$\mathbf{E} = E_0 \exp\left[i(-kz + \omega t)\right]\hat{\boldsymbol{x}}, \tag{6.22}$$

$$\mathbf{H} = H_0 \exp\left[i(-kz + \omega t)\right]\hat{\boldsymbol{y}}, \tag{6.23}$$

with $E_0/H_0 = Z$. The expressions above are in phasor form. The instantaneous fields are the real parts

$$\boxed{\begin{aligned}\mathbf{E} &= E_0\cos(-kz+\omega t)\hat{\boldsymbol{x}}, \\ \mathbf{H} &= H_0\cos(-kz+\omega t)\hat{\boldsymbol{y}}.\end{aligned}} \qquad \begin{aligned}(6.24)\\(6.25)\end{aligned}$$

They are sinusoidal spatial functions moving towards $+z$ as time increases. Figure 6.1 (left panel) schematically illustrates spatial variation of the two field components. The right panel in Fig. 6.1 shows another way of presenting the same plane wave.

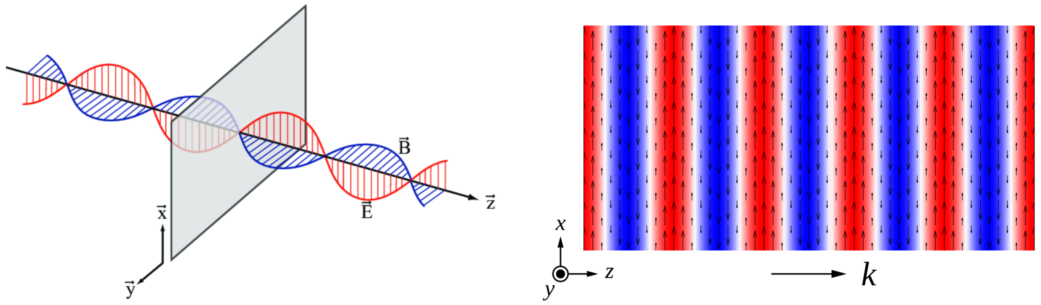

Figure 6.1: Electromagnetic plane wave. Left: Schematic representation. Right: Field representation with electric field shown by arrows and magnetic field shown by color (red for positive values and blue for negative values).

The phase $(-kz+\omega t)$ varies when space and/or time varies. One can track traveling speed of a constant phase, *i.e.* $-kz+\omega t = \text{Const}$. It requires the overall phase change owing to a displacement in space Δz and a time lapse Δt should be zero, *i.e.*

$$-k\Delta z + \omega\Delta t = 0. \tag{6.26}$$

One sees *phase velocity* of wave traveling as

$$\frac{\Delta z}{\Delta t} = \frac{\omega}{k} = \frac{\omega}{\omega/v} = v = [\text{definition following Eq. 6.15}] = \frac{1}{\sqrt{\epsilon\mu}} = \frac{1}{\sqrt{\epsilon_r\mu_r}}\frac{1}{\sqrt{\epsilon_0\mu_0}}. \tag{6.27}$$

In free space, one has $\epsilon_r = 1$ and $\mu_r = 1$. v is then simply $\frac{1}{\sqrt{\epsilon_0\mu_0}}$. Maxwell found that this calculated velocity is rather close to the measured light speed[2]. Thereby, he claimed "light is electromagnetic wave"!

It is worth commenting the factor $\sqrt{\epsilon_r\mu_r}$ in Eq. 6.27. It determines the factor of slowing down of electromagnetic wave velocity in a medium, as compared to that in vacuum. For light, which has frequency over 300 THz, the factor has a special meaning. At such a high frequency, materials usually have no magnetic response, *i.e.* $\mu_r = 1$. Therefore the factor becomes $\sqrt{\epsilon_r}$, which we commonly refer to as *refractive index*, or

$$n = \sqrt{\epsilon_r}\ . \tag{6.28}$$

Refractive index defines how much speed of light slows down in a medium, which in turn determines, as we will see in the next chapter, how light is reflected and refracted across a material interface.

6.4 Electromagnetic plane wave, generalized

In the previous section, we have derived time-harmonic electromagnetic plane wave solution based on the assumption that the wave is propagating along z direction and with

[2] Speed of light was measured by English astronomer James Bradley to a good accuracy already in 1729.

electric field directed along x. Generally, a homogeneous medium accommodates plane waves propagating in any directions. The more general expression for plane-wave solution in terms of $\mathbf{E}$ (spatial dependence only) is

$$\boxed{\mathbf{E} = \mathbf{E}_0 \exp(-i\mathbf{k} \cdot \mathbf{r}),} \tag{6.29}$$

and a similar equation for $\mathbf{H}$ field. $\mathbf{k}$ is called *wave vector*; its magnitude is wave number (k) and its direction is wave propagation direction. In Cartesian coordinate, the dot product in exponent can be fully expanded, the above expression becomes

$$\mathbf{E} = \mathbf{E}_0 \exp\left[-i(k_x x + k_y y + k_z z)\right], \tag{6.30}$$

with $k_x^2 + k_y^2 + k_z^2 = k^2$. Upon substituting Eq. 6.29 and the similar expression for $\mathbf{H}$ into time-harmonic Maxwell's equations (Eqs. 6.10-6.13), one can realize that the vectors $\mathbf{E}$, $\mathbf{H}$, and $\mathbf{k}$ form a mutually orthogonal triplet. A plane wave therefore always has electric and magnetic field components transverse to the propagation direction. We call such an electromagnetic wave a *transverse electromagnetic (TEM) wave.*

6.5 Plane-wave properties

6.5.1 Polarization

Polarization of an electromagnetic plane wave is its $\mathbf{E}$ field direction. A simple plane wave has *linear polarization* – the direction of its $\mathbf{E}$ field is oriented in one direction and not changing with respect to position or time. One expresses $\mathbf{E}$ field for such a plane wave, for example, as

$$\mathbf{E} = E_{x0} \exp\left[i(-kz + \omega t)\right] \hat{\boldsymbol{x}}. \tag{6.31}$$

In a more general scenario, a place wave can be superposition of two linearly polarized plane waves *sharing the same wave vector*. For example, apart from one x-polarized plane wave, *i.e.* Eq. 6.31, there can co-exist a y-polarized plane wave a constant phase difference $\delta\phi$ to the x-polarized plane wave component. So the overall plane wave is written as

$$\mathbf{E} = E_{x0} \exp\left[i(-kz + \omega t)\right] \hat{\boldsymbol{x}} + E_{y0} \exp\left[i(-kz + \omega t + \delta\phi)\right] \hat{\boldsymbol{y}}. \tag{6.32}$$

If $\delta\phi = 0$ and $E_{x0} = E_{0y}$, the combination is a plane wave with an amplitude $\sqrt{2}E_{x0}$ and a linear polarization directed 45° with respect to x axis.

Another special case is when phase of the y-component plane wave lags by $\pi/2$, *i.e.* $\delta\phi = -\pi/2$. The combined plane wave is then

$$\mathbf{E} = E_{x0} \exp\left[i(-kz + \omega t)\right] \hat{\boldsymbol{x}} + E_{y0} \exp\left[i\left(-kz + \omega t - \frac{\pi}{2}\right)\right] \hat{\boldsymbol{y}}. \tag{6.33}$$

One can examine time-evolution of $\mathbf{E}$ field at a constant z position, say $z = 0$. The above expression becomes

$$\mathbf{E} = E_{x0} \exp(i\omega t)\hat{\boldsymbol{x}} + E_{y0} \exp\left[i\left(\omega t - \frac{\pi}{2}\right)\right] \hat{\boldsymbol{y}}. \tag{6.34}$$

The instantaneous field is the real part of the phasor. Therefore,

$$\mathbf{E} = E_{x0} \cos(\omega t)\hat{\boldsymbol{x}} + E_{y0} \cos\left(\omega t - \frac{\pi}{2}\right) \hat{\boldsymbol{y}} = E_{x0} \cos(\omega t)\hat{\boldsymbol{x}} + E_{y0} \sin\left(\omega t\right) \hat{\boldsymbol{y}}. \tag{6.35}$$

If $E_{x0} = E_{y0} \equiv E_0$, one observes that the overall vector $\mathbf{E}$ has an amplitude E_0 and its tip rotating anti-clockwise in xy plane as time t increases. The rotation can be characterized

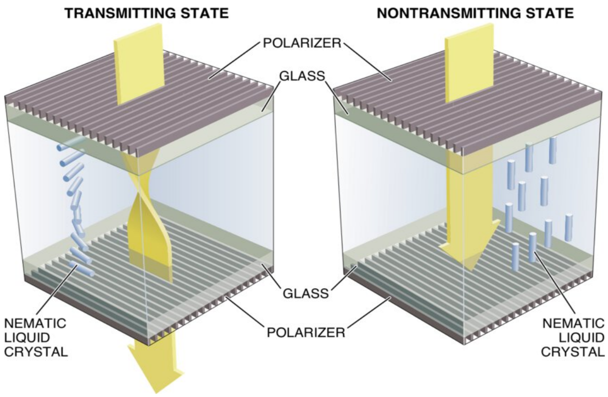

Figure 6.2: Operation principle of an LCD display. Left: ON state (no voltage applied); Right: OFF state (voltage applied). [Source: S.W. Depp and W.E. Howard, Scientific American 268, p. 90 (1993).]

with hand, in this case using one's right-hand — with thumb pointing towards wave propagation direction (z), the rest fingers' curling direction is the $\mathbf{E}$ field's rotation direction. Therefore, the plane wave has *right-hand circular polarization.*

If the y-component plane wave has a leading phase of $\delta\phi = \pi/2$, the rotation of $\mathbf{E}$ will fit to a left-hand gesture. In that case, one says it has *left-hand circular polarization.* Generally, $E_{x0} \neq E_{y0}$ and $\delta\phi \neq \pm\pi/2$, and the overall $\mathbf{E}$ will in general have an *elliptical polarization.* The tip of $\mathbf{E}$ vector traces an ellipse.

Polarizer is a device to purposely stop transmission of a certain linearly polarized electromagnetic wave, while letting the other perpendicularly polarized wave to transmit through. One type of such polarizer is made of an array of fine parallel metal wires, usually embedded in or supported by a glass piece. If an incident wave is polarized along the wire direction, its electric field will excite current in the wires which is then dissipated as heat. For visible light, a cheaper option is to use oriented long polymer molecules, which one can fabricate by simply stretching softened polymer material in one direction. In this case it is polarization/displacement current which causes dissipation. The latter you may find in eyeglasses for watching 3D movies.

There are ways to convert polarization from one linear polarization to another linear polarization, or to circular polarization, and vice versa. A pixel in liquid-crystal display (LCD) is made of a thin layer of liquid crystal (LC) material sandwiched between two orthogonally placed polarizers. Refer to Fig. 6.2. The first polarizer selectively passes one light polarization emitted by backlight. At ON state (left panel), light polarization tends to rotate with the twisting of the LC molecules; light after the LC layer therefore has polarization rotated by 90 degrees, and hence transmits through the second polarizer. At OFF state, a voltage is applied across the LC layer which forces its molecules directed vertically. Light polarization will no longer experience rotation by the LC layer; light will be stopped by the second polarizer.

6.5.2 Group velocity

An electromagnetic plane wave travels with a phase velocity v, which is dependent on medium's permittivity and permeability. The material constants in general vary according to frequency (which is known as material dispersion). Therefore electromagnetic waves at different frequencies travel at different speeds. This effect limits how fast electromagnetic signals can be transmitted. Electromagnetic signals are always carried by a band of frequencies. The envelope of the combined wave can bear information which can then be read by a receiver, given knowledge of a proper key/protocol. The speed of transmission of the envelope is called "group velocity" v_g.

The formula for computing group velocity can be obtained through the following simplified case study. Consider a wave consisting two plane waves of equal amplitudes E_0, one with angular frequency $\omega_0 + \Delta\omega$ and the other $\omega_0 - \Delta\omega$, where $\Delta\omega \ll \omega_0$. Two plane-wave components have also slightly different phase velocities, or correspondingly (since $v = \omega/k$) different wave numbers $k_0 + \Delta k$ and $k_0 - \Delta k$ respectively. Polarizations are the same, which is omitted in the following expression. The combined wave is

$$\begin{aligned} E(z,t) &= E_0 \cos[(\omega + \Delta\omega)t - (k_0 + \Delta k)z] \\ &\quad + E_0 \cos[(\omega - \Delta\omega)t - (k_0 - \Delta k)z] \\ &= 2E_0 \cos(\Delta\omega t - \Delta k z) \cos(\omega_0 t - k_0 z). \end{aligned} \tag{6.36}$$

On RHS, the factor $\cos(\omega_0 t - k_0 z)$ is a quick-varying function (signal-carrying wave, or carrier), while the factor $\cos(\Delta\omega t - \Delta k z)$ is a slow-varying envelope function (modulating wave), representing signal. The velocity of the envelope is determined by tracing a constant phase of the envelope function, *i.e.* $\Delta\omega t - \Delta k z =$ Const. It requires that the overall phase change owing to a displacement in space Δz and a time lapse Δt should be zero, *i.e.*

$$\Delta\omega\Delta t - \Delta k \Delta z = 0. \tag{6.37}$$

The group velocity is then $v_g = \frac{\Delta z}{\Delta t} = \frac{\Delta\omega}{\Delta k}$. In differential form,

$$\boxed{v_g = \frac{d\omega}{dk}.} \tag{6.38}$$

6.5.3 Electromagnetic power

Electromagnetic wave transports power. Ultimately, radiation from the Sun sustains all events and lives on the Earth. Modulation and detection of electromagnetic power is the basis for e.g. optical information communication and mobile networking. Here, we learn how to quantify power of an electromagnetic wave, from its field quantities **E** and **H**.

