Konstantin K. Likharev

Essential Graduate Physics

Lecture Notes and Problems

Open online access at mirror sites

http://commons.library.stonybrook.edu/egp/

https://essentialgraduatephysics.org/

https://sites.google.com/site/likharevegp/

Part EM:
Classical Electrodynamics

Last edit: July 2, 2024

B/W paperback copies of this volume are also available on *Amazon.com*:
https://www.amazon.com/gp/product/B0D81M69BD

About the author:
https://you.stonybrook.edu/likharev/

Table of Contents

* * *

Supplemental file **Exercise Problems with Model Solutions** (300 problems, 420 pp.)
is available online:
https://essentialgraduatephysics.org/Files/EM%20exercises.pdf .

B/W paperback copies of these materials are available on *Amazon.com*:
https://www.amazon.com/gp/product/B0D7SKPQF9 .

Additional file **Test Problems with Model Solutions** (52 problems, 46 pp.)
is available for course instructors from the author upon request – see *Front Matter*.

* * *

Introductory Remarks

The structure of this classical electrodynamics course is quite traditional. Namely, in order to address the most important subjects of the field, which involve not only charged point particles but also conducting, dielectric, and magnetic media, the electromagnetic interactions are discussed in parallel with simple models of the electric and magnetic properties of most common materials.

Also following tradition, I use this part of my series (notably Chapter 2) as a convenient platform for the discussion of various methods of the solution of partial differential equations, including the use of the most important systems of curvilinear orthogonal coordinates and special functions.

One more traditional part of classical electrodynamics is an introduction to special relativity (in Chapter 9) because although this topic includes a substantial classical mechanics component, it is the electrodynamics that makes a relativistic analysis unavoidable.

Chapter 1. Electric Charge Interaction

This chapter reviews the basics of electrostatics – the description of interactions between stationary (or relatively slowly moving) electric charges. Much of this material should be known to the reader from their undergraduate studies;[1] *because of that, the explanations are very brief.*

1.1. The Coulomb law

A quantitative discussion of classical electrodynamics, starting from the electrostatics, requires common agreement on the meaning of the following notions:[2]

- *electric charges* q_k, as revealed, most explicitly, by observation of *electrostatic interaction* between the charged particles;

- *point charges* – the charged particles so small that their position in space, for the given problem, may be completely described (in the given reference frame) by their radius-vectors $\mathbf{r}_k$; and

- *electric charge conservation* – the fact that the algebraic sum of all charges q_k inside any closed volume is conserved unless the charged particles cross the volume's border.

I will assume that these notions are well known to the reader. Using them, the *Coulomb law*[3] for the interaction of two stationary point charges may be formulated as follows:

Coulomb law

$$\mathbf{F}_{kk'} = \kappa q_k q_{k'} \frac{\mathbf{r}_k - \mathbf{r}_{k'}}{\left|\mathbf{r}_k - \mathbf{r}_{k'}\right|^3} \equiv \kappa \frac{q_k q_{k'}}{R_{kk'}^2} \mathbf{n}_{kk'}, \quad \text{with } \mathbf{R}_{kk'} \equiv \mathbf{r}_k - \mathbf{r}_{k'}, \quad \mathbf{n}_{kk'} \equiv \frac{\mathbf{R}_{kk'}}{R_{kk'}}, \quad R_{kk'} \equiv \left|\mathbf{R}_{kk'}\right|, \tag{1.1}$$

where $\mathbf{F}_{kk'}$ denotes the electrostatic (*Coulomb*) force exerted on the charge number k by the charge number k', separated from it by distance $R_{kk'}$ – see Fig. 1.

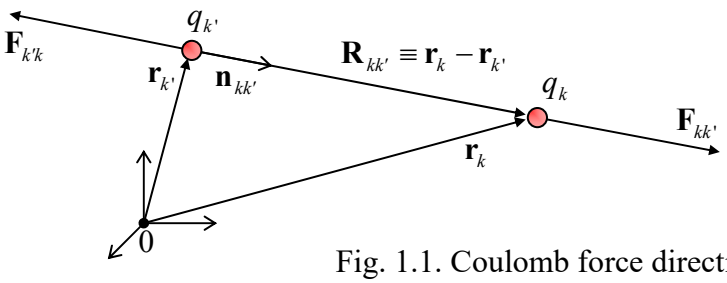


Fig. 1.1. Coulomb force directions (for the case $q_k q_{k'} > 0$).

[1] For remedial reading, I can recommend, for example, D. Griffiths, *Introduction to Electrodynamics*, 4th ed., Pearson, 2015.

[2] On top of the more general notions of the *classical Newtonian space*, *point particles* and *forces*, as used in classical mechanics – see, e.g., CM Sec. 1.1.

[3] Formulated in 1785 by Charles-Augustin de Coulomb, on the basis of his earlier experiments, in turn rooted in prior studies of electrostatic phenomena, with notable contributions by William Gilbert, Otto von Guericke, Charles François de Cisternay Du Fay, Benjamin Franklin, and Henry Cavendish.

I am confident that this law is very familiar to the reader, but a few comments may still be due:

(i) Flipping the indices k and k', we see that Eq. (1) complies with the 3rd Newton law: the reciprocal force is equal in magnitude but opposite in direction: $\mathbf{F}_{k'k} = -\mathbf{F}_{kk'}$.

(ii) Since the vector $\mathbf{R}_{kk'} \equiv \mathbf{r}_k - \mathbf{r}_{k'}$, by its definition, is directed from point $\mathbf{r}_{k'}$ toward point $\mathbf{r}_k$ (Fig. 1), Eq. (1) correctly describes the experimental fact that charges of the same sign (i.e. with $q_k q_{k'} > 0$) repulse, while those with opposite signs ($q_k q_{k'} < 0$) attract each other.

(iii) In some textbooks, the Coulomb law (1) is given with the qualifier "in free space" or "in vacuum". However, actually, Eq. (1) remains valid even in the presence of any other charges – for example, of internal charges in a quasi-continuous medium that may surround the two charges (number k and k') under consideration. The confusion stems from the fact, to be discussed in detail in Chapter 3 below, that in some cases it is convenient to *formally* represent the effect of the other charges as an *effective* (rather than actual!) modification of the Coulomb law.

(iv) The constant κ in Eq. (1) depends on the system of units we use. In the *Gaussian* units, κ is set to 1, for the price of introducing a special unit of charge (the *statcoulomb*) that would make experimental data compatible with Eq. (1) if the force $\mathbf{F}_{kk'}$ is measured in the Gaussian units (*dynes*). On the other hand, in the *International System* ("SI") of units, the charge's unit is one *coulomb* (abbreviated C), and κ is different from 1:

$$\kappa\big|_{\rm SI} = \frac{1}{4\pi\varepsilon_0}, \qquad (1.2)$$

κ in SI units

where $\varepsilon_0 \approx 8.854\times10^{-12}$ is called the *electric constant*.[4]

Unfortunately, the continuing struggle between zealous proponents of these two systems of units bears all the not-so-nice features of a religious war, with a similarly slim chance for any side to win it in any foreseeable future. In my humble view, each of these systems has its advantages and handicaps (to be noted on several occasions below), and every educated physicist should have no problem with using any of them. Following insisting recommendations of international scientific unions, I am using the SI units throughout my series. However, for the readers' convenience, in this course (where the difference between the Gaussian and SI systems is especially significant) I will write the most important formulas with the constant (2) clearly displayed – for example, the combination of Eqs. (1) and (2) as

$$\mathbf{F}_{kk'} = \frac{1}{4\pi\varepsilon_0} q_k q_{k'} \frac{\mathbf{r}_k - \mathbf{r}_{k'}}{\left|\mathbf{r}_k - \mathbf{r}_{k'}\right|^3}, \qquad (1.3)$$

so the formal transfer to the Gaussian units may be performed just by dropping the front fraction. (In the rare cases when the transfer is not obvious, I will duplicate formulas in the Gaussian units.)