We start from the two curl equations in Maxwell's equations, *i.e.* Eqs. 6.7 and 6.8. Through taking dot-product of Eq. 6.7 with **H** and respectively Eq. 6.8 with **E**, one obtains

$$\mathbf{H} \cdot (\nabla \times \mathbf{E}) = -\mathbf{H} \cdot \frac{\partial \mathbf{B}}{\partial t}, \tag{6.39}$$

$$\mathbf{E} \cdot (\nabla \times \mathbf{H}) = \mathbf{E} \cdot \mathbf{J} + \mathbf{E} \cdot \frac{\partial \mathbf{D}}{\partial t}. \tag{6.40}$$

Subtracting the two equations, LHS becomes

$$\mathbf{H} \cdot (\nabla \times \mathbf{E}) - \mathbf{E} \cdot (\nabla \times \mathbf{H}) = \nabla \cdot (\mathbf{E} \times \mathbf{H}). \tag{6.41}$$

Among the terms on RHS, one has

$$\mathbf{H} \cdot \frac{\partial \mathbf{B}}{\partial t} = \mathbf{H} \cdot \frac{\partial(\mu \mathbf{H})}{\partial t} = \frac{1}{2}\frac{\partial(\mu \mathbf{H} \cdot \mathbf{H})}{\partial t} = \frac{\partial}{\partial t}\left(\frac{1}{2}\mu H^2\right), \tag{6.42}$$

$$\mathbf{E} \cdot \frac{\partial \mathbf{D}}{\partial t} = \mathbf{E} \cdot \frac{\partial(\epsilon \mathbf{E})}{\partial t} = \frac{1}{2}\frac{\partial(\epsilon \mathbf{E} \cdot \mathbf{E})}{\partial t} = \frac{\partial}{\partial t}\left(\frac{1}{2}\epsilon E^2\right), \tag{6.43}$$

and

$$\mathbf{E} \cdot \mathbf{J} = \sigma E^2. \tag{6.44}$$

Therefore,

$$\nabla \cdot (\mathbf{E} \times \mathbf{H}) = -\frac{\partial}{\partial t}\left(\frac{1}{2}\epsilon E^2 + \frac{1}{2}\mu H^2\right) - \sigma E^2. \tag{6.45}$$

Take volume integral and use divergence theorem for LHS. One obtains

$$\oint_S (\mathbf{E}\times\mathbf{H})\cdot d\mathbf{s} = -\frac{\partial}{\partial t}\int_V \left(\frac{1}{2}\epsilon E^2 + \frac{1}{2}\mu H^2\right) dv - \int_V \sigma E^2 dv. \tag{6.46}$$

The terms on RHT have clear physical meanings. The first term on RHS is decrease of energy stored in electric and magnetic fields in an enclosed volume per unit time. The second term on RHS is ohmic power loss owing to conducting current in the volume, if conductivity is non-zero. Consequently, the surface integral on LHS must correspond to sum of power losses in the volume. By reasoning, $\mathbf{E}\times\mathbf{H}$ must be *surface power density* flowing outwards from the volume, according to conservation of energy. We define the vector cross product as *Poynting vector* $\mathscr{P}$, as

$$\boxed{\mathscr{P} = \mathbf{E}\times\mathbf{H}. \qquad (\text{unit: W/m}^2)} \tag{6.47}$$

In general, $\mathscr{P}$, like $\mathbf{E}$ or $\mathbf{H}$, has a time-dependent value. However, the oscillations occur so fast such that one cares more for *time-averaged Poynting vector* $\mathscr{P}_{av}$. If an electromagnetic wave has oscillating period T, $\mathscr{P}_{av}$ can be expressed as

$$\boxed{\mathscr{P}_{av} = \frac{1}{T}\int_0^T \mathscr{P}\, dt.} \tag{6.48}$$

Using *phasor expressions* of $\mathbf{E}$ and $\mathbf{H}$ fields, after a purely mathematical derivation, one can arrive at an very simple expression of $\mathscr{P}_{av}$ as

$$\mathscr{P}_{av} = \frac{1}{2}\Re(\mathbf{E}\times\mathbf{H}^*), \qquad (\mathbf{E} \text{ and } \mathbf{H} \text{ are phasors}) \tag{6.49}$$

where * is complex conjugate.

$\mathscr{P}_{av}$ can be computed on any surface where an electromagnetic field exists. One can then integrate $\mathscr{P}_{av}$ over the surface area (through dot product) to calculate total power flowing through the surface (unit W), *i.e.*

$$\boxed{P = \int_S \mathscr{P}_{av}\cdot d\mathbf{s}.} \tag{6.50}$$

As a specific example, one can integrate $\mathscr{P}_{av}$ over an spherical surface enclosing a source to find out the power emitted by the source. As another example, for a laser beam in free space or light propagating in a waveguide (e.g. optical fiber), electromagnetic wave is propagating in one direction. One can then compute $\mathscr{P}_{av}$ on a cut-plane normal to the wave propagation direction, and thereby to compute the total beam power.

For a plane wave, its power per unit cross-sectional area, which is usually referred to as *intensity* I, can be readily calculated through the phasor expressions of $\mathbf{E}$ and $\mathbf{H}$ fields (Eqs. 6.22 and 6.23). Since $\mathbf{E}$ and $\mathbf{H}$ are everywhere perpendicular to each other, $\mathscr{P}_{av}$ is

$$\begin{aligned}
\mathscr{P}_{av} &= \frac{1}{2}\Re(\mathbf{E}\times\mathbf{H}^*) \\
&= \frac{1}{2}\Re\{E_0 \exp[i(-kz+\omega t)]\,\hat{\boldsymbol{x}}\} \times \{H_0 \exp[-i(-kz+\omega t)]\,\hat{\boldsymbol{y}}\} \\
&= \frac{1}{2}E_0 H_0 \hat{\boldsymbol{z}} = \frac{1}{2}E_0\frac{E_0}{Z}\hat{\boldsymbol{z}} = \frac{1}{2}\sqrt{\frac{\epsilon}{\mu}}E_0^2\hat{\boldsymbol{z}}. \\
&= \frac{1}{2}\epsilon_0 c\sqrt{\frac{\epsilon_r}{\mu_r}}E_0^2\hat{\boldsymbol{z}}.
\end{aligned} \tag{6.51}$$

Hence wave intensity (unit W/m^2) is

$$\boxed{I = \frac{1}{2}\epsilon_0 c\sqrt{\frac{\epsilon_r}{\mu_r}}E_0^2.} \tag{6.52}$$

6.6 Dipole radiation

We have so far excluded the source terms in the Maxwell's equations and examined how the equations lead to a source-free wave equation with basic plane-wave solutions. With sources present, Maxwell's equations can be used to determine electromagnetic field generated by the sources. One elemental source is a harmonically oscillating current in a very short straight piece of conductor, which we call as an *electric dipole*, or sometimes *Hertzian dipole*.

Derivation of electromagnetic field based on time-varying charge and/or current sources requires introduction of new variables called *electric* and *magnetic potentials*. The latter, magnetic potential, has been intentionally skipped in Chap. 4. In general, a point charge (q, scalar) generates an electric potential distribution (V, scalar) in its surrounding. Spatial variation in V (more specifically ∇V) leads to electric field distribution in space. Magnetic field can be obtained by taking curl of the electric field. In analogy, a very short linear current ($Id\mathbf{l}$, vector) generates a magnetic potential distribution ($\mathbf{A}$, vector, directed along $d\mathbf{l}$) in its surrounding. Curl of $\mathbf{A}$ leads to magnetic field distribution, and electric field can be consequently obtained by taking curl of the magnetic field.

Assume a current source $i(t) = \Re(I \exp i\omega t)$, oriented along z in free space at origin, and its length is much shorter than operating wavelength. The vector potential generated is

$$\mathbf{A} = \frac{\mu_0 Idl}{4\pi}\frac{\exp(-ik_0R)}{R}\hat{\boldsymbol{z}}, \tag{6.53}$$

where R is magnitude of position vector $\mathbf{R}$, and k_0 is free-space wave number.

Magnetic field can be obtained by taking curl of the vector potential, in spherical coordinate. It turns out to have only azimuthal component, as

$$\begin{aligned}\mathbf{H} &= \frac{1}{\mu_0}\nabla\times\mathbf{A} \\ &= -\frac{Idl}{4\pi}k_0^2\sin\theta\left[\frac{1}{ik_0R}+\frac{1}{(ik_0R)^2}\right]\exp(-ik_0R)\ \hat{\boldsymbol{\phi}}.\end{aligned} \tag{6.54}$$

Here θ is the angle between $\mathbf{R}$ and z axis.

Electric field is obtained as

$$\begin{aligned}\mathbf{E} &= \frac{1}{i\omega\epsilon_0}\nabla\times\mathbf{H} \\ &= -\frac{Idl}{4\pi}Z_0k_0^2 2\cos\theta\left[\frac{1}{(ik_0R)^2}+\frac{1}{(ik_0R)^3}\right]\exp(-ik_0R)\ \hat{\boldsymbol{r}} \\ &\quad -\frac{Idl}{4\pi}Z_0k_0^2\sin\theta\left[\frac{1}{ik_0R}+\frac{1}{(ik_0R)^2}+\frac{1}{(ik_0R)^3}\right]\exp(-ik_0R)\ \hat{\boldsymbol{\theta}}.\end{aligned} \tag{6.55}$$

Radiation field by an electric dipole is shown in Fig. 6.3. At a very distant place, *i.e.* $R \gg \lambda/2\pi$ or $k_0R \gg 1$, one considers only the $1/(ik_0R)$ term in the brackets. The far-field electromagnetic wave is therefore

$$\mathbf{E} = i\frac{Idl}{4\pi}Z_0k_0\sin\theta\frac{\exp(-ik_0R)}{R}\ \hat{\boldsymbol{\theta}}, \tag{6.56}$$

$$\mathbf{H} = i\frac{Idl}{4\pi}k_0\sin\theta\frac{\exp(-ik_0R)}{R}\ \hat{\boldsymbol{\phi}}. \tag{6.57}$$

Electric and magnetic fields are perpendicular to each other, with amplitudes different by a factor of Z_0, and propagate away from the source. The maximum radiation is along $\theta = \pi/2$ (equator) direction, and zero radiation is found along the polar axis. The radiated wave asymptotically approaches a plane wave for a local observer at a far distance.

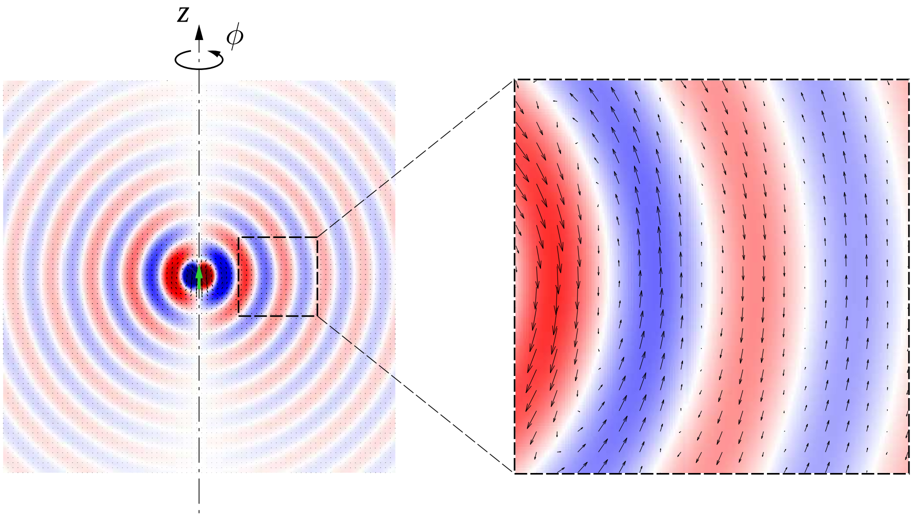

Figure 6.3: Left: Radiation from an electric dipole (green arrow). Electric field (with θ and r components) is shown in arrows. Magnetic field (with only ϕ component) is shown in color (red for positive and blue for negagive). Right: Zoom-in view.

The near-field electromagnetic wave, if one focuses on electric field, is dominated by the $1/(ik_0R)^3$ terms in brackets. At near-field regime, *i.e.* $R \ll \lambda/2\pi$, electric-field strength varies strongly as a function of distance (e.g. increases 8 times when distance is reduced by half).

Radiation power by oscillating electric dipole

Calculate total electromagnetic power emitted by an electric dipole of length d carrying a current I with an angular frequency ω. (Exercise 11-4)

Solution: Use far field for calculation. The phasor expressions for a dipole is shown in Eqs. 6.56&6.57. One can well use substitution $C = \frac{Idl}{4\pi}k_0$ to simplify expressions. From the time-averaged Poynting vector formula, Eq. 6.49, one can calculate the surface power density on a spherical surface at $R = R_0$ enclosing the dipole as

$$\begin{aligned}\mathscr{P}_{av} &= \frac{1}{2}\Re(\mathbf{E}\times\mathbf{H}^*)\\ &= \frac{1}{2}\Re\left\{\left[iCZ_0\sin\theta\frac{\exp(-ik_0R_0)}{R_0}\,\hat{\boldsymbol{\theta}}\right]\times\left[-iC\sin\theta\frac{\exp(ik_0R_0)}{R_0}\,\hat{\boldsymbol{\phi}}\right]\right\}\\ &= \frac{1}{2}\frac{C^2Z_0}{R_0^2}\sin^2\theta\,\hat{\boldsymbol{r}}.\end{aligned}$$

Total power is integral of this surface power density on the spherical surface at $R = R_0$.

$$\begin{aligned}P &= \oint_S \mathscr{P}_{av}\cdot d\mathbf{s} = \oint_S \frac{1}{2}\frac{C^2Z_0}{R_0^2}\sin^2\theta\,\hat{\boldsymbol{r}}\cdot d\mathbf{s}\\ &= \frac{1}{2}\frac{C^2Z_0}{R_0^2}\int_\theta\int_\phi \sin^2\theta R_0^2\sin\theta d\theta d\phi = \frac{1}{2}C^2Z_0\int_\theta \sin^3 d\theta\int_\phi d\phi\\ &= \frac{1}{2}C^2Z_0\cdot\frac{4}{3}\cdot(2\pi) = \frac{4}{3}\pi C^2Z_0.\end{aligned}$$

The result is irrelevant to R_0 value, as long it is relatively far from the dipole.