Besides Eq. (3), another key experimental law of electrostatics is the *linear superposition principle*: the electrostatic forces exerted on some point charge (say, q_k) by other charges add up as vectors, forming the net force

[4] Since 2018, one coulomb is defined, in the "legal" metrology, as a certain exactly fixed number of the fundamental electric charges e, and the "legal" SI value of ε_0 is not more exactly equal to $10^7/4\pi c^2$ (where c is the speed of light) as it was before, but remains extremely close to that fraction, with the relative difference of the order of 10^{-10} – see appendix *UCA*: *Selected Units and Constants*. In this series, this minute difference is ignored.

$$\mathbf{F}_k = \sum_{k' \neq k} \mathbf{F}_{kk'}, \tag{1.4}$$

where the summation is extended over all charges but q_k, and the partial force $\mathbf{F}_{kk'}$ is described by Eq. (3). The fact that the sum is restricted to $k' \neq k$ means that a *point charge, in statics, does not interact with itself*. This fact may look obvious from Eq. (3), whose right-hand side diverges at $\mathbf{r}_k \to \mathbf{r}_{k'}$, but becomes less evident (though still true) in quantum mechanics – where the charge of even an elementary particle is effectively spread around some volume, together with the particle's wavefunction.[5]

Now we may combine Eqs. (3) and (4) to get the following expression for the net force $\mathbf{F}$ acting on a *probe charge* q located at point $\mathbf{r}$:

$$\mathbf{F}(\mathbf{r}) = q \frac{1}{4\pi\varepsilon_0} \sum_{\mathbf{r}_{k'} \neq \mathbf{r}} q_{k'} \frac{\mathbf{r} - \mathbf{r}_{k'}}{\left|\mathbf{r} - \mathbf{r}_{k'}\right|^3}. \tag{1.5}$$

This equality implies that it makes sense to introduce the notion of the *electric field* (as an entity independent of q), whose distribution in space is characterized by the following vector:

Electric field: definition

$$\boxed{\mathbf{E}(\mathbf{r}) \equiv \frac{\mathbf{F}(\mathbf{r})}{q},} \tag{1.6}$$

formally called the *electric field strength* – but much more frequently, just the "electric field". In these terms, Eq. (5) becomes

Electric field of point charges

$$\mathbf{E}(\mathbf{r}) = \frac{1}{4\pi\varepsilon_0} \sum_{\mathbf{r}_{k'} \neq \mathbf{r}} q_{k'} \frac{\mathbf{r} - \mathbf{r}_{k'}}{\left|\mathbf{r} - \mathbf{r}_{k'}\right|^3}. \tag{1.7}$$

Being just convenient is electrostatics, the notion of the field becomes unavoidable for the description of time-dependent phenomena (such as electromagnetic waves, see Chapter 7 and on), where the electromagnetic field shows up as a specific form of matter, different from the usual "material" particles – even though quantum electrodynamics (to be reviewed in QM Chapter 9) offers their joint description.

Many real-world problems involve multiple point charges located so closely that it is possible to approximate them with a continuous charge distribution. Indeed, let us consider a group of many ($dN >> 1$) close charges, located at points $\mathbf{r}_{k'}$, all within an elementary volume d^3r'. For relatively distant field observation points, with $|\mathbf{r} - \mathbf{r}_{k'}| >> dr'$, the geometrical factor in the corresponding terms of Eq. (7) is essentially the same. As a result, these charges may be treated as a single elementary charge $dQ(\mathbf{r}')$. Since at $dN >> 1$, this elementary charge is proportional to the elementary volume d^3r', we can define the local 3D *charge density* $\rho(\mathbf{r}')$ by the following relation:

$$\rho(\mathbf{r}')d^3r' \equiv dQ(\mathbf{r}') \equiv \sum_{r_{k'} \in d^3r'} q_{k'}, \tag{1.8}$$

and rewrite Eq. (7) as an integral (over the volume containing all essential charges):

Electric field of continuous charge

$$\boxed{\mathbf{E}(\mathbf{r}) = \frac{1}{4\pi\varepsilon_0} \int \rho(\mathbf{r}') \frac{\mathbf{r} - \mathbf{r}'}{\left|\mathbf{r} - \mathbf{r}'\right|^3} d^3r'.} \tag{1.9}$$

[5] Note that some widely used approximations, e.g., the density functional theory (DFT) of multiparticle systems, essentially violate this law, thus limiting their accuracy and applicability – see, e.g., QM Sec. 8.4.

Note that for a continuous, smooth charge density $\rho(\mathbf{r}')$, the integral in Eq. (9) does not diverge at $\mathbf{R} \equiv \mathbf{r} - \mathbf{r}' \to 0$, because in this limit, the fraction under the integral increases as R^{-2}, i.e. slower than the decrease of the elementary volume d^3r', proportional to R^3.

Let me emphasize the dual use of Eq. (9). In the case when $\rho(\mathbf{r})$ is a continuous function representing the average charge defined by Eq. (8), Eq. (9) is not valid at distances $|\mathbf{r} - \mathbf{r}_{k'}|$ of the order of the distance between the adjacent point charges, i.e. does not describe rapid variations of the electric field at these distances. Such *approximate*, smoothly changing field $\mathbf{E}(\mathbf{r})$, is called *macroscopic*; we will repeatedly return to this notion in the following chapters. On the other hand, Eq. (9) may be also used for the description of the *exact* (frequently called *microscopic*) field of discrete point charges, by employing the notion of Dirac's delta function, which is the mathematical description of a very sharp function equal to zero everywhere but one point, and still having a finite integral (equal to 1).[6] Indeed, in this formalism, a set of point charges $q_{k'}$ located in points $\mathbf{r}_{k'}$ may be represented by the pseudo-continuous density

$$\rho(\mathbf{r}') = \sum_{k'} q_{k'} \delta(\mathbf{r}' - \mathbf{r}_{k'}). \tag{1.10}$$

Plugging this expression into Eq. (9), we return to its exact, discrete version (7). In this sense, Eq. (9) is exact, and we may use it as the general expression for the electric field.

1.2. The Gauss law

Due to the extension of Eq. (9) to point ("discrete") charges, it may seem that we do not need anything besides it to solve any problem of electrostatics. In practice, however, this is not quite true – first of all, because the direct use of Eq. (9) frequently leads to complex calculations. Indeed, let us try to solve a problem that is conceptually very simple: find the electric field induced by a spherically symmetric charge distribution with density $\rho(r')$ – see Fig. 2.

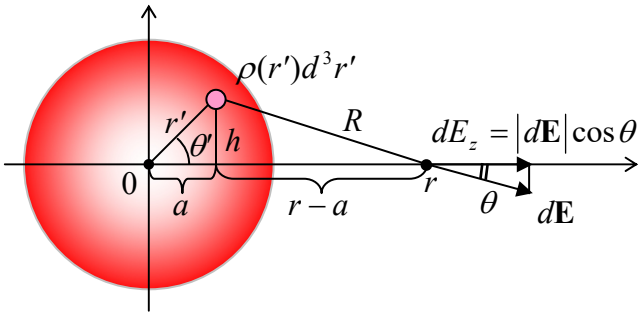


Fig. 1.2. One of the simplest problems of electrostatics: the electric field produced by a spherically-symmetric charge distribution.

We may immediately use the problem's symmetry to argue that the electric field should be also spherically symmetric, with only one component in the spherical coordinates: $\mathbf{E}(\mathbf{r}) = E(r)\mathbf{n}_r$, where $\mathbf{n}_r \equiv \mathbf{r}/r$ is the unit vector in the direction of the field observation point **r.** Taking this direction for the polar axis of a spherical coordinate system, we can use the evident axial symmetry of the system to reduce Eq. (9) to

[6] See, e.g., MA Sec. 14. The 2D (*areal*) charge density σ and the 1D (*linear*) density λ may be defined absolutely similarly to the 3D (*volumic*) density ρ: $dQ = \sigma d^2r$, $dQ = \lambda dr$. Note that the approximations in that either $\sigma \neq 0$ or $\lambda \neq 0$ imply that ρ is formally infinite at the charge location; for example, the model in that a plane $z = 0$ is charged with areal density $\sigma \neq 0$, means that $\rho = \sigma\delta(z)$, where $\delta(z)$ is Dirac's delta function.

$$E = \frac{1}{4\pi\varepsilon_0} 2\pi \int_0^{\pi} \sin\theta' d\theta' \int_0^{\infty} r'^2 dr' \frac{\rho(r')}{R^2} \cos\theta, \tag{1.11}$$

where θ, θ', and R are the geometrical parameters marked in Fig. 2. Since θ and R may be readily expressed via r' and θ', using the auxiliary parameters a and h,

$$\cos\theta = \frac{r-a}{R}, \qquad R^2 = h^2 + (r - r'\cos\theta)^2, \qquad \text{where } a \equiv r'\cos\theta', \quad h \equiv r'\sin\theta', \tag{1.12}$$

Eq. (11) may be eventually reduced to an explicit integral over r' and θ', and worked out analytically, but that would require some effort.

For other problems, the integral (9) may be much more complicated, defying an analytical solution. One could argue that with the present-day abundance of computers and numerical algorithm libraries, one can always resort to numerical integration. This argument may be enhanced by the fact that numerical *integration* is based on the replacement of the required integral by a discrete sum, and the summation is much more robust to the (unavoidable) rounding errors than the finite-difference schemes typical for the numerical solution of *differential* equations. These arguments, however, are only partly justified, since in many cases the numerical approach runs into a problem sometimes called the *curse of dimensionality* – the exponential dependence of the number of needed calculations on the number of independent parameters of the problem.[7] Thus, despite the proliferation of numerical methods in physics, analytical results have an everlasting value, and we should try to get them whenever we can. For our current problem of finding the electric field generated by a fixed set of electric charges, large help may come from the so-called *Gauss law*.

To derive it, let us consider a single point charge q inside a smooth closed surface S (Fig. 3), and calculate the product $E_n d^2r$, where d^2r is an elementary area of the surface (which may be well approximated with a plane fragment of that area), and $E_n \equiv \mathbf{E}\cdot\mathbf{n}$ is the component of the electric field at that point, normal to the plane.

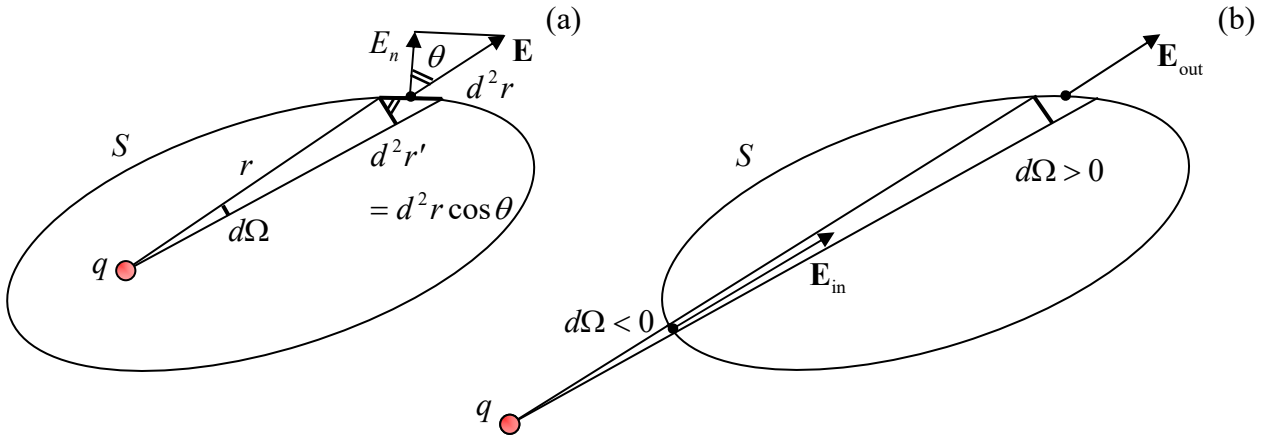


Fig. 1.3. Deriving the Gauss law: a point charge q (a) inside the volume V, and (b) outside of that volume.

This component may be calculated as $E\cos\theta$, where θ is the angle between the vector $\mathbf{E}$ and the unit vector $\mathbf{n}$ normal to the surface. Now let us notice that the product $\cos\theta\, d^2r$ is nothing more than the

[7] For a more detailed discussion of this problem, see, e.g., CM Sec. 5.8.

area d^2r' of the projection of d^2r onto the plane normal to the vector **r** connecting the charge q with the considered point of the surface (Fig. 3), because the angle between the elementary areas d^2r' and d^2r is also equal to θ. Using the Coulomb law for **E**, we get

$$E_n d^2 r = E\cos\theta\, d^2 r = \frac{1}{4\pi\varepsilon_0}\frac{q}{r^2} d^2 r'. \tag{1.13}$$

But the ratio d^2r'/r^2 is nothing more than the elementary solid angle $d\Omega$ under which the areas d^2r' and d^2r are seen from the charge point, so $E_n d^2r$ may be represented just as a product of $d\Omega$ by a constant $(q/4\pi\varepsilon_0)$. Summing these products over the whole surface, we get

$$\oint_S E_n d^2 r = \frac{q}{4\pi\varepsilon_0}\oint_S d\Omega \equiv \frac{q}{\varepsilon_0}, \tag{1.14}$$

since the full solid angle equals 4π. (The integral on the left-hand side of this relation is called the *flux of electric field* through the surface S.)