6.7 Boundary conditions

Similar to electrostatics and magnetostatics, boundary conditions can be obtained by examining the Maxwell's equations in integral form just across a material interface. The line-integral equations lead to conditions for tangential field components; and the surface-integral equations lead to conditions for normal field components. It turns out the boundary conditions are the same as in static cases, i.e.

$$\boxed{\begin{aligned} E_{t1} &= E_{t2}, & (6.58)\\ H_{t1} - H_{t2} &= J_s, & (6.59)\\ D_{n1} - D_{n2} &= \rho_s, & (6.60)\\ B_{n1} &= B_{n2}. & (6.61)\end{aligned}}$$

For lossless dielectric media (no free charge or surface current), one sees that all E_t, H_t, D_n and B_n fields are continuous across interface. Conditions for the other fields (e.g. D_t, B_t, E_n, and H_n) can be obtained through constitutive relations.

For perfect conductors (gold, silver, copper, etc), we assume they have an infinite conductivity. Electric field within them is zero as otherwise there would be infinite current density. In *time-varying situation*, magnetic field is also zero since (E, D) and (B, H) mediate each other. Therefore, all fields in a perfect conductor are zero. In general, an incoming electromagnetic wave cannot penetrate into a perfect conductor; it is reflected backward (principle of household mirror). Field at the interface, according to boundary conditions in Eqs. 6.58-6.61, must fulfill

$$\boxed{\begin{aligned} E_t &= 0, & (6.62)\\ H_t &= J_s, & (6.63)\\ D_n &= \rho_s, & (6.64)\\ B_n &= 0. & (6.65)\end{aligned}}$$

Physically, an effective surface current (J_s) and surface charge (ρ_s) are generated on conductor surface to completely shield the incident electromagnetic wave. In practice, owing to finite conductivities of metals, field can penetrate slightly into metals. The depth of penetration is called *skin depth*, which we do not discuss in depth here.

Chapter 7

Reflection and Refraction

In the previous chapter we have learned basic form of electromagnetic wave propagation in a homogeneous medium. Media usually have finite extent. This chapter describes reflection and refraction (transmission) of an electromagnetic wave upon meeting a planar interface. When applied to waves at optical frequencies (wavelength 400-700 nm), the discussion is relevant to many visual effects that we observe in our daily lives.

Imagine a plane wave propagating in medium 1 (ϵ_1, μ_1) and meets a flat interface into medium 2 (ϵ_2, μ_2). We consider the general case where the propagation direction has an angle of θ_i against the surface normal of the interface plane. We call θ_i as *angle of incidence*. Very importantly, analysis of such a physical problem always starts with identifying the *plane of incidence*. This plane is defined by two vectors — the wavevector of the incident plane wave $\mathbf{k}$ and the surface-normal unit vector of the interface plane $\hat{\boldsymbol{n}}$. The plane of incidence is always perpendicular to the interface plane. Then, we can impose an appropriate Cartesian coordinates accordingly. In the following, we say the interface is $z = 0$ plane and the plane of incidence is $y = 0$ plane. The wave is propagating in xz plane. The structure as well as wave are invariant in y direction, which means we are now dealing with a 2D problem.

Let's see how the problem can be simplified by setting $\frac{\partial}{\partial y} = 0$ in the source-free time-harmonic Maxwell's equations. Only two curl equations are needed (Eqs. 6.12 and 6.13). Each of the curl equations can be broken into three equations in the Cartesian coordinate. That is, Eqs. 6.12&6.12 become respectively

$$\begin{cases} \dfrac{\partial E_z}{\partial y} - \dfrac{\partial E_y}{\partial z} = -i\omega\mu H_x \\ \dfrac{\partial E_x}{\partial z} - \dfrac{\partial E_z}{\partial x} = -i\omega\mu H_y \\ \dfrac{\partial E_y}{\partial x} - \dfrac{\partial E_x}{\partial y} = -i\omega\mu H_z \end{cases} , \qquad \begin{cases} \dfrac{\partial H_z}{\partial y} - \dfrac{\partial H_y}{\partial z} = i\omega\epsilon E_x \\ \dfrac{\partial H_x}{\partial z} - \dfrac{\partial H_z}{\partial x} = i\omega\epsilon E_y \\ \dfrac{\partial H_y}{\partial x} - \dfrac{\partial H_x}{\partial y} = i\omega\epsilon E_z \end{cases} .$$

By getting rid of y dependences, *i.e.* $\frac{\partial}{\partial y} = 0$, one has

$$\begin{cases} -\dfrac{\partial E_y}{\partial z} = -i\omega\mu H_x \\ \dfrac{\partial E_x}{\partial z} - \dfrac{\partial E_z}{\partial x} = -i\omega\mu H_y \\ \dfrac{\partial E_y}{\partial x} = -i\omega\mu H_z \end{cases}, \qquad \begin{cases} -\dfrac{\partial H_y}{\partial z} = i\omega\epsilon E_x \\ \dfrac{\partial H_x}{\partial z} - \dfrac{\partial H_z}{\partial x} = i\omega\epsilon E_y \\ \dfrac{\partial H_y}{\partial x} = i\omega\epsilon E_z \end{cases}.$$

It is noticed the equations can be re-organized into two groups, one containg field components (E_y, H_x, H_z) and the other containing (H_y, E_x, E_z). These two sets of field components represent two light polarizations, each propagating without affecting the other. The two equation groups are

$$\text{TE} \begin{cases} -\dfrac{\partial E_y}{\partial z} = -i\omega\mu H_x, & (7.1) \\ \dfrac{\partial E_y}{\partial x} = -i\omega\mu H_z, & (7.2) \\ \dfrac{\partial H_x}{\partial z} - \dfrac{\partial H_z}{\partial x} = i\omega\epsilon E_y, & (7.3) \end{cases} \qquad \text{TM} \begin{cases} \dfrac{\partial E_x}{\partial z} - \dfrac{\partial E_z}{\partial x} = -i\omega\mu H_y, & (7.4) \\ -\dfrac{\partial H_y}{\partial z} = i\omega\epsilon E_x, & (7.5) \\ \dfrac{\partial H_y}{\partial x} = i\omega\epsilon E_z & (7.6) \end{cases}$$

Refer to Fig. 7.1. Wave with the first set of field components in Eqs. 7.1-7.3 has *transverse-electric (TE)* components (electric field perpendicular to the plane of incidence), and wave with the second set of field components in Eqs. 7.4-7.6 has *transverse-magnetic (TM)* polarization (magnetic field perpendicular to the plane of incidence). An arbitrarily polarized incident plane wave can always be decomposed into TE and TM polarizations. Study of an electromagnetic plane wave crossing an interface can be studied separately according to its polarization.

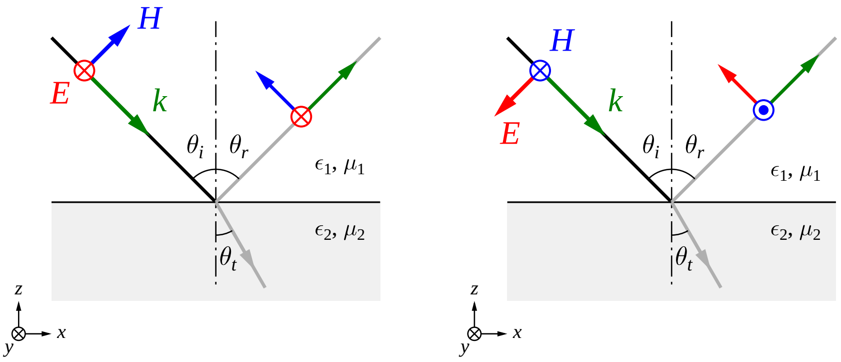

Figure 7.1: Left: Plane wave incidence with TE polarization. Right: TM polarization. Convention: upon reflection, reflected wave has electric field oriented in the same direction as incident wave (best understood at normal incidence).

7.1 TE polarization

A closer look at the equations for TE components tells that H_x and H_z can be written in terms of E_y. If expressions for H_x and H_y are substituted into Eq. 7.3, one has the wave equation based on only E_y

$$\frac{\partial^2 E_y}{\partial x^2} + \frac{\partial^2 E_y}{\partial z^2} + k^2 E_y = 0. \tag{7.7}$$

Refer to Fig. 7.1 (left panel). In medium 1 (upper domain), the incident plane wave is a solution of the equation as

$$E_y^i = E_{y0}^i \exp(-ik_x^i x + ik_z^i z), \tag{7.8}$$

where superscript i is used for denoting incident wave. Without loss of generality, we set interface location at $z = 0$ as well as zero initial phase for plane wave solutions. Both k_x and k_z are positive so the plane wave is propagating along $+x$ and $-z$ directions, conforming to Fig. 7.1. There must be ${k_x^i}^2 + {k_z^i}^2 = k_1^2 = k_0^2 \epsilon_{r1}\mu_{r1}$. Moreover, $k_x^i = k_1 \sin\theta_i$ and $k_z^i = k_1 \cos\theta_i$.

The reflected wave, according to convention indicated in Fig. 7.1, shall have a general plane wave solution as

$$E_y^r = E_{y0}^r \exp(-ik_x^r x - ik_z^r z). \tag{7.9}$$

k_x^r and k_z^r are positive. We have reversed the sign before $(ik_z^r z)$, because the reflected wave has to propagate away from the interface. In medium 2 (lower domain), the transmitted wave is a plane wave solution to Eq. 7.7 as

$$E_y^t = E_{y0}^t \exp(-ik_x^t x + ik_z^t z). \tag{7.10}$$

k_x^t and k_z^t are positive, and there must be ${k_x^t}^2 + {k_z^t}^2 = k_2^2 = k_0^2 \epsilon_{r2}\mu_{r2}$.

E_y is tangential to the interface. Its value has to be continuous across the interface, *i.e.* $(E_y^i + E_y^r)|_{z=0^+} = E_y^t|_{z=0^-}$. From Eqs. 7.8-7.10, one has

$$E_{y0}^i \exp(-ik_x^i x + ik_z^i 0) + E_{y0}^r \exp(-ik_x^r x - ik_z^r 0) = E_{y0}^t \exp(-ik_x^t x + ik_z^t 0). \tag{7.11}$$

or

$$E_{y0}^i \exp(-ik_x^i x) + E_{y0}^r \exp(-ik_x^r x) = E_{y0}^t \exp(-ik_x^t x). \tag{7.12}$$

The expression effectively is asking that E_y has to be continuous across the interface *at all x positions* along the interface. The only chance for this to be satisfied is

$$k_x^i = k_x^r = k_x^t. \tag{7.13}$$

The incident, reflected, and transmitted waves have to vary in phase along x direction. The above condition decides the reflection and transmission angles. First, $k_x^i = k_x^r$ leads to $k_1 \sin\theta_i = k_1 \sin\theta_r$, which means $\theta_i = \theta_r$. For transmission, $k_x^i = k_x^t$ leads to $k_1 \sin\theta_i = k_2 \sin\theta_t$, which can be further written as

$$\sqrt{\epsilon_{r1}\mu_{r1}} \sin\theta_i = \sqrt{\epsilon_{r2}\mu_{r2}} \sin\theta_t. \tag{7.14}$$

In optics, materials are usually non-magnetic, i.e. $\mu_r = 1$ and $\epsilon_r = n^2$, where n is defined as *refractive index*. The above equation is simply the so-called Snell's law

$$\boxed{n_1 \sin\theta_i = n_2 \sin\theta_t.} \tag{7.15}$$

Equation 7.12 can now be simplified to

$$E_{y0}^i + E_{y0}^r = E_{y0}^t. \tag{7.16}$$

E_{y0}^i is known (incidence). To calculate the other two amplitudes, we need another equation, which one can establish using continuity of another tangential field component, H_x. According to Eq. 7.1, H_x can be calculated based on E_y. Imposing $(H_x^i + H_x^r) = H_x^t$ at $z = 0$ and taking $k_x^i = k_x^r = k_x^t \equiv k_x$, one has

$$\frac{k_z^i}{\mu_1} E_{y0}^i - \frac{k_z^r}{\mu_1} E_{y0}^r = \frac{k_z^t}{\mu_2} E_{y0}^t. \tag{7.17}$$

Since $k_z^i = k_z^r$, one has

$$\frac{k_z^i}{\mu_1} \left(E_{y0}^i - E_{y0}^r\right) = \frac{k_z^t}{\mu_2} E_{y0}^t, \tag{7.18}$$

or

$$E^i_{y0} - E^r_{y0} = \frac{k^t_z}{k^i_z}\frac{\mu_1}{\mu_2}E^t_{y0}, \tag{7.19}$$

or (through $k^i_z = k_1 \cos\theta_i = k_0\sqrt{\epsilon_{r1}\mu_{r1}}\cos\theta_i$ and intrinsic impedance $Z = \sqrt{\mu/\epsilon}$),

$$E^i_{y0} - E^r_{y0} = \frac{Z_1 \cos\theta_t}{Z_2 \cos\theta_i}E^t_{y0}. \tag{7.20}$$

From Eqs. 7.16 and 7.20, one can obtain *reflection and transmission coefficients* respectively as

$$\boxed{\begin{aligned} r_{\text{TE}} &= \frac{E^r_{y0}}{E^i_{y0}} = \frac{Z_2\cos\theta_i - Z_1\cos\theta_t}{Z_2\cos\theta_i + Z_1\cos\theta_t}, && (7.21)\\ t_{\text{TE}} &= \frac{E^t_{y0}}{E^i_{y0}} = \frac{2Z_2\cos\theta_i}{Z_2\cos\theta_i + Z_1\cos\theta_t}. && (7.22)\end{aligned}}$$

The two coefficients fulfill the relation $1 + r_{\text{TE}} = t_{\text{TE}}$.

At optical frequencies ($\mu_r = 1$ for both media, $\epsilon_r = n^2$), one has

$$r_{\text{TE}} = \frac{n_1\cos\theta_i - n_2\cos\theta_t}{n_1\cos\theta_i + n_2\cos\theta_t}, \tag{7.23}$$

$$t_{\text{TE}} = \frac{2n_1\cos\theta_i}{n_1\cos\theta_i + n_2\cos\theta_t}. \tag{7.24}$$

7.2 TM polarization

Refer to Fig. 7.1 (right panel). For TM polarization, one has a 2D wave equation similar to Eq. 7.7 based on H_y. The incident plane wave in medium 1 is

$$H^i_y = H^i_{y0}\exp(-ik^i_x x + ik^i_z z). \tag{7.25}$$

The reflected wave, according to convention indicated in Fig. 7.1 (right panel) is

$$H^r_y = -H^r_{y0}\exp(-ik^r_x x - ik^r_z z), \tag{7.26}$$

And the transmitted wave is

$$H^t_y = H^t_{y0}\exp(-ik^t_x x + ik^t_z z). \tag{7.27}$$

The continuity condition for H_y across the interface ($z = 0$) gives rise to

$$H^i_{y0}\exp(-ik^i_x x) - H^r_{y0}\exp(-ik^r_x x) = H^t_{y0}\exp(-ik^t_x x). \tag{7.28}$$

Again, one must have $k^i_x = k^r_x = k^t_x$. And from this equation, one knows the reflection and transmission angles, which are exactly the same as in the TE case. *Snell's law applies regardless of polarization.*

Equation 7.28 is further reduced to

$$H^i_{y0} - H^r_{y0} = H^t_{y0}. \tag{7.29}$$

H^i_{y0} is input and its value is known. To find out the other two amplitudes, we need another equation – the continuity of E_x component at the interface. From Eq. 7.5, one can obtain E_x field expressions in terms of H_y amplitudes. By letting E_x amplitudes at two sides equal at $z = 0$, one has

$$H^i_{y0} + H^r_{y0} = \frac{Z_2\cos\theta_t}{Z_1\cos\theta_i}H^t_{y0}. \tag{7.30}$$

From Eqs. 7.29 and 7.30, one has

$$\frac{H_{y0}^r}{H_{y0}^i} = \frac{Z_2\cos\theta_t - Z_1\cos\theta_i}{Z_2\cos\theta_t + Z_1\cos\theta_i}, \quad \text{and} \quad \frac{H_{y0}^t}{H_{y0}^i} = \frac{2Z_1\cos\theta_i}{Z_2\cos\theta_t + Z_1\cos\theta_i}. \tag{7.31}$$

Reflection and transmission coefficients are defined after ratios between electric fields as,

$$\boxed{r_{\text{TM}} = \frac{E_0^r}{E_0^i} = \frac{H_{y0}^r}{H_{y0}^i} = \frac{Z_2\cos\theta_t - Z_1\cos\theta_i}{Z_2\cos\theta_t + Z_1\cos\theta_i},} \tag{7.32}$$

$$\boxed{t_{\text{TM}} = \frac{E_0^t}{E_0^i} = \frac{Z_2}{Z_1}\frac{H_{y0}^t}{H_{y0}^i} = \frac{2Z_2\cos\theta_i}{Z_2\cos\theta_t + Z_1\cos\theta_i}.} \tag{7.33}$$

The two coefficients fulfill the relation $1 + r_{\text{TM}} = t_{\text{TM}}\frac{\cos\theta_t}{\cos\theta_i}$.

At optical frequencies ($\mu_r = 1$ for both media, $\epsilon_r = n^2$), one has

$$r_{\text{TM}} = \frac{n_1\cos\theta_t - n_2\cos\theta_i}{n_1\cos\theta_t + n_2\cos\theta_i}, \tag{7.34}$$

$$t_{\text{TM}} = \frac{2n_1\cos\theta_i}{n_1\cos\theta_t + n_2\cos\theta_i}. \tag{7.35}$$

7.3 Reflectance and transmittance of beam power

A beam of electromagnetic wave (*e.g.* a laser beam) can in many cases be treated as a plane wave. Refer to Fig. 7.2. The total power of an incident beam will be divided upon reflection and transmission upon meeting a material interface.

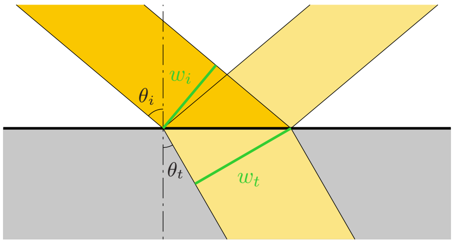

Figure 7.2: Reflectance and transmittance of a light beam.

Ratio between reflected beam power with respect to that of incidence, defined as *reflectance*, is calculated as $R = \frac{I_r S_r}{I_i S_i}$. I is plane-wave intensity (Eq. 6.52), and S refers to beam cross-sectional area. Note that the reflected beam has the same cross-sectional area as the incident beam, or $S_i = S_r$. Recall that plane-wave intensity is $I = \frac{1}{2}\epsilon_0 c\sqrt{\frac{\epsilon_r}{\mu_r}}E_0^2$. Reflected and incident beams are in identical medium. Hence,

$$\boxed{R = \frac{I_r S_r}{I_i S_i} = \frac{{E_0^r}^2}{{E_0^i}^2} = r^2.} \tag{7.36}$$

It is valid for both TE and TM cases.

The transmitted beam has an increased beam width *in one dimension*, from the incident beam width w_i to w_t (see Fig. 7.2), and $\frac{w_t}{w_i} = \frac{\cos\theta_t}{\cos\theta_i}$. Hence, $\frac{S_t}{S_i} = \frac{\cos\theta_t}{\cos\theta_i}$. In addition, the transmitted beam is in a different medium. Therefore, transmittance, or ratio between transmitted beam power to incident power, is

$$\boxed{T = \frac{I_t S_t}{I_i S_i} = \frac{{E_0^t}^2\sqrt{\frac{\epsilon_{r2}}{\mu_{r2}}}}{{E_0^i}^2\sqrt{\frac{\epsilon_{r1}}{\mu_{r1}}}}\frac{\cos\theta_t}{\cos\theta_i} = [\text{for optics}] = t^2\frac{n_2\cos\theta_t}{n_1\cos\theta_i}.} \tag{7.37}$$

If both media are lossless, from conservation of energy, one has $R + T = 1$.

7.4 Normal incidence

At $\theta_i = 0$, the incident wave's electric and magnetic fields are both parallel to the interface. Therefore, they are indifferent from each other. One can also see that the reflection and transmission coefficients become the same (setting also $\theta_t = 0$), as

$$r = \frac{Z_2 - Z_1}{Z_2 + Z_1}, \tag{7.38}$$

$$t = \frac{2Z_2}{Z_2 + Z_1}. \tag{7.39}$$

At optical frequencies ($\mu_r = 1$), one has $Z = \sqrt{\mu_r/\epsilon_r} = 1/n$. Therefore,

$$r = \frac{n_1 - n_2}{n_1 + n_2}, \tag{7.40}$$

$$t = \frac{2n_1}{n_1 + n_2}. \tag{7.41}$$

It is now easier to see that when an electromagnetic wave, or rather light, is reflected by an interface, r can be negative or equivalently the reflected wave acquires a π phase change if the second medium has a higher refractive index.

7.5 Total internal reflection

Here we limit our discussion to optics where materials have no magnetic responses. From Snell's law, $n_1 \sin\theta_i = n_2 \sin\theta_t$, one sees that for two fixed media with $n_1 > n_2$, one shall have in general $\sin\theta_i < \sin\theta_t$ or $\theta_i < \theta_t$ for the equation to be satisfied. However, $\sin\theta_t$ has an upper limit of 1, or equivalently θ_t has an upper limit of $\pi/2$. This critical condition can always be met by an appropriate incident angle, which we call critical incident angle θ_c. Beyond the critical angle, no transmitted light is able to satisfy the required phase matching condition along the interface. Therefore, incident light will be totally reflected. We call such phenomenon as *total internal reflection*. Critical angle can be determined by setting $\theta_t = \pi/2$. The Snell's relation gives $n_1 \sin\theta_c = n_2$. Therefore, one has

$$\boxed{\sin\theta_c = \frac{n_2}{n_1},} \tag{7.42}$$

which is applicable to both TE and TM incidence cases.

7.6 Brewster angle

There exists a special scenario for TM polarization where an incident electromagnetic wave can pass through an interface without any reflection. This happens only at a certain incident angle, which is referred to as *Brewster angle* θ_B. The scenario is fulfilled by setting the numerator in Eq. 7.34 to zero. At optical frequencies ($\mu_r = 1$ for all media), one can use refractive indices. From numerator in Eq. 7.34, one has

$$n_1 \cos\theta_t = n_2 \cos\theta_B. \tag{7.43}$$

Multiplying the equation with Snell's law $n_2 \sin\theta_t = n_1 \sin\theta_B$, one has $\sin\theta_t \cos\theta_t = \sin\theta_t \cos\theta_t$, or $\sin(2\theta_t) = \sin(2\theta_B)$. Therefore $\theta_B + \theta_t = \pi/2$. At Brewster incident angle, transmitted TM light beam would form a right angle to reflected light beam (of zero intensity). Put this condition back to Eq. 7.43, one has

$$\boxed{\tan\theta_B = \frac{n_2}{n_1}.} \tag{7.44}$$

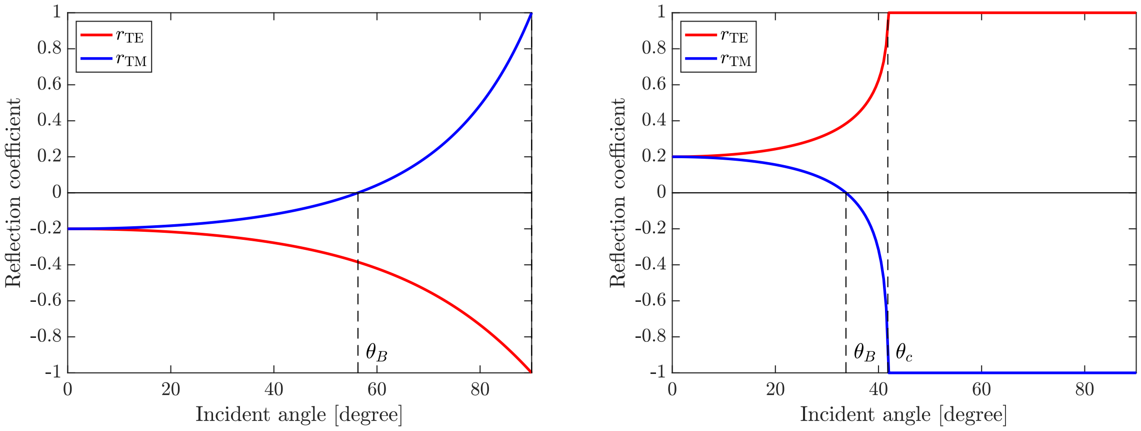

Figure 7.3: Left: Reflection coefficient for light travelling from air ($n_1 = 1$) to glass ($n_2 = 1.5$). Right: From glass to air.

Extended discussion: Figure 7.3 (left panel) shows graphically dependence of reflection coefficient on incident angle when light travels from air (refractive index $n_1 = 1$) to glass ($n_2 = 1.5$). Reflection coefficients for TE and TM polarizations are calculated through Eqs. 7.23 and 7.34, respectively. TE light has negative reflection coefficient; therefore, the reflected light undergoes π phase shift. Reflection coefficient for TM light is negative at incident angle smaller than Brewster angle $\theta_B = 56.3°$ (π phase shift upon reflection) and positive afterwards (no phase change upon reflection).

Reversely, if light travels from glass to air, reflection-coefficient curves for two polarizations are shown in Fig. 7.3 (right panel). Both polarizations experience total internal reflection after incident angle of $\theta_c = 41.8°$. Before the critical angle, TE light has positive reflection coefficient; therefore, no phase shift will occur for reflected light. Reflection coefficient for TM light is positive at incident angle smaller than Brewster angle $\theta_B = 33.7°$ (no phase shift upon reflection) and negative afterwards (π phase change upon reflection).

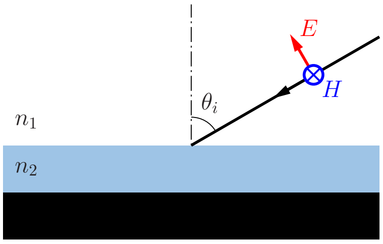

Figure 7.4: Schematic for a light beam incident on a detector.

Light power received by photodetector

Example: See Fig. 7.4. A light beam (plane electromagnetic wave) with indicated polarization is incident on a photodetector at an incident angle $\theta_i = 60°$. The light transmits from air (index $n_1 = 1$) to a lossless glass protecting layer (blue color, index $n_2 = 1.5$). The refracted light then gets completely absorbed by the detecting layer (black color). The incident light has an electric field amplitude $E_0 = 50$ V/m. Calculate electromagnetic power density detected by the photodetector in unit mW/cm^2. [Re-exam 2019]

Suggested steps: (1) Determine light polarization (TM). (2) Calculate through Snell's law θ_t in the glass layer. (3) Calculate transmission coefficient t, hence transmitted electric-field amplitude. (4) Calculate transmitted light intensity I_t. (5) Calculate surface power density on detecting layer based on I_t and θ_t (beam cross-sectional area spreads out on detecting surface by a factor of $1/\cos\theta_t$.) Final solution: 0.166 mW/cm^2.