Relation (14) expresses the Gauss law for one point charge. However, it is only valid if the charge is located *inside* the volume V limited by the surface S. To find the flux created by a charge located *outside* of this volume, we still can use Eq. (13), but have to be careful with the signs of the elementary contributions $E_n dA$. Let us use the common convention to direct the unit vector **n** out of the closed volume we are considering (the so-called *outer normal*), so the elementary product $E_n d^2r = (\mathbf{E}\cdot\mathbf{n})d^2r$ and hence $d\Omega = E_n d^2r'/r^2$ is positive if the vector **E** is pointing out of the volume (like in the example shown in Fig. 3a and at the upper-right area in Fig. 3b), and negative in the opposite case (for example, at the lower-left area in Fig. 3b). As the latter panel shows, if the charge is located outside of the volume, for each positive contribution $d\Omega$ there is always an equal and opposite contribution to the integral. As a result, at the integration over the solid angle, the positive and negative contributions cancel exactly, so

$$\oint_S E_n d^2 r = 0. \tag{1.15}$$

The real power of the Gauss law is revealed by its generalization to the case of several, especially many charges. Since the calculation of flux is a linear operation, the linear superposition principle (4) means that the flux created by several charges is equal to the (algebraic) sum of individual fluxes from each charge, for which either Eq. (14) or Eq. (15) are valid, depending on whether the charge is in or out of the volume. As a result, for the total flux, we get:

$$\oint_S E_n d^2 r = \frac{Q_V}{\varepsilon_0} \equiv \frac{1}{\varepsilon_0}\sum_{\mathbf{r}_j \in V} q_j \equiv \frac{1}{\varepsilon_0}\int_V \rho(\mathbf{r}')d^3r', \tag{1.16}$$

Gauss law

where Q_V is the net charge inside volume V. This is the full version of the Gauss law.[8]

In order to appreciate the problem-solving power of the law, let us revisit the problem shown in Fig. 2, i.e. the field of a spherical charge distribution. Due to its symmetry, which had already been discussed above, if we apply Eq. (16) to a sphere of a certain radius r, the electric field has to be normal

[8] The law is named after the famed Carl Gauss (1777-1855), even though it was first formulated earlier (in 1773) by Joseph-Louis Lagrange who was also the father-founder of analytical mechanics – see, e.g., CM Chapter 2.

to the sphere at each point (i.e., $E_n = E$), and its magnitude has to be the same at all points: $E_n = E(r)$. As a result, the flux calculation is elementary:

$$\oint E_n d^2 r = 4\pi r^2 E(r). \tag{1.17}$$

Now applying the Gauss law (16), we get:

$$4\pi r^2 E(r) = \frac{1}{\varepsilon_0} \int_{r'<r} \rho(r') d^3 r' = \frac{4\pi}{\varepsilon_0} \int_0^r r'^2 \rho(r') dr', \tag{1.18}$$

so, finally,

$$E(r) = \frac{1}{r^2 \varepsilon_0} \int_0^r r'^2 \rho(r') dr' \equiv \frac{1}{4\pi\varepsilon_0} \frac{Q_r}{r^2}, \tag{1.19}$$

where Q_r is the full charge inside the sphere of radius r:

$$Q_r \equiv \int_{r'<r} \rho(r') d^3 r' = 4\pi \int_0^r \rho(r') r'^2 dr'. \tag{1.20}$$

In particular, this formula shows that the field *outside* of a sphere of a finite radius R is exactly the same as if all its charge $Q = Q(R)$ is concentrated in the sphere's center. (Note that this important result is only valid for a spherically symmetric charge distribution.) For the field *inside* the sphere, finding the electric field still requires the explicit integration (20), but this 1D integral is much simpler than the 2D integral (11), and in some important cases may be readily worked out analytically. For example, if the charge Q is uniformly distributed inside a sphere of radius R,

$$\rho(r') = \rho = \frac{Q}{V} = \frac{Q}{(4\pi/3)R^3}, \tag{1.21}$$

then the integration is elementary:

$$E(r) = \frac{\rho}{r^2 \varepsilon_0} \int_0^r r'^2 dr' = \frac{\rho r}{3\varepsilon_0} = \frac{1}{4\pi\varepsilon_0} \frac{Qr}{R^3}. \tag{1.22}$$

We see that in this case, the field is growing linearly from the center to the sphere's surface, and only at $r > R$ starts to decrease in agreement with Eq. (19) with constant $Q(r) = Q$. Note also that the electric field is continuous for all r (including $r = R$) – as for all systems with finite volumic density,

In order to underline the importance of the last condition, let us consider one more elementary but very important example of Gauss law's application. Let a thin plane sheet (Fig. 4) be charged uniformly, with a finite *areal* density σ = const. In this case, it is fruitful to use the Gauss volume in the form of a planar "pillbox" of thickness $2z$ (where z is the Cartesian coordinate perpendicular to the plane) and certain area A – see the dashed lines in Fig. 4. Due to the symmetry of the problem, it is evident that the electric field should be: (i) directed along the z-axis, (ii) constant on each of the upper and bottom sides of the pillbox, (iii) equal and opposite on these sides, and (iv) parallel to the side surfaces of the box. As a result, the full electric field flux through the pillbox's surface is just $2AE(z)$, so the Gauss law (16) yields $2AE(z) = Q_A/\varepsilon_0 \equiv \sigma A/\varepsilon_0$, and we get a very simple but important formula

$$E(z) = \frac{\sigma}{2\varepsilon_0} = \text{const}. \tag{1.23}$$

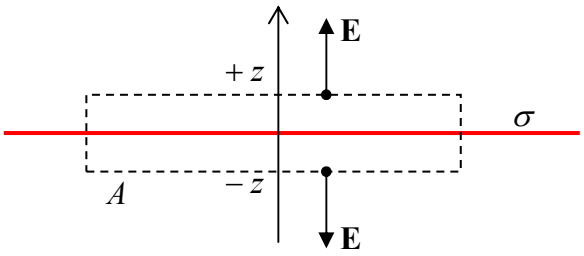


Fig. 1.4. The electric field of a charged plane.

Notice that, somewhat counter-intuitively, the field magnitude does not depend on the distance from the charged plane. From the point of view of the Coulomb law (5), this result may be explained as follows: the farther the observation point from the plane, the weaker the effect of each elementary charge, $dQ = \sigma d^2 r$, but the more such elementary charges give contributions to the z-component of vector $\mathbf{E}$, because they are "seen" from the observation point at relatively small angles to the z-axis.

Note also that though the magnitude $E \equiv |\mathbf{E}|$ of this electric field is constant, its component E_n normal to the plane (for our coordinate choice, E_z) changes its sign at the plane, experiencing a *discontinuity* (jump) equal to

$$\Delta E_z \equiv E_z(z=+0) - E_z(z=-0) = \frac{\sigma}{\varepsilon_0}. \tag{1.24}$$

This jump disappears if the surface is not charged. Returning for a split second to our charged sphere problem (Fig. 2), solving it we have considered the volumic charge density ρ to be finite everywhere, including the sphere's surface, so on it $\sigma = 0$, and the electric field should be continuous – as it is.

Admittedly, the *integral form* (16) of the Gauss law is immediately useful only for highly symmetrical geometries, such as in the two problems discussed above. However, it may be recast into an alternative, *differential form* whose field of useful applications is much wider. This form may be obtained from Eq. (16) using the *divergence theorem* of the vector algebra, which is valid for any space-differentiable vector, in particular $\mathbf{E}$, and for the volume V limited by any closed surface S:[9]

$$\oint_S E_n d^2 r = \int_V (\nabla \cdot \mathbf{E}) d^3 r, \tag{1.25}$$

where ∇ is the *del* (or "nabla") *operator* of spatial differentiation.[10] Combining Eq. (25) with the Gauss law (16), we get

$$\int_V \left(\nabla \cdot \mathbf{E} - \frac{\rho}{\varepsilon_0} \right) d^3 r = 0. \tag{1.26}$$

For a given spatial distribution of electric charge (and hence of its electric field), this equation should be valid for any choice of the volume V. This can hold only if the function under the integral vanishes at each point, i.e. if[11]

[9] See, e.g., MA Eq. (12.2). Note also that the scalar product under the volumic integral in Eq. (25) is nothing else than the divergence of the vector $\mathbf{E}$ – see, e.g., MA Eq. (8.4), hence the theorem's name.

[10] See, e.g., MA Secs. 8-10.