7.7 Reflection by perfect conductor

Electromagnetic wave interaction with a perfect conductor surface is an extreme scenario of the reflection/transmission phenomenon discussed above. Perfect conductor ($\sigma = \infty$) has zero intrinsic impedance, *i.e.* $Z = 0$. Let's see why. A conductor (not necessary perfect yet) with permittivity ϵ and conductivity σ can be viewed as a material with complex permittivity. This can be understood from Maxwell's equation 6.13. Presence of an electric field will generate a current density $\mathbf{J} = \sigma\mathbf{E}$. Hence the equation becomes

$$\nabla \times \mathbf{H} = \sigma\mathbf{E} + i\omega\epsilon\mathbf{E} = i\omega\left(\epsilon + \frac{\sigma}{i\omega}\right)\mathbf{E}. \tag{7.45}$$

A conductor therefore has a complex effective permittivity

$$\epsilon_c = \epsilon + \frac{\sigma}{i\omega} = \epsilon\left(1 + \frac{\sigma}{i\omega\epsilon}\right). \tag{7.46}$$

A criteria for being a *good conductor* is $\sigma \gg \omega\epsilon$. Hence, $\epsilon_c \approx \frac{\sigma}{i\omega}$. Intrinsic impedance for a good and even perfect conductor becomes

$$Z = \sqrt{\frac{\mu}{\epsilon_c}} \approx \sqrt{\frac{\mu}{\frac{\sigma}{i\omega}}} = \sqrt{\frac{i\omega\mu}{\sigma}} = [\text{if } \textit{perfect} \text{ conductor, i.e. } \sigma = \infty] = 0. \tag{7.47}$$

The formulas for calculating reflection and transmission coefficients, *i.e.* Eqs. 7.21-7.22 and 7.32-7.33, are still valid in the limiting case of $Z_2 = 0$. One finds reflection and transmission coefficients for an electromagnetic wave incident on a perfect conductor surface simply as

$$r = -1, \qquad t = 0. \tag{7.48}$$

The wave is totally reflected back (with $\theta_r = \theta_i$), acquring a phase shift of π. No field will penetrate into a perfect conductor.

A more detailed analysis through boundary conditions governing field on perfect-conductor surface (Section 6.7) can lead to exact field expressions for reflected wave, given an incident wave. For TE polarization, if the incident wave is (*i.e.* Eq. 7.8)

$$E_y^i = E_{y0}\exp(-ik_x x + ik_z z), \tag{7.49}$$

the reflected wave is

$$E_y^r = -E_{y0}\exp(-ik_x x - ik_z z). \tag{7.50}$$

The total electric field in the upper domain will form a standing wave.

In case of normal incidence ($k_x = 0$, $k_z = k_0$, assuming air for medium 1), one has total electric field

$$E_y = E_y^i + E_y^r = E_{y0}\exp(ik_0z) - E_{y0}\exp(-ik_0z) = 2iE_{y0}\sin(k_0z). \tag{7.51}$$

Appending time-harmonic dependence and taking real part, one has instantaneous standing-wave field as

$$E_y = -2E_{y0}\sin(k_0z)\sin(\omega t). \tag{7.52}$$

Chapter 8

Inteference and Diffraction

"Interference" is a phenomenon that associated with all types of waves, including electromagnetic wave. Interference is a direct result of the *superposition principle* — When multiple waves are present in a common spatial region, total wave amplitude can locally be strengthened or weakened, depending on phase values of the individual incoming waves. A criteria for having interference effect is that each of the input waves should be *coherent*, or in the simplest case, can be described as cosine functions. While interference can occur for dissimilar wave frequencies, the most basic interference phenomena are achieved with waves with the same frequency. Specifically to electromagnetic waves, the input waves shall also have the same polarization. In optics, interference is closely associated with "fringe pattern", or mixture of bright and dark light intensities which one can observe with naked eyes. It was such fringe patterns resulted from sunlight interacting with fine objects that made Thomas Young in early 1800s claim that *light is a kind of wave.* Young's experiments challenged Isaac Newton's century-old description of light as particles, and revived the wave theory of light (which was put forward by Christiaan Huygens earlier in 1678). "Diffraction", as far as this chapter concerns, generally refers to the tendency of broadening of an electromagnetic beam, such as light passing through a narrow slit. As will be shown in this chapter, diffraction is intimately connected to interference, and can be treated as interference of infinite number of point sources. In the last part of the chapter, we will examine how light interacts with a transmissive-type "grating", a periodic arrangement of narrow slits. One will see that the transmitted wave is rather interference of diffracted beams from individual slits.

8.1 Huygens' wave theory of light

Christiaan Hygens formulated intially wave theory of light. Huygens argued that any point at a light wave's wavefront acts as a point source that emits spherical wave. Superposition of these spherical "wavelets" collectively determines wavefronts at a later spatial position. A numerical verification of the *Huygens principle* is illustrated in Fig. 8.1 (left panel).

It turns out that passing light through a fine pinhole or slit serves as a simple yet effective method to get a clean point- or respectively line-like light source. Since almost all sorts of light exhibit certain degree of coherence if observed within a very small spatial region, light passing through a pinhole or narrow slit can be coherent. Based on fine slits, Thomas Young was able to experimentally demonstrate interference effect of sunlight (Fig. 8.1, right panel). A slit on the first screen creates a line-like source with cylindrical wavefront. The two slits on the second screens serve as two *coherent* line sources, which can effectively interfere and create a fringe pattern.

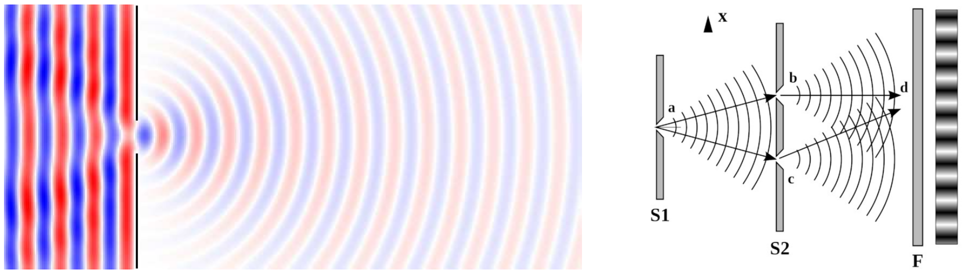

Figure 8.1: Left: numerical verification of Huygens' principle. A plane electromagnetic wave is incident on a metal screen with a linear slit opening (infinite along paper-normal direction). The slit has a vertical width equal to the wavelength. The wave passing through the slit has a nearly circular (more exactly cylindrical) wavefront. Color represents the plane wave's electric field, which is polarized in paper-normal direction. Right: Young's double-slit experiment.

8.2 Interference: two coherence sources

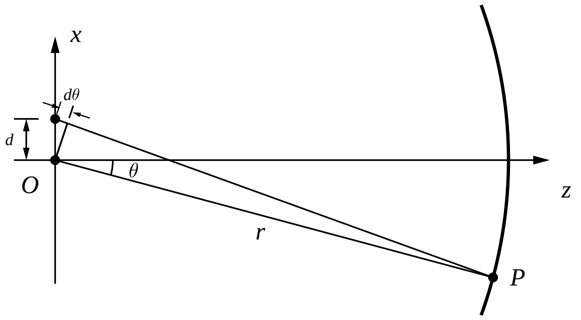

Figure 8.2: Interference of two point (line) sources.

Refer to Fig. 8.2. Two line sources (slits) are separated by a distance of d located at origin (d is very small). Each source emits a cylindrical wave towards a cylindrical screen at distance $z = r$. We assume $r \gg d$. Wave amplitude from each slit, as revealed in Fig. 8.1 (left panel), decreases as the wave propagates. The decrease in amplitude is about the same for both sources. By ignoring the common decrease in amplitude, one can treat waves from the two sources as plane waves of equal amplitudes at P on the screen. In addition, we consider P to be very close to wave propagation axis, or $\theta \approx 0$. Therefore, optical path lengths from the two sources to P differ by $d \sin\theta \simeq d\theta$. Electric field ($E_y$ component only) at P is superposition of fields from two sources, as

$$\begin{aligned} E &= E_0 \exp[i(-kr+\omega t)] + E_0 \exp\{i[-k(r+d\theta)+\omega t]\} \\ &= E_0 \exp[i(-kr+\omega t)] + E_0 \exp[i(-kr+\omega t - kd\theta)] \\ &= E_0 \exp\left[i\left(-kr+\omega t-\frac{kd\theta}{2}+\frac{kd\theta}{2}\right)\right] + E_0 \exp\left[i\left(-kr+\omega t-\frac{kd\theta}{2}-\frac{kd\theta}{2}\right)\right] \\ &= E_0 \exp\left[i\left(-kr+\omega t-\frac{kd\theta}{2}\right)\right]\left[\exp\left(-i\frac{kd\theta}{2}\right)+\exp\left(i\frac{kd\theta}{2}\right)\right]. \end{aligned}$$

The instantensous field is real part of the above, hence

$$\begin{aligned} E &= E_0 \cos\left(-kr+\omega t-\frac{kd\theta}{2}\right)\left[\cos\left(-\frac{kd\theta}{2}\right)+\cos\left(\frac{kd\theta}{2}\right)\right] \\ &= 2E_0 \cos\left(-kr+\omega t-\frac{kd\theta}{2}\right)\cos\left(\frac{kd\theta}{2}\right) \qquad (8.1) \end{aligned}$$

Imagine one places a cylindrical detecting screen with radius r centered to origin. The first cosine function on RHS represents quick oscillation of field in time, giving to all-white exposure to the screen; the second cosine function modulates the field along θ direction, resulting in *fringes*. Note that for very small θ, the cylindrical screen is approximately a flat screen. Wave intensity I is proportional to E^2. On the screen, one has recorded wave intensity

$$I = 4I_0 \cos^2\left(\frac{kd\theta}{2}\right) = 2I_0\left[1+\cos(kd\theta)\right]. \tag{8.2}$$

I_0 is the intensity on the screen due to a single source. Bright fringes have peak intensity $4I_0$, and they happen when the condition $kd\theta = m\cdot(2\pi)$ (m is integer) is fulfilled. Physically, fringe maxima correspond to path-length difference $d\theta$ at multiple of wavelengths. At $\theta = 0$ (when $m = 0$), there is always maximum intensity. The (angular) distance between two fringe maxima is calculated as $kd\Delta\theta = 2\pi$, or

$$\boxed{\Delta\theta = \frac{\lambda}{d}.} \tag{8.3}$$

8.3 Interference: many coherent sources

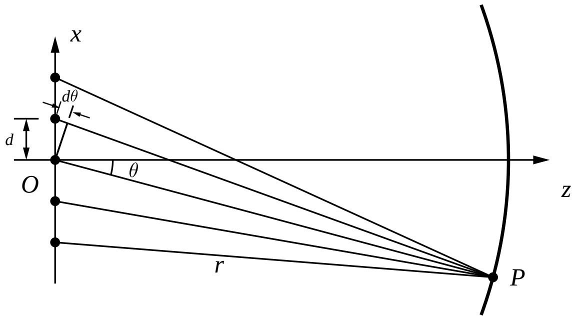

Figure 8.3: Interference of many point (line) sources.

Refer to Fig. 8.3. If there are N such sources with source separation d, one has E phasor field at observation point

$$\begin{aligned} E &= \sum_{n=0}^{N-1} E_0 \exp\left[i(-kr+\omega t+n\delta)\right] \quad (\delta = kd\theta) && (8.4)\\ &= E_0\exp\left[i(-kr+\omega t)\right]\sum_{n=0}^{N-1}\exp(in\delta), && (8.5)\end{aligned}$$

where the geometric series

$$\begin{aligned} \sum_{n=0}^{N-1}\exp(in\delta) &= \frac{\exp(iN\delta)-1}{\exp(i\delta)-1} && (8.6)\\ &= \frac{\exp\left(\frac{iN\delta}{2}\right)}{\exp\left(\frac{i\delta}{2}\right)}\frac{\exp\left(\frac{iN\delta}{2}\right)-\exp\left(-\frac{iN\delta}{2}\right)}{\exp\left(\frac{i\delta}{2}\right)-\exp\left(-\frac{i\delta}{2}\right)} && (8.7)\\ &= \exp\left[\frac{i(N-1)\delta}{2}\right]\frac{\sin\left(\frac{N\delta}{2}\right)}{\sin\left(\frac{\delta}{2}\right)}. && (8.8)\end{aligned}$$

Insert the above into Eq. 8.5. One has

$$E = E_0 \exp\left[i\left(-kr + \omega t + \frac{(N-1)\delta}{2}\right)\right]\frac{\sin\left(\frac{N\delta}{2}\right)}{\sin\left(\frac{\delta}{2}\right)}. \tag{8.9}$$

Taking real part, one has instantaneous field as

$$E = E_0 \cos\left[\left(-kr + \omega t + \frac{(N-1)\delta}{2}\right)\right]\frac{\sin\left(\frac{N\delta}{2}\right)}{\sin\left(\frac{\delta}{2}\right)}. \tag{8.10}$$

Again, imagine there is a cylindrical screen with radius r centered to origin. The cosine function represents quick varying field in time, giving rise to all-white exposure. The ratio between two sine functions modulates the field, giving rise fringes. Intensity distribution on the screen is

$$I = I_0 \frac{\sin^2\left(\frac{N\delta}{2}\right)}{\sin^2\left(\frac{\delta}{2}\right)}. \tag{8.11}$$

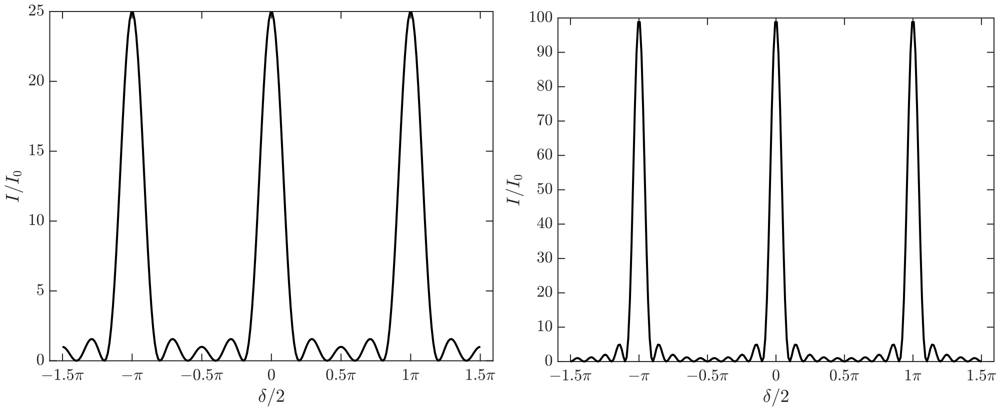

Figure 8.4: Intensity on a cylindrical screen for $N = 5$ (left) and $N = 10$ (right).