[11] In the Gaussian units, just as in the initial Eq. (6), ε_0 has to be replaced with $1/4\pi$, so the Maxwell equation (27) looks like $\nabla \cdot \mathbf{E} = 4\pi\rho$, while Eq. (28) stays the same.

Inhomogeneous Maxwell equation for **E**

$$\nabla \cdot \mathbf{E} = \frac{\rho}{\varepsilon_0}. \tag{1.27}$$

Note that in sharp contrast with the integral form (16), Eq. (27) is *local*: it relates the electric field's divergence to the charge density *at the same point*. This equation, being the differential form of the Gauss law, is frequently called one of the famed *Maxwell equations*[12] – to be discussed again and again later in this course.

In the mathematical terminology, Eq. (27) is *inhomogeneous*, because it has a right-hand side independent (at least explicitly) of the field **E** that it describes. Another, *homogeneous* Maxwell equation's "embryo" (this one valid for the stationary case only!) may be obtained by noticing that the curl of the point charge's field, and hence that of *any* system of charges, equals zero:[13]

Homogeneous Maxwell equation for **E**

$$\nabla \times \mathbf{E} = 0. \tag{1.28}$$

(We will arrive at two other Maxwell equations, for the magnetic field, in Chapter 5, and then generalize all the equations to their full, time-dependent form at the end of Chapter 6. However, Eq. (27) will stay the same.)

Just to get a better gut feeling of Eq. (27), let us apply it to the same example of a uniformly charged sphere (Fig. 2). Vector algebra tells us that the divergence of a spherically symmetric vector function $\mathbf{E}(\mathbf{r}) = E(r)\mathbf{n}_r$ may be simply expressed in spherical coordinates:[14] $\nabla\cdot\mathbf{E} = [d(r^2E)/dr]/r^2$. As a result, Eq. (27) yields a linear ordinary differential equation for the scalar function $E(r)$:

$$\frac{1}{r^2}\frac{d}{dr}(r^2E) = \begin{cases} \rho/\varepsilon_0, & \text{for } r \le R, \\ 0, & \text{for } r \ge R, \end{cases} \tag{1.29}$$

which may be readily integrated on each of these segments:

$$E(r) = \frac{1}{\varepsilon_0}\frac{1}{r^2} \times \begin{cases} \rho\int r^2 dr = \rho r^3/3 + c_1, & \text{for } r \le R, \\ c_2, & \text{for } r \ge R. \end{cases} \tag{1.30}$$

To determine the integration constant c_1, we can use the following boundary condition: $E(0) = 0$. (It follows from the problem's spherical symmetry: in the center of the sphere, the electric field has to vanish, because otherwise, where would it be directed?) This requirement gives $c_1 = 0$. The second constant, c_2, may be found from the continuity condition $E(R - 0) = E(R + 0)$, which has already been discussed above, giving $c_2 = \rho R^3/3 \equiv Q/4\pi$. As a result, we arrive at our previous results (19) and (22).

We can see that in this particular, highly symmetric case, using the differential form of the Gauss law is a bit more complex than its integral form. (For our second example, shown in Fig. 4, it would be even less natural.) However, Eq. (27) and its generalizations are more convenient for asymmetric charge

[12] Named after the genius of classical electrodynamics and statistical physics, James Clerk Maxwell (1831-1879).

[13] This follows, for example, from the direct application of MA Eq. (10.11) to any spherically-symmetric vector function of type $\mathbf{f}(\mathbf{r}) = f(r)\mathbf{n}_r$ (in particular, to the electric field of a point charge placed at the origin), giving $f_\theta = f_\varphi = 0$ and $\partial f_r/\partial\theta = \partial f_r/\partial\varphi = 0$ so all components of the vector $\nabla\times\mathbf{f}$ vanish. Since nothing prevents us from placing the reference frame's origin at the point charge's location, this result remains valid for any position of the charge.

[14] See, e.g., MA Eq. (10.10) for the particular case $\partial/\partial\theta = \partial/\partial\varphi = 0$.

distributions, and are invaluable in cases where the distribution $\rho(\mathbf{r})$ is not known *a priori* and has to be found in a self-consistent way. (We will start discussing such cases in the next chapter.)

1.3. Scalar potential and electric field energy

One more help for solving problems of electrostatics (and electrodynamics as a whole) may be obtained from the notion of the *electrostatic potential*, which is just the electrostatic potential energy U of a probe point charge q placed into the field in question, normalized by its charge:

$$\phi \equiv \frac{U}{q}. \tag{1.31}$$

Electro-static potential

As we know from classical mechanics,[15] the notion of U (and hence ϕ) makes the most sense for the case of *potential forces* – for example, those depending just on the particle's position. Eqs. (6) and (9) show that stationary electric fields fall into this category. For such a field, the potential energy may be defined as a scalar function $U(\mathbf{r})$ that allows the force to be calculated as its gradient (with the opposite sign):

$$\mathbf{F} = -\nabla U\,. \tag{1.32}$$

Dividing both sides of this equation by the probe charge, and using Eqs. (6) and (31), we get[16]

$$\mathbf{E} = -\nabla\phi\,. \tag{1.33}$$

Electrostatic field as a gradient

To calculate the scalar potential, let us start from the simplest case of a single point charge q placed at the origin. For it, Eq. (7) takes the simple form

$$\mathbf{E} = \frac{1}{4\pi\varepsilon_0} q \frac{\mathbf{r}}{r^3} \equiv \frac{1}{4\pi\varepsilon_0} q \frac{\mathbf{n}_r}{r^2}. \tag{1.34}$$

It is straightforward to verify that the last fraction in the last form of Eq. (34) is equal to $-\nabla(1/r)$.[17] Hence, according to the definition (33), for this particular case

$$\phi = \frac{1}{4\pi\varepsilon_0}\frac{q}{r}. \tag{1.35}$$

Potential of a point charge

(In the Gaussian units, this result is spectacularly simple: $\phi = q/r$.) Note that we could add an arbitrary constant to this potential (and indeed to *any* other distribution of ϕ discussed below) without changing the field, but it is convenient to define the potential energy so it would approach zero at infinity.

In order to justify the introduction and the forthcoming exploration of U and ϕ, let me demonstrate (I hope, unnecessarily :-) how useful the notions are, on a very simple example. Let two similar charges q be launched from afar, with the same initial speed $v_0 << c$ each, straight toward each other (i.e. with the zero impact parameter) – see Fig. 5. Since, according to the Coulomb law, the

[15] See, e.g., CM Sec. 1.4.

[16] Eq. (28) could be also derived from this relation because according to vector algebra, any gradient field has no curl – see, e.g., MA Eq. (11.1).

[17] This may be done either by Cartesian components or using the well-known expression $\nabla f = (df/dr)\mathbf{n}_r$ valid for any spherically-symmetric scalar function $f(r)$ – see, e.g., MA Eq. (10.8) for the particular case $\partial/\partial\theta = \partial/\partial\varphi = 0$.

charges repel each other with increasing force, they will stop at some minimum distance r_{min} from each other, and then fly back. We could of course find r_{min} directly from the Coulomb law. However, for that, we would need to write the 2nd Newton law for each particle (actually, due to the problem symmetry, they would be similar), then integrate them over time to find the particle velocity v as a function of distance, and only then recover r_{min} from the requirement $v = 0$.

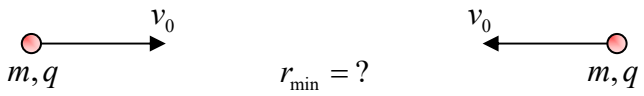


Fig. 1.5. A simple problem of charged particle motion.

The notion of potential allows this problem to be solved in one line. Indeed, in the field of potential forces, the system's total energy $\mathcal{E} = T + U \equiv T + q\phi$ is conserved. In our non-relativistic case $v << c$, the kinetic energy T is just $mv^2/2$. Hence, equating the total energy of two particles at the points $r = \infty$ and $r = r_{min}$, and using Eq. (35) for ϕ, we get

$$2\frac{mv_0^2}{2} + 0 = 0 + \frac{1}{4\pi\varepsilon_0}\frac{q^2}{r_{min}}, \tag{1.36}$$

immediately giving us the final answer: $r_{min} = q^2/4\pi\varepsilon_0 m v_0^2$. So, the notion of scalar potential is indeed very useful.

With this motivation, let us calculate ϕ for an arbitrary configuration of charges. For a single charge in an arbitrary position (say, at point $\mathbf{r}_{k'}$), $r \equiv |\mathbf{r}|$ in Eq. (35) should be evidently replaced with $|\mathbf{r} - \mathbf{r}_{k'}|$. Now, the linear superposition principle (3) allows for an easy generalization of this formula to the case of an arbitrary set of discrete charges,

$$\phi(\mathbf{r}) = \frac{1}{4\pi\varepsilon_0}\sum_{\mathbf{r}_{k'}\neq\mathbf{r}}\frac{q_{k'}}{|\mathbf{r} - \mathbf{r}_{k'}|}. \tag{1.37}$$

Finally, using the same arguments as in Sec. 1, we can use this result to argue that in the case of an arbitrary continuous charge distribution

Potential of a charge distribution

$$\phi(\mathbf{r}) = \frac{1}{4\pi\varepsilon_0}\int\frac{\rho(\mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|}d^3r'. \tag{1.38}$$

Again, Dirac's delta function allows using the last equation to recover Eq. (37) for discrete charges as well, so Eq. (38) may be considered as the general expression for the electrostatic potential.

For most practical calculations, using this expression and then applying Eq. (33) to the result, is preferable to using Eq. (9), because ϕ is a scalar, while $\mathbf{E}$ is a 3D vector, mathematically equivalent to three scalars. Still, this approach may lead to technical problems similar to those discussed in Sec. 2. For example, applying it to the spherically symmetric distribution of charge (Fig. 2), we get the integral

$$\phi = \frac{1}{4\pi\varepsilon_0}2\pi\int_0^{\pi}\sin\theta' d\theta'\int_0^{\infty}r'^2 dr'\frac{\rho(r')}{R}\cos\theta, \tag{1.39}$$

which is not much simpler than Eq. (11).

The situation may be much improved by recasting Eq. (38) into a differential form. For that, it is sufficient to plug the definition of ϕ, Eq. (33), into Eq. (27):

$$\nabla \cdot (-\nabla \phi) = \frac{\rho}{\varepsilon_0}. \tag{1.40}$$

The left-hand side of this equation is nothing else than the Laplace operator of ϕ (with the minus sign), so we get the famous *Poisson equation*[18] for the electrostatic potential:

$$\boxed{\nabla^2 \phi = -\frac{\rho}{\varepsilon_0}}. \tag{1.41}$$

Poisson equation for ϕ

(In the Gaussian units, the Poisson equation is $\nabla^2\phi = -4\pi\rho$.) This differential equation is so convenient for applications that even its particular case for $\rho = 0$,

$$\boxed{\nabla^2 \phi = 0}, \tag{1.42}$$

Laplace equation for ϕ

has earned a special name – the *Laplace equation*.[19]

In order to get a gut feeling of the Poisson equation's value as a problem-solving tool, let us return to the spherically-symmetric charge distribution (Fig. 2) with a constant charge density ρ. Exploiting this symmetry, we can represent the potential as $\phi(r)$, and hence use the following simple expression for its Laplace operator:[20]

$$\nabla^2 \phi = \frac{1}{r^2}\frac{d}{dr}\left(r^2 \frac{d\phi}{dr}\right), \tag{1.43}$$

so for the points inside the charged sphere ($r \le R$) the Poisson equation yields

$$\frac{1}{r^2}\frac{d}{dr}\left(r^2 \frac{d\phi}{dr}\right) = -\frac{\rho}{\varepsilon_0}, \qquad \text{i.e. } \frac{d}{dr}\left(r^2 \frac{d\phi}{dr}\right) = -\frac{\rho}{\varepsilon_0} r^2. \tag{1.44}$$

Integrating the last form of the equation over r once, with the natural boundary condition $d\phi/dr\big|_{r=0} = 0$ (because of the condition $E(0) = 0$, which has been discussed above), we get

$$\frac{d\phi}{dr}(r) = -\frac{\rho}{r^2 \varepsilon_0}\int_0^r r'^2 dr' = -\frac{\rho r}{3\varepsilon_0} \equiv -\frac{1}{4\pi\varepsilon_0}\frac{Qr}{R^3}. \tag{1.45}$$

Since this derivative is nothing more than $-E(r)$, in this formula we can readily recognize our previous result (22). Now we may like to carry out the second integration to calculate the potential itself:

$$\phi(r) = -\frac{Q}{4\pi\varepsilon_0 R^3}\int_0^r r' dr' + c_1 = -\frac{Qr^2}{8\pi\varepsilon_0 R^3} + c_1. \tag{1.46}$$

[18] Named after Siméon Denis Poisson (1781-1840), also famous for the *Poisson distribution* – one of the central results of the probability theory – see, e.g., SM Sec. 5.2.

[19] Named after the famous mathematician (and astronomer) Pierre-Simon Laplace (1749-1827) who, together with Alexis Clairault, is credited for the development of the very concept of potential.

[20] See, e.g., MA Eq. (10.8) for $\partial/\partial\theta = \partial/\partial\varphi = 0$.

Before making any judgment on the integration constant c_1, let us solve the Poisson equation (in this case, just the Laplace equation) for the range outside the sphere ($r > R$):

$$\frac{1}{r^2}\frac{d}{dr}\left(r^2\frac{d\phi}{dr}\right) = 0 \cdot \tag{1.47}$$

Its first integral,

$$\frac{d\phi}{dr}(r) = \frac{c_2}{r^2}, \tag{1.48}$$

also gives the electric field (with the minus sign). Now using Eq. (45) and requiring the field to be continuous at $r = R$, we get

$$\frac{c_2}{R^2} = -\frac{Q}{4\pi\varepsilon_0 R^2}, \qquad \text{i.e. } \frac{d\phi}{dr}(r) = -\frac{Q}{4\pi\varepsilon_0 r^2}, \tag{1.49}$$

in an evident agreement with Eq. (19). Integrating this result again,

$$\phi(r) = -\frac{Q}{4\pi\varepsilon_0}\int\frac{dr}{r^2} = \frac{Q}{4\pi\varepsilon_0 r} + c_3, \qquad \text{for } r > R, \tag{1.50}$$

we can select $c_3 = 0$, so $\phi(\infty) = 0$, in accordance with the usual (though not compulsory) convention. Now we can finally determine the constant c_1 in Eq. (46) by requiring that this equation and Eq. (50) give the same value of ϕ at the boundary $r = R$. (According to Eq. (33), if the potential had a jump, the electric field at that point would be infinite.) The final answer may be represented as

$$\phi(r) = \frac{Q}{4\pi\varepsilon_0 R}\left(\frac{R^2 - r^2}{2R^2} + 1\right), \qquad \text{for } r \le R. \tag{1.51}$$

This calculation shows that using the Poisson equation to find the electrostatic potential distribution for highly symmetric problems may be a bit more cumbersome than directly finding the electric field – say, from the Gauss law. However, we will repeatedly see below that if the electric charge distribution is not fixed in advance, using Eq. (41) may be the only practicable way to proceed.