In the case of $N = 5$ (or 10), we have the fringes (dependence on $\delta/2$) shown in Fig. 8.4. There are high-intensity peaks (bright fringes) separated by relatively weaker peaks and dark fringes. The highest peaks occur when denominator in Eq. 8.11 becomes zero, or $\frac{\delta}{2} = m\pi$ (m is an integer). The peak intensities are not infinite because the numerator also becomes zero at these conditions. By taking a limit, one sees that the peak intensity is $N^2 I_0$. From the above condition, one gets the angular distance between two highest peaks as (again)

$$\boxed{\Delta\theta = \frac{\lambda}{d}\,.} \tag{8.12}$$

8.4 Diffraction: wave through a slit

An electromagnetic wave can't be limited to an arbitrarily small cross-sectional size; it tends to spread out as a result of diffraction. Refer to Fig. 8.5 (left panel), the classic example for diffraction is the single-slit experiment, where a *coherent* light passing through a narrow slit is found to diverge. The phenomenon can be seen with our naked eyes if the slit width is comparable to the light wavelength (a rule of thumb is $< 100\lambda$). The beam after the slit is no longer a clear image of the slit, but predominantly a high-intensity line, plus some weaker lines on two sides. The dominant high-intensity line has a width larger than slit width — the beam is diverging with a *divergence angle* θ_d. Similar phenomenon occurs when light passes through a small pinhole (Fig. 8.5, right panel). Generally speaking, the more tightly one wants to limit a coherent light beam, the larger will be the divergence angle.

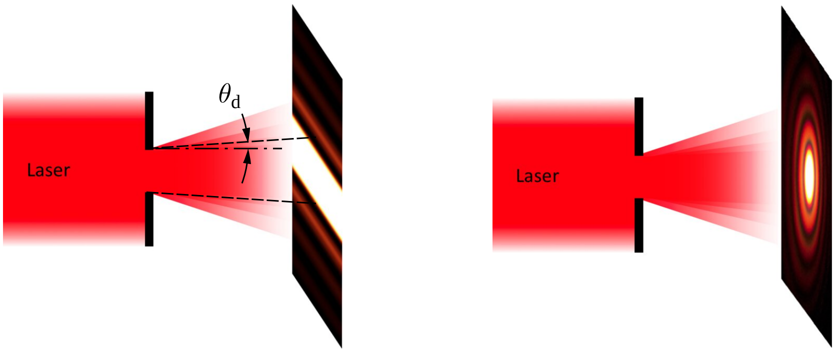

Figure 8.5: Left: Light diffraction by a narrow slit. Right: Diffraction by a pinhole.

Physically, light at the slit opening can be treated as infinite number of coherent point sources (Huygen's principle). Therefore, wave transmission through a single slit can be treated as interference of waves emitted by these point sources. This equivalence allows us to proceed with the solutions which we obtained in the previous section. The problem setting is now: there exist linearly arranged N ($N \to \infty$) point sources with nearest-neighbor separation d ($d \to 0$) such that the product $Nd = D$, where D is the slit width. With this setting, intensity at the detecting screen due to a *single* source I_0 approaches zero. Along $\theta = 0$ direction, one expects the maximum intensity, *i.e.* $I_m = N^2 I_0$, owing to constructive interference. Further, since $d \to 0$, one has $\delta \to 0$; therefore $\sin(\delta/2) = \delta/2$. If one continues from Eq. 8.11, one has

$$I_{\text{slit}} = I_0 \frac{\sin^2\left(\frac{N\delta}{2}\right)}{\sin^2\left(\frac{\delta}{2}\right)} = I_0 \frac{\sin^2\left(\frac{N\delta}{2}\right)}{\left(\frac{\delta}{2}\right)^2} = N^2 I_0 \frac{\sin^2\left(\frac{N\delta}{2}\right)}{\left(\frac{N\delta}{2}\right)^2}. \tag{8.13}$$

If one additionally defines $u = \frac{N\delta}{2} = \frac{Nkd\sin\theta}{2} = \frac{Nkd\theta}{2} = \frac{\pi D\theta}{\lambda}$, one can write the wave intensity after a slit as

$$\boxed{I_{\text{slit}} = I_m \frac{\sin^2 u}{u^2}. \quad \left(= I_m \,\text{sinc}^2 u\right)} \tag{8.14}$$

The numerator tells that there shall be bright and dark fringes, while the denominator damps the overall intensity as u deviates from 0. A plot of the intensity after the slit is shown in Fig. 8.6. It is not a sharp image of the slit — boundaries are blurred, and there are fringes extending to both sides.

The main intensity peak has half-width $u = \pi$ or $\frac{\pi D\theta}{\lambda} = \pi$, which defines the divergence angle of the beam as

$$\boxed{\theta_d = \frac{\lambda}{D}.} \tag{8.15}$$

Smaller aperture and longer wavelength lead to larger divergence. This value for divergence angle is exact for line slits, and gives a good approximation to aperture in other shapes, or even beam in other intensity profiles (we have only considered input light with constant amplitude in the aperture).

8.5 Grating: wave through many slits

Grating is an optical device that can, in general, convert a laser beam into several beams propagating in different directions. For a transmissive-type grating, it can simply be made

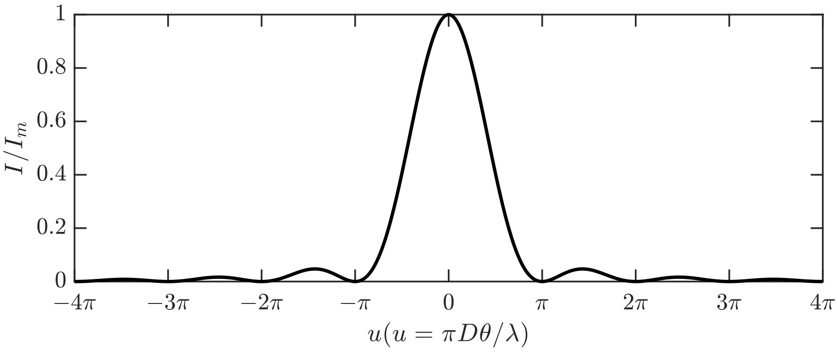

Figure 8.6: Intensity recorded on cylindrical screen for light passing through a slit.

of a linear array of identical slits. So far, we have obtained solutions for light intensity *interferenced* by many point sources, as well as for light intensity *diffracted* by a single slit. We can make use of these knowledges to get solution for light passing through a transmissive grating. The way to do this is simply to replace I_0 in Eq. 8.11 by I_{slit} in Eq. 8.14. The result is

$$I_{\text{grating}} = I_{\text{slit}} \frac{\sin^2\left(\frac{N\delta}{2}\right)}{\sin^2\left(\frac{\delta}{2}\right)} = I_m \frac{\sin^2 u}{u^2} \frac{\sin^2\left(\frac{N\delta}{2}\right)}{\sin^2\left(\frac{\delta}{2}\right)}. \tag{8.16}$$

As in the previous section, $u = \frac{\pi D\theta}{\lambda}$ and $\delta = kd\sin\theta = kd\theta = \frac{2\pi d\theta}{\lambda}$. N is number of slits; D is slit width; and d is slit separation (grating period).

For illustration, we choose $N = 5$ and $D = d/4$. The latter leads to $\delta = 8u$. Light intensity after the grating is as shown in Fig. 8.7. A beam becomes multiple beams propagating in different directions with enhanced peak intensities (maximum $N^2 = 25$ times compared to a single slit case). Its outer envelope is defined by the profile in Fig. 8.6 (diffraction by a single slit), and the fine high-intensity beams are owing to interference by N slits as in Fig. 8.4 ($N = 5$, left panel). Figure. 8.8 shows diffraction of a laser beam as viewed from side. Only the zero$^{\text{th}}$-order (straight through) and first-order diffracted beams are visible. The angular separation between two beams is the same as that for interference by multiple point sources, *i.e.*,

$$\boxed{\Delta\theta = \frac{\lambda}{d}\,.} \tag{8.17}$$

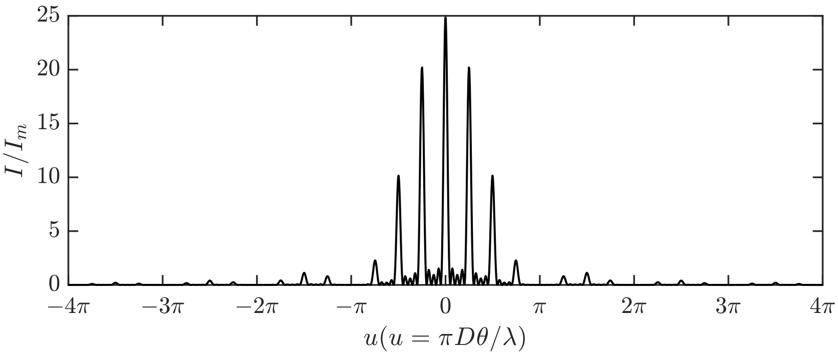

Figure 8.7: Intensity after a grating recorded on cylindrical screen for $N = 5$ and $D = d/4$.

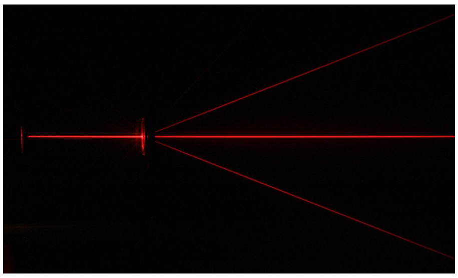

Figure 8.8: Diffraction of a laser beam by a grating. (Source: Wikipedia)

Chapter 9

Waveguides

9.1 Introduction

Modern society is built upon electrification and information technology, both of which are based on transportation of electromagnetic waves. Electromagnetic radiation from their sources, such as antennas, usually goes in all directions. Even a coherent laser beam tends to diverge as it propagates, which can be further adversely affected by diffraction upon meeting obstacles. *Waveguide* is an indispensable device for delivering electromagnetic wave from its source to a specified destination with minimal loss. Depending on operation frequency, electromagnetic waveguides can appear differently. The exact theory for analyzing them can also vary: at extremely low frequency, one can use circuit theory, where one solves for voltage and current; at high frequency, one shall resort to field theory, where one solves for electric and magnetic fields. In Fig. 9.1, four representative types of electromagnetic waveguides are illustrated. Generally, waveguides can have different geometry or material designs in their *transverse* or cross-sectional domain, but are uniform along their *longitudinal* or axial direction.

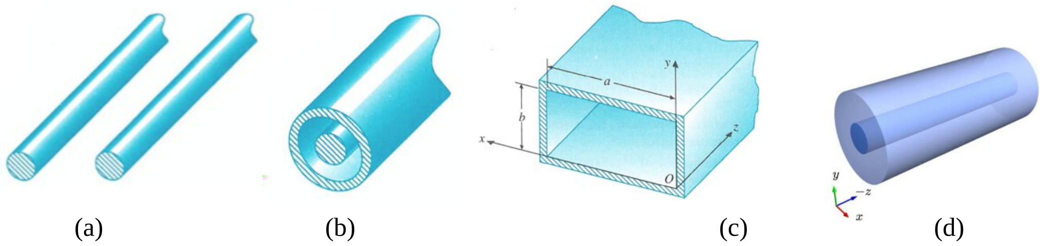

Figure 9.1: (a) Two-conductor cable. (b) Co-axial waveguide. (c) Hollow metallic waveguide. (d) Optical fiber.

The waveguides shown in panels (a) and (b) have two conductors; they are therefore referred to as two-conductor waveguides. From electromagnetic theory, two-conductor waveguides support transverse-electric-magnetic (TEM) mode (detailed discussion is omitted in this text). This property leads to well-defined voltage value at any position of the waveguide. Furthermore, TEM mode has no lower limit of operating frequency (so-called *cut-off frequency*), which allows such waveguides to have dimensions much smaller compared to guided electromagnetic wavelength. The two-conductor waveguide shown in panel (a) is commonly used for low-frequency (up to few kHz) applications. They can be found in ubiquitous power-line networks for transporting electricity (normally 50 Hz), as well as traditional telephone networks ($<$ 3400 Hz). We know already that, such a two-pole cable, when connected to source and a load, makes a simple electric circuit. At 50 Hz, electromagnetic wavelength is 6000 km, which is much larger than the circuit size. When this condition fulfills, we say the circuit is operating at static limit. That is, one can use static field as well as circuit theory to analyze the circuit. At each time instance, the load (or neutral) line in household electrical network has everywhere identical voltage. Drawbacks of this waveguide are two-fold. First, the propagating mode has electric and magnetic fields exposed in the cable surroundings; the field can interact with nearby objects and incur power loss. Second, at even higher frequencies, alternating current in the two wires can simply radiate out electromagnetic wave. Panel (b) shows another type of two-conductor waveguide, which is commonly referred to as co-axial cable. This particular design allows electric field to be well confined in the dielectric spacer between the center conductor and the outer cylindrical conductor shell. Electromagnetic shielding not only prevents signal from being disturbed by nearby objects, but also prevents potential radiation loss at high frequencies. Therefore, co-axial cables can be used in circuits that operate in a frequency up to a few GHz. Applications include radio, TV, and scientific instruments, where requirement on data rate is relatively high. Note that electromagnetic wavelength can now be in centimeters, *i.e.* smaller than circuit dimension; voltage and current in a circuit are no longer of the same values but propagate like waves in the circuit. A special circuit theory, called transmission-line theory (not discussed in this text), is developed to model such networks.