Returning now to the general theory of electrostatic phenomena, let us calculate the potential energy U of an arbitrary system of point electric charges q_k. Despite the apparently simple relation (31) between U and ϕ, the result is not that straightforward. Indeed, let us assume that the charge distribution has a finite spatial extent, so at large distances from it (formally, at $\mathbf{r} = \infty$) the electric field tends to zero, so the electrostatic potential tends to a constant. Selecting this constant, for convenience, to equal zero, we may calculate U as a sum of the energy increments ΔU_k created by bringing the charges, one by one, from infinity to their final positions $\mathbf{r}_k$ – see Fig. 6.[21] According to the integral form of Eq. (32), such a contribution is

$$\Delta U_k = -\int_{\infty}^{\mathbf{r}_k}\mathbf{F}(\mathbf{r})\cdot d\mathbf{r} = -q_k\int_{\infty}^{\mathbf{r}_k}\mathbf{E}(\mathbf{r})\cdot d\mathbf{r} \equiv q_k\phi(\mathbf{r}_k), \tag{1.52}$$

[21] Indeed, by the very definition of the potential energy of a system, it should not depend on the way we are arriving at its final configuration.

where $\mathbf{E}(\mathbf{r})$ is the total electric field, and $\phi(\mathbf{r})$ is the total electrostatic potential during this process, besides the field created by the very charge q_k that is being moved.

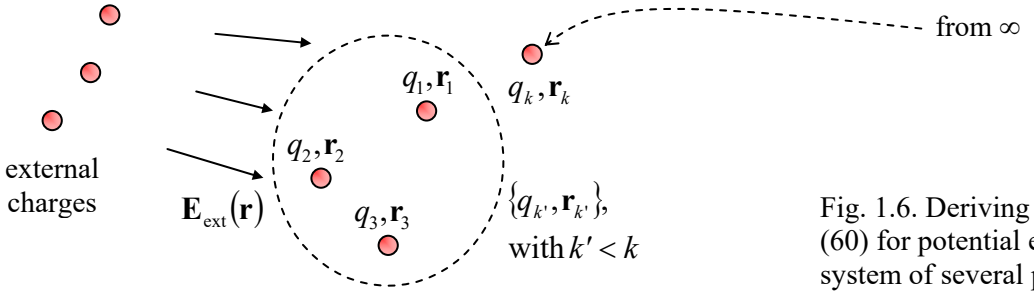


Fig. 1.6. Deriving Eqs. (55) and (60) for potential energies of a system of several point charges.

This expression shows that the increment ΔU_k, and hence the total potential energy U, depends on the source of the electric field $\mathbf{E}$. If the field is dominated by an *external field* $\mathbf{E}_{\text{ext}}$, induced by some *external* charges, not being a part of the charge configuration under our analysis (whose energy we are calculating, see Fig. 6), then the spatial distribution $\phi(\mathbf{r})$ is determined by this field, i.e. does not depend on how many charges we have already brought in, so Eq. (52) is reduced to

$$\Delta U_k = q_k \phi_{\text{ext}}(\mathbf{r}_k), \quad \text{where } \phi_{\text{ext}}(\mathbf{r}) \equiv -\int_{\infty}^{\mathbf{r}} \mathbf{E}_{\text{ext}}(\mathbf{r}') \cdot d\mathbf{r}'. \tag{1.53}$$

Summing up these contributions, we get what is called the charge system's energy *in* the external field:[22]

$$U_{\text{ext}} \equiv \sum_k \Delta U_k = \sum_k q_k \phi_{\text{ext}}(\mathbf{r}_k). \tag{1.54}$$

Now repeating the argumentation that has led us to Eq. (9), we see that for a continuously distributed charge, this sum turns into an integral:

$$\boxed{U_{\text{ext}} = \int \rho(\mathbf{r}) \phi_{\text{ext}}(\mathbf{r}) d^3 r}. \tag{1.55}$$

Energy: external field

(As was discussed above, using the delta-functional representation of point charges, we may always return from here to Eq. (54), so Eq. (55) may be considered as a final, universal result.)

The result is different in the opposite limit when the electric field $\mathbf{E}(\mathbf{r})$ is created only by the very charges whose energy we are calculating. In this case, $\phi(\mathbf{r}_k)$ in Eq. (52) is the potential created only by the charges with numbers $k' = 1, 2, \ldots, (k-1)$ that are already in place when the k^{th} charge is moved in (in Fig. 6, the charges inside the dashed boundary), and we may use the linear superposition principle to write

$$\Delta U_k = q_k \sum_{k'<k} \phi_{k'}(\mathbf{r}_k), \qquad \text{so that } U = \sum_k U_k = \sum_{\substack{k,k' \\ (k'<k)}} q_k \phi_{k'}(\mathbf{r}_k). \tag{1.56}$$

This result is so important that it is worthy of rewriting in several other forms. First, we may use Eq. (35) to represent Eq. (56) in a more symmetric form:

[22] An alternative, perhaps more accurate term for U_{ext} is the energy of the system's *interaction with* the external field.

$$U = \frac{1}{4\pi\varepsilon_0} \sum_{\substack{k,k' \\ (k'<k)}} \frac{q_k q_{k'}}{|\mathbf{r}_k - \mathbf{r}_{k'}|}. \tag{1.57}$$

The expression under this sum is evidently symmetric with respect to the index swap, so it may be extended into a different form,

$$U = \frac{1}{4\pi\varepsilon_0} \frac{1}{2} \sum_{\substack{k',k \\ (k'\neq k)}} \frac{q_k q_{k'}}{|\mathbf{r}_k - \mathbf{r}_{k'}|}, \tag{1.58}$$

where the interaction between each couple of charges is described by two equal terms under the sum, and the front coefficient ½ is used to compensate for this *double-counting*. The convenience of the last form is that it may be readily generalized to the continuous case:

$$U = \frac{1}{4\pi\varepsilon_0} \frac{1}{2} \int d^3r \int d^3r' \frac{\rho(\mathbf{r})\rho(\mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|}. \tag{1.59}$$

(As before, in this case, the restriction expressed in the discrete charge case as $k \neq k'$ is not important, because if the charge density is a continuous function, the integral (59) does not diverge at point $\mathbf{r} = \mathbf{r}'$.)

To represent this result in one more form, let us notice that according to Eq. (38), the inner integral over r' in Eq. (59), divided by $4\pi\varepsilon_0$, is just the full electrostatic potential at point $\mathbf{r}$, and hence

Energy: charge interaction

$$\boxed{U = \frac{1}{2} \int \rho(\mathbf{r})\phi(\mathbf{r}) d^3r}. \tag{1.60}$$

For the discrete charge case, this result is

$$U = \frac{1}{2} \sum_k q_k \phi(\mathbf{r}_k), \tag{1.61}$$

but here it is important to remember that the "full" potential's value $\phi(\mathbf{r}_k)$ should exclude the (infinite) contribution from the point charge k itself. Comparing the last two formulas with Eqs. (54) and (55), we see that the electrostatic energy of charge interaction within the system, as expressed via the charge-by-potential product, is twice less than that of the energy of charge interaction with a fixed ("external") field. This is the result of the fact that in the case of mutual interaction of the charges, the electric field $\mathbf{E}$ in the basic Eq. (52) is proportional to the charge's magnitude, rather than constant.[23]

Now we are ready to address an important conceptual question: can we locate this interaction energy in space? This task may seem trivial: Eqs. (58)-(61) seem to imply that non-zero contributions to U come only from the regions where the electric charges are located. However, one of the most beautiful features of physics is that sometimes completely different interpretations of the same mathematical result are possible. To get an alternative view of our current result, let us write Eq. (60) for a volume V so large that the electric field on the limiting surface S is negligible, and plug into it the charge density expressed from the Poisson equation (41):

$$U = -\frac{\varepsilon_0}{2} \int_V \phi \nabla^2 \phi \, d^3r. \tag{1.62}$$

[23] The nature of this additional factor ½ is absolutely the same as in the well-known formula $U = (\frac{1}{2})\kappa x^2$ for the potential energy of an elastic spring providing the returning force $F = -\kappa x$, proportional to its displacement x from the equilibrium position.