The waveguide shown in panel (c) is called *hollow metallic waveguide*, which is the main topic of this chapter. It is made of a single conductor. Such waveguide is used in typical microwave circuits with frequency up to 300 GHz. The co-axial cable in panel (b) suffers from heavy resistive loss at these high frequencies. Unlike waveguides in panels (a)&(b) where there is a well-defined voltage between the two conductors at any longitudinal position, a hollow metallic waveguide, being a single conductor, has no well-defined voltage at a longitudinal position. Instead of signal input by biasing a voltage on two conductors for waveguides in panels (a) or (b), a high-frequency signal is first converted to electromagnetic wave through antenna, and the generated wave is coupled into such a hollow metallic waveguide for signal transportation. Although one can straightforwardly understand wave transportation in such a waveguide as wave reflection by all “mirror-like” inner facets, a quantitative understanding requires complete field analysis. From electromagnetic theory, this waveguide does not support a TEM mode. As a direct consequence, the waveguide has a lower limit in operating frequency. As we will learn in next section, a hollow metallic waveguide has its cross-sectional size comparable to or larger than the operating electromagnetic wavelength.

The waveguide shown in panel (d) is an optical fiber. It is an all-dielectric waveguide developed for transmitting electromagnetic wave at optical frequencies ($>$ 150 THz). At these frequencies, mirrors made of metallic materials are too lossy for long-distance communication. Instead, one relies on the phenomenon of “total internal reflection” between two dielectric materials to confine and channel optical waves. In order to do so, the core material has a slightly higher refractive index compared to the outer cladding material. Usually the core dimension is a few times as large as the operating wavelength (in the material). Operating frequency is limited by the transparency range of the material used. Telecommunication networks use optical wavelengths close to 1.55 μm, where silica-based fibers have the lowest loss ($\sim$ 0.2 dB/km). Often, a field theory is required to completely understand behavior of such optical fibers.

9.2 Rectangular hollow metallic waveguide

As mentioned, the principle for electromagnetic wave guidance in a rectangular hollow metallic waveguide is wave reflection by metal surface. We assume metal considered here is perfect conductor. In general, a confined wave will experience repeated reflections by up to four mirrors at lateral sides. From Section 6.7 as well as 7.7, we learned that a perfect conductor requires specific boundary conditions for electromagnetic field at its surface. The boundary conditions have to be fulfilled at all interfaces where wave reflections occur. The requirement on boundary conditions dictates that only certain specific propagation "modes" exist in such a waveguide. What this implies in practice is that, when one couples an electromagnetic wave into a rectangular hollow metallic waveguide, one has to carefully choose polarization, plane of incidence, and incident angle in order for the wave to propagate in the waveguide. Propagating modes in a rectangular waveguide can be categorized into two groups: *transverse-electric* (TE) modes and *transverse-magnetic* (TM) modes. Refer to geometric and coordinate setups in Fig. 9.2. We elaborate below formation and properties of the so-called *fundamental TE mode*, physically corresponding to a (laterally) standing wave formed from reflections by two opposite mirrors (left and right surfaces). For a reason that is to be clarified in the following sub-section, this particular mode is referred to as TE_{10} mode.

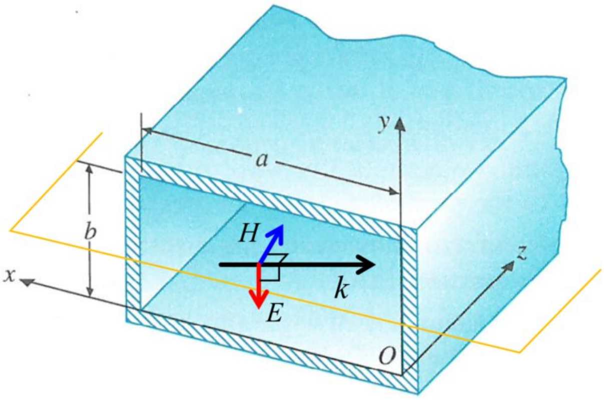

Figure 9.2: A rectangular hollow-metallic waveguide.

9.2.1 TE_{10} mode

Mode field derivation

Refer to geometrical parameters in Fig. 9.2. We can take a xz cut-plane through the center ($y = b/2$) of a rectangular hollow-metallic waveguide, and treat the cut-plane as incident plane for a plane wave (more appropriately "wave propagation plane"). Furthermore, we send out along the plane a TE-polarized plane wave, *i.e.* with E_y electric field component and H_x and H_z magnetic field components. Wave frequency ω, and in turn wave number $k = \frac{\omega}{c}$, are known. Note that given this polarization, boundary conditions at top and bottom side walls (surfaces at $y = 0$ and b) are always satisfied, because only normal electric field and tangential magnetic field with respect to these two surfaces exist. The remaining task is to find out under what conditions (incident angle θ_i, and in turn $k_z = k \sin\theta_i$ and $k_x = k\cos\theta_i$) the boundary conditions at left and right side walls (surfaces at $x = 0$ and a) can be satisfied.

If we single out the above-mentioned wave propagation plane, we have wave-reflection scenario shown in Fig. 9.3. Incident plane wave in phasor can be written as

$$E_i = E_0 \exp(ik_x x - ik_z z + i\omega t), \tag{9.1}$$

where the incident wave vector has been decomposed into two components, *i.e.* $k_z = k\sin\theta_i$ and $k_x = k\cos\theta_i$. The reflected plane waves shall be

$$E_r = -E_0 \exp(-ik_x x - ik_z z + i\omega t). \tag{9.2}$$

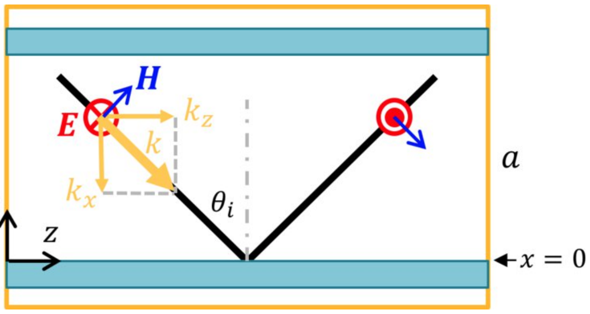

Figure 9.3: xz cut plane on the rectangular hollow metallic waveguide.

Notice the field amplitude has be flipped in sign (recall Subsection 7.7), which ensures tangential electric field is zero at metal interface ($x = 0$). The total electric field is

$$E = E_i + E_r = E_0 \left[\exp(ik_x x) - \exp(-ik_x x)\right] \exp(-ik_z z + i\omega t) \tag{9.3}$$

$$= E_0 \left[2i \sin(k_x x)\right] \exp(-ik_z z + i\omega t). \tag{9.4}$$

Take the real part, one has instantaneous field

$$E = 2E_0 \sin(k_x x) \sin(-k_z z + \omega t). \tag{9.5}$$

Standing wave is formed along x direction, whose spatial dependent is decided by $\sin(k_x x)$, or k_x. One can double-check that the (tangential) electric field is zero at $x = 0$, *i.e.* surface of the bottom perfect conductor. At surface of the upper conductor $x = a$, one must also fulfill the boundary condition of zero tangential electric field. In order to achieve that, we must impose

$$\boxed{k_x a = m\pi. \qquad (m = 1, 2, 3...)} \tag{9.6}$$

In other words, in order for an electromagnetic wave to propagate in such a waveguide, k_x can only take discrete values

$$k_x = \frac{m\pi}{a}. \qquad (m = 1, 2, 3, ...) \tag{9.7}$$

This relation decides the specific k_x values, and in turn the incident angle θ_i as well as (of course) k_z, in order for the wave to propagate in the waveguide. Hence, electric field (y component) for TE modes inside waveguide shall have the form

$$E_y = 2E_0 \sin\left(\frac{m\pi}{a}x\right) \sin(-k_z z + \omega t). \tag{9.8}$$

When $m = 1$, the field has "1" π phase change along x direction inside the waveguide, while "0" phase change happens along y coordinate. By convention, we use the number of π phase changes along two lateral-coordinate directions as subscripts to denote each mode. Therefore, we have TE$_{10}$ (fundamental mode), TE$_{20}$ mode, etc.

The magnetic field can be obtained similarly, except that it has both x and z components. The incident and reflected magnetic fields are

$$\mathbf{H}_i = H_0(-\sin\theta_i \hat{\boldsymbol{x}} - \cos\theta_i \hat{\boldsymbol{z}}) \exp(ik_x x - ik_z z + i\omega t), \tag{9.9}$$

$$\mathbf{H}_r = H_0(\sin\theta_i \hat{\boldsymbol{x}} - \cos\theta_i \hat{\boldsymbol{z}}) \exp(-ik_x x - ik_z z + i\omega t). \tag{9.10}$$

where the amplitude $H_0 = \frac{E_0}{Z_0}$ with free-space impedance $Z_0 = \sqrt{\frac{\mu_0}{\epsilon_0}}$. The x- and z-components of magnetic field can be summed. After taking real part and with the help of trigonometric identity $\cos(\alpha + \beta) = \cos\alpha\cos\beta - \sin\alpha\sin\beta$, they are respectively

$$H_x = 2H_0 \sin\theta_i \sin(k_x x) \sin(-k_z z + \omega t), \tag{9.11}$$

$$H_z = -2H_0 \cos\theta_i \cos(k_x x)\cos(-k_z z + \omega t). \tag{9.12}$$

The amplitude for H_x can be re-written as $2H_0 \sin\theta_i = 2\frac{E_0}{Z_0}\frac{k_z}{k} = \frac{2E_0}{\mu_0}\frac{k_z}{\omega}$ (note $k = k_0 = \frac{\omega}{c}$, and both Z_0 and c can be expressed in ϵ_0 and μ_0). H_x and H_y can be re-written using E_0 as

$$H_x = \frac{2E_0}{\mu_0}\frac{k_z}{\omega}\sin(k_x x)\sin(-k_z z + \omega t), \tag{9.13}$$

$$H_z = -\frac{2E_0}{\mu_0}\frac{k_x}{\omega}\cos(k_x x)\cos(-k_z z + \omega t). \tag{9.14}$$

The condition for magnetic field, *i.e.* the normal component of magnetic field (H_x) shall be zero at the boundary of a perfect conductor, is automatically satisfied with the condition identified in Eq. 9.7. The whole wave solution (multiple of them as m can vary) in Eqs. 9.8, 9.13, and 9.14 represents legitimate electromagnetic modes propagating in the rectangular hollow-core metallic waveguide.

Mode field plot

Based on field solutions in Eqs. 9.8, 9.13, and 9.14 one can sketch out TE$_{10}$ mode field in the rectangular hollow metallic waveguide, as shown in Fig. 9.4.

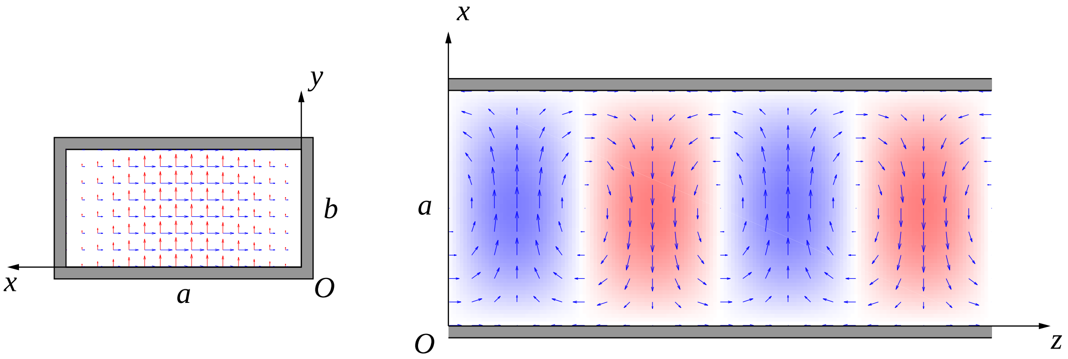

Figure 9.4: TE$_{10}$ mode. Left: Electric (red arrows) and magnetic (blue arrows) fields as seen from cross-sectional plane (xy) of the hollow metallic waveguide. Right: Electric (red-blue color map) and magnetic (blue arrows) fields on xz cut plane. Red color denotes positive value and blue color denotes negative value.

From Fig. 9.4 (left panel), one sees that electric field E_y has different magnitude at different lateral x positions. If one integrates E_y along y at different x positions, one has different voltages. That is, wave in a single-conductor waveguide can't be represented by V (and I) parameters.

One also notices from Fig. 9.4 (left panel) that, surface-normal electric field exists at top and bottom metal surfaces ($y = 0, b$). This is due to effective surface charge generated by the mode field. One can calculate surface charge density through Gauss's law based on the electric field. Tangential magnetic field exists at all four sides of the waveguide. From the field, one can calculate surface current density through Ampère's law. Alternatively, one can readily calculate such surface charge or current densities through the boundary conditions for electromagnetic waves listed in Section 6.7, more specifically Eqs. 6.63&6.64.

Modal dispersion and cut-off frequency

One sees that the waveguiding condition in a rectangular hollow metallic waveguide is defined by Eq. 9.7. Combined with the relation

$$k_x^2 + k_z^2 = k^2, \tag{9.15}$$

where $k = \omega/c$, one has

$$\left(\frac{m\pi}{a}\right)^2 + k_z^2 = \left(\frac{\omega}{c}\right)^2. \quad \text{(dispersion relation, TE}_{n0}\text{ modes)} \tag{9.16}$$

Given a fixed geometry (a) and mode order (m), one have a definite relation between electromagnetic wave frequency ω versus k_z. For waveguides, k_z is usually referred to as *propagation constant*. This relation is called *dispersion relation*. When plotted out as a curve, it is called dispersion curve. Below in Fig. 9.5 we plot dispersion curves for first three TE modes guided in a waveguide with $a = 3$ cm.

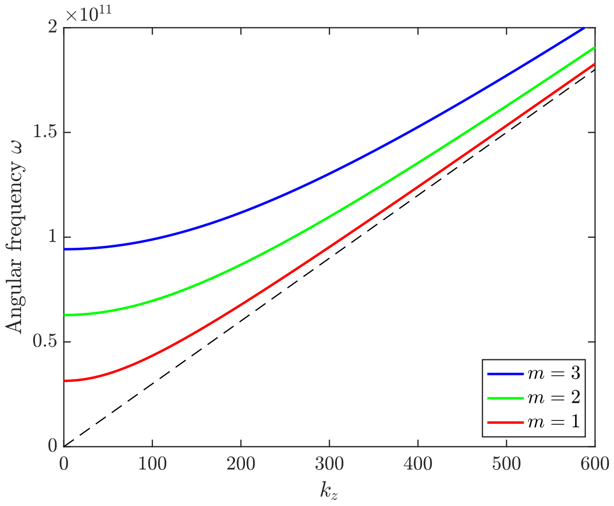

Figure 9.5: First three TE modes of a rectangular waveguide with $a = 3$ cm. Dashed line is the dispersion relation for a plane wave propagating in free space (no waveguide) along z direction (usually referred to as light line).

One sees that there are cut-off frequencies for each guided mode. According to the dispersion relation Eq. 9.16, the lowest frequency happens when $k_z = 0$; or physically, wave is not propagating along z (just along x, forming standing wave). According to Eq. 9.16, the cut-off frequency for TE$_{m0}$ mode is

$$\omega_c = \frac{m\pi c}{a}. \tag{9.17}$$

When translated to wavelength, the cut-off wavelength is

$$\lambda_c = \frac{2a}{m}. \tag{9.18}$$

TE$_{10}$ mode has the lowest cut-off frequency among all modes. For that reason, one can refer to the cut-off frequency of TE$_{10}$ mode as the waveguide's cut-off frequency. Below this frequency, electromagnetic wave simply can't propagate in a hollow metallic waveguide. Knowledge about cut-off frequencies is important when one wants to design a waveguide with single-mode operation at a particular frequency. Single-mode waveguide has a higher data transmission speed compared to multi-mode waveguides.

Cut-off frequency of hollow-metallic waveguide

Refer to Fig. 9.2. In the case of $a = 3$ cm and $b = 1$ cm, calculate cut-off frequency of its TE$_{10}$ mode.

Solution: Direction application of Eq. 9.17 leads to $\omega_c = \frac{1\cdot\pi c}{a} = 31.4$ GHz. $\lambda_c = 6$ cm (or waveguide width a is half of wavelength). b is not used.

Phase and group velocities

Phase velocity of a mode is defined now by k_z. That is

$$v = \frac{\omega}{k_z}. \tag{9.19}$$

In comparison, plane-wave has phase velocity $v = \omega/k$. Since k_z is projection of $\mathbf{k}$ along z direction, guided modes have larger phase velocities as compared to plane-wave velocity at the same frequency. Information on phase velocity is contained in the dispersion relation, as plotted in Fig. 9.5.

Group velocity for each mode is calculated as

$$v_g = \frac{d\omega}{dk_z}. \tag{9.20}$$

It can be interpreted as slope of the dispersion curves in Fig. 9.5. Although the propagating modes have higher phase velocities than speed of light, their group velocities are always less than speed of light. For each mode, at low-frequency limit (close to cut-off frequency), $v_g \to 0$; at very high frequency, v_g approaches speed of light.

9.2.2 TE$_{mn}$ modes

In the above subsection, we have discussed TE$_{m0}$ modes with a focus on the fundamental TE$_{10}$ mode. These modes correspond to waves propagating in xz plane with an E_y electric field component (TE polarization). In general, there are other TE modes, called TE$_{mn}$ modes, where the subscripts m and n are integers referring to the number of π phase changes along x and y directions, respectively. Among these modes, TE$_{0n}$ modes are rather easy to picture (refer to Fig. 9.2): one uses instead yz cut-plane as wave propagation plane with E_x electric field component (still TE polarization). Then it is the waveguide height b that decides the fulfillment of boundary conditions. The condition for lateral wave number, dispersion relation, and cut-off frequency are the same as those in Eqs. 9.7, 9.16, and 9.17, except one shall replace k_x by k_y, a by b (and m by n). In case of waveguides with $b < a$ (as in Fig. 9.2), the cut-off frequencies for TE$_{0n}$ modes are higher than those for TE$_{m0}$ modes.

For general TE$_{mn}$ modes ($m \neq 0, n \neq 0$), the wave propagation plane can be considered as an intermediate plane between xz and yz planes. The wave vector has both k_x and k_y components, besides k_z. Still, the incident wave can have a TE polarization; but now the electric field has both x and y components. The propagating modes shall in general fulfill

$$\boxed{k_x = \frac{m\pi}{a}, \quad \text{and} \quad k_y = \frac{n\pi}{b}. \qquad (m, n = 0, 1, 2, 3, ...)} \tag{9.21}$$

Note m and n can't be zero at the same time. Dispersion relation can be obtained from

$$\boxed{\left(\frac{m\pi}{a}\right)^2 + \left(\frac{n\pi}{b}\right)^2 + k_z^2 = \left(\frac{\omega}{c}\right)^2. \qquad \text{(dispersion relation, TE}_{mn}\text{ modes)}} \tag{9.22}$$

Cut-off frequencies for these modes are therefore

$$\boxed{\omega_c = \pi c\sqrt{\left(\frac{m}{a}\right)^2 + \left(\frac{n}{b}\right)^2}. \qquad \text{(cut-off frequency, TE}_{mn}\text{ modes)}} \tag{9.23}$$

We have so far borrowed analogy to the definition of "TE polarization" for plane wave interaction with a planar interface to call the modes discussed as TE modes. A more common way to tell a TE mode is that such a mode has electric field in waveguide's transverse (cross-sectional) domain, *i.e.* no component in waveguide's longitudinal direction.

9.2.3 TM$_{mn}$ modes

Besides TE modes, there exist also TM modes in a rectangular hollow metallic waveguide. The modes can similarly be pictured as TM-polarized plane wave being reflected by side walls, following a certain propagation plane. Like TE modes, they can be classified by two integer subscripts, as TM$_{mn}$ modes, where m and n denotes the number of π phase changes of modal field along x and y directions, respectively. However, here neither m nor n can be zero. Effectively speaking, one can't get a legitimate mode solution satisfying all boundary conditions if one tries to use xz or yz cut-plane as wave propagation plane.

It turns out that the condition for lateral wave vector components in order to have a legitimate TM mode is the same as that for the TE mode (with the same m and n values), *i.e.* Eq. 9.21. It follows that dispersion relation as well as cut-off frequency for each TM mode are also identical to those of the corresponding TE mode, *i.e.* Eqs. 9.22 and 9.23.

Field components for TM$_{mn}$ modes are summarized as follows

$$E_x = \frac{k_z k_x}{k_x^2 + k_y^2} E_0 \cos(k_x x) \sin(k_y y) \sin(-k_z z + \omega t) \tag{9.24}$$

$$E_y = \frac{k_z k_y}{k_x^2 + k_y^2} E_0 \sin(k_x x) \cos(k_y y) \sin(-k_z z + \omega t) \tag{9.25}$$

$$E_z = E_0 \sin(k_x x) \sin(k_y y) \cos(-k_z z + \omega t) \tag{9.26}$$

$$H_x = -\frac{k k_y}{Z_0(k_x^2 + k_y^2)} E_0 \sin(k_x x) \cos(k_y y) \sin(-k_z z + \omega t), \tag{9.27}$$

$$H_y = \frac{k k_x}{Z_0(k_x^2 + k_y^2)} E_0 \cos(k_x x) \sin(k_y y) \sin(-k_z z + \omega t). \tag{9.28}$$

The lowest-order mode, *i.e.* TM$_{11}$ mode has its mode field plotted in Fig. 9.6. From the left panel, one sees that the waveguide mode has the maximum z-directed power close to the side walls, which decreases towards waveguide center.

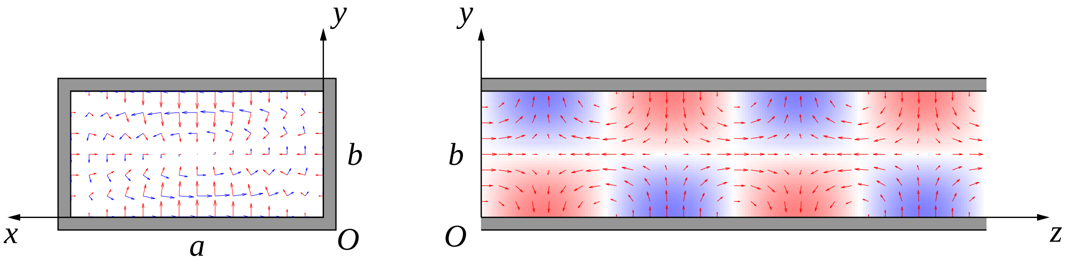

Figure 9.6: TM$_{11}$ mode. Left: Electric (red arrows) and magnetic (blue arrows) fields as seen from cross-sectional plane (xy) of the hollow metallic waveguide. Right: Electric field (red arrows) and x-component of magnetic field (red-blue color map) on yz cut plane at $x = a/2$. Red color denotes positive value and blue color denotes negative value.

Appendix A: Solution to exercises

Chap. 2 Electrostatics

1. $\mathbf{E} = E\hat{\boldsymbol{r}} = 450(\frac{\sqrt{2}}{2}\hat{\boldsymbol{x}} + \frac{\sqrt{2}}{2}\hat{\boldsymbol{y}})$ V/m.
2. $\mathbf{E} = E\hat{\boldsymbol{r}} = 545.45(-0.7785\hat{\boldsymbol{x}} - 0.6228\hat{\boldsymbol{y}} + 0.0778\hat{\boldsymbol{z}})$ V/m.
3. $\mathbf{E} = 587.88\hat{\boldsymbol{z}}$ V/m.
4. $\mathbf{F} = F\hat{\boldsymbol{r}} = 45(\frac{\sqrt{2}}{2}\hat{\boldsymbol{x}} + \frac{\sqrt{2}}{2}\hat{\boldsymbol{y}})$ nN; $W_E = -q_2\Delta V = 3.182$ nJ.
5. $\mathbf{E}_1 = E\hat{\boldsymbol{r}} = 28.125(\frac{\sqrt{2}}{2}\hat{\boldsymbol{x}} + \frac{\sqrt{2}}{2}\hat{\boldsymbol{y}})$ V/m; $\mathbf{E}_2 = 0$ V/m.
6. $V_1 = 15.91$ V; $V_2 = 45$ V; $V_3 = 45$ V.
7. $\Delta V = -48$ V.
8. $W_F = 0.212$ mJ.
9. $\mathbf{E}_1 = 28.125(\frac{\sqrt{2}}{2}\hat{\boldsymbol{x}} + \frac{\sqrt{2}}{2}\hat{\boldsymbol{y}})$ V/m; $E_2 = 20$ V/m and $E_3 = 1800$ V/m (same direction). (Use *generalized* Gauss's law.)
10. $\mathbf{E}_2 = 10\hat{\boldsymbol{x}} - 8\hat{\boldsymbol{y}} + 2.4\hat{\boldsymbol{z}}$ V/m.
11. $C = 0.214$ nF/m.

Chap. 3 Electric circuit

1. $J_s = 10^{-4}$ A/m along x direction. (Charge passing through a unit length in the plane per unit time.)
2. $I = 1$ mA, in the same direction as the rotation.
3. Same as the previous problem.

Chap. 4 Magnetostatics

1. $\mathbf{B} = -2.828 \times 10^{-5}\hat{\boldsymbol{y}}$ T. (Refer to Example "Magnetic field by line current" and set appropriate integration limits.)
2. $\mathbf{B} = 8 \times 10^{-5}\hat{\boldsymbol{z}}$ T. (Use the result in the previous problem. Four sides of the current loop have "similar contributions to the field at the point.)
3. $\mathbf{B} = 8.886 \times 10^{-5}\hat{\boldsymbol{z}}$ T. (Refer to Example "Magnetic field by circular current loop".)
4. $\mathbf{B} = -4.189\hat{\boldsymbol{z}}$ nT.
5. $\mathbf{B} = 1.257 \times 10^{-7}\hat{\boldsymbol{z}}$ T. (Use knowledge in the previous problem.)
6. $\mathbf{B} = 1.742 \times 10^{-11}\hat{\boldsymbol{z}}$ T. (Similar to the previous problem. Each differential line charge, when rotating about center of the line, forms a current loop. Note: suggested solution for Pre-exam 2 in 2017 is incorrect.)
7. $\mathbf{F}_m = 1\hat{\boldsymbol{z}}$ nN.
8. $\mathbf{F}_m = -\hat{\boldsymbol{x}} + 0.6\hat{\boldsymbol{y}} - 0.6\hat{\boldsymbol{z}} = 1.311(-0.7625\hat{\boldsymbol{x}} + 0.4575\hat{\boldsymbol{y}} - 0.4575\hat{\boldsymbol{z}})$ nN.
9. The particle experiences at this moment a magnetic force $\mathbf{F}_m = 10\hat{\boldsymbol{y}}$ μN. Later on, the particle will move in a circular motion with a radius of 10 cm, counterclockwise as observed from $z = +\infty$.
10. $F_m = 0.2$ mN per meter length of conductor. Direction: towards the other conductor.

11. $\mathbf{B} = -6.283 \times 10^{-5} \hat{\boldsymbol{y}}$ T.

Chap. 5 Magnetic induction

1. 2.667×10^{-7} V. Current direction: counterclockwise. (Use either Eq. 5.1 or Eq. 5.15.)

Chap. 6 Electromagnetic wave

(to be updated)