This expression may be integrated by parts as[24]

$$U = -\frac{\varepsilon_0}{2}\left[\oint_S \phi(\nabla\phi)_n\, d^2r - \int_V (\nabla\phi)^2\, d^3r\right]. \tag{1.63}$$

According to our condition of negligible field $\mathbf{E} = -\nabla\phi$ at the surface, the first integral vanishes, and we get a very important formula

$$U = \frac{\varepsilon_0}{2}\int (\nabla\phi)^2\, d^3r = \frac{\varepsilon_0}{2}\int E^2\, d^3r. \tag{1.64}$$

This result, if represented in the following equivalent form:[25]

$$U = \int u(\mathbf{r})\, d^3r, \quad \text{with } u(\mathbf{r}) \equiv \frac{\varepsilon_0}{2}E^2(\mathbf{r}), \tag{1.65}$$

Energy: electric field

certainly invites an interpretation very much different than Eq. (60): it is natural to consider $u(\mathbf{r})$ as the *spatial density of the electric field energy*, which is continuously distributed over all the space where the field exists – rather than just its part where the charges are located.

Let us have a look at how these two alternative pictures work for our testbed problem, a uniformly charged sphere. If we start with Eq. (60), we may limit the integration by the sphere volume ($0 \le r \le R$) where $\rho \ne 0$. Using Eq. (51), and the spherical symmetry of the problem (giving $d^3r = 4\pi r^2 dr$), we get

$$U = \frac{1}{2}4\pi\int_0^R \rho\phi\, r^2 dr = \frac{1}{2}4\pi\rho\frac{Q}{4\pi\varepsilon_0 R}\int_0^R\left(\frac{R^2 - r^2}{2R^2} + 1\right)r^2 dr = \frac{6}{5}\frac{1}{4\pi\varepsilon_0}\frac{Q^2}{2R}. \tag{1.66}$$

On the other hand, if we use Eq. (65), we need to integrate the energy density everywhere, i.e. both inside and outside of the sphere:

$$U = \frac{\varepsilon_0}{2}4\pi\left(\int_0^R E^2 r^2 dr + \int_R^\infty E^2 r^2 dr\right). \tag{1.67}$$

Using Eqs. (19) and (22) for, respectively, the external and internal regions, we get

$$U = \frac{\varepsilon_0}{2}4\pi\left[\int_0^R\left(\frac{Qr}{4\pi\varepsilon_0}\right)^2 r^2 dr + \int_R^\infty\left(\frac{Q}{4\pi\varepsilon_0 r^2}\right)^2 r^2 dr\right] = \left(\frac{1}{5}+1\right)\frac{1}{4\pi\varepsilon_0}\frac{Q^2}{2R}. \tag{1.68}$$

This is (fortunately :-) the same answer as given by Eq. (66), but to some extent, Eq. (68) is more informative because it shows how exactly the electric field's energy is distributed between the interior and exterior of the charged sphere.[26]

---

[24] This transformation follows from the divergence theorem MA (12.2) applied to the vector function $\mathbf{f} = \phi\nabla\phi$, taking into account the differentiation rule MA Eq. (11.4a): $\nabla\cdot(\phi\nabla\phi) = (\nabla\phi)\cdot(\nabla\phi) + \phi\nabla\cdot(\nabla\phi) = (\nabla\phi)^2 + \phi\nabla^2\phi$.

[25] In the Gaussian units, the standard replacement $\varepsilon_0 \to 1/4\pi$ turns the last of Eqs. (65) into $u(\mathbf{r}) = E^2/8\pi$.

[26] Note that $U \to \infty$ at $R \to 0$. Such divergence appears at the application of Eq. (65) to any point charge. Since it does not affect the force acting on the charge, the divergence does not create any technical difficulty for analysis of charge statics or non-relativistic dynamics, but it points to a possible conceptual problem of classical electrodynamics as a whole at describing point charges. This issue will be discussed at the very end of the course (Sec. 10.6).

We see that, as we could expect, within the realm of *electrostatics*, Eqs. (60) and (65) are equivalent. However, when we examine *electrodynamics* (in Chapter 6 and beyond), we will see that the latter equation is more general and that it is more adequate to associate the electric energy with the field itself rather than its sources – in our current case, the electric charges.

Finally, let us calculate the potential energy of a system of charges in the general case when both the internal interaction of the charges and their interaction with an external field are important. One might fancy that such a calculation should be very hard since, in both ultimate limits, when one of these interactions dominates, we have gotten different results. However, once again we get help from the almighty linear superposition principle: in the general case, for the total electric field we may write

$$\mathbf{E}(\mathbf{r}) = \mathbf{E}_{\text{int}}(\mathbf{r}) + \mathbf{E}_{\text{ext}}(\mathbf{r}), \qquad \phi(\mathbf{r}) = \phi_{\text{int}}(\mathbf{r}) + \phi_{\text{ext}}(\mathbf{r}), \tag{1.69}$$

where the index "int" now marks the field induced by the charge system under analysis, i.e. the variables participating (without indices) in Eqs. (56)-(65). Now let us imagine that our system is being built up in the following way: first, the charges are brought together at $\mathbf{E}_{\text{ext}} = 0$, giving the potential energy U_{int} expressed by Eq. (60), and then $\mathbf{E}_{\text{ext}}$ is slowly increased. Evidently, the energy contribution from the latter process cannot depend on the internal interaction of the charges, and hence may be expressed in the form (55). As a result, the total potential energy[27] is the sum of these two components:

$$U = U_{\text{int}} + U_{\text{ext}} = \frac{1}{2}\int \rho(\mathbf{r})\phi_{\text{int}}(\mathbf{r})d^3r + \int \rho(\mathbf{r})\phi_{\text{ext}}(\mathbf{r})d^3r\,. \tag{1.70}$$

Now making the transition from the potentials to the fields, absolutely similar to that performed in Eqs. (62)-(65), we may rewrite this expression as

$$U = \int u(\mathbf{r})d^3r, \qquad \text{with } u(\mathbf{r}) \equiv \frac{\varepsilon_0}{2}\left[E_{\text{int}}^2(\mathbf{r}) + 2\mathbf{E}_{\text{int}}(\mathbf{r})\cdot\mathbf{E}_{\text{ext}}(\mathbf{r})\right]. \tag{1.71}$$

One might think that this result, more general than Eq. (65) and perhaps less familiar to the reader, is something entirely new; however, it is not. Indeed, let us add to, and subtract $E_{\text{ext}}^2(\mathbf{r})$ from the sum in the brackets, and use Eq. (69) for the total electric field $\mathbf{E}(\mathbf{r})$; then Eq. (71) takes the form

$$U = \frac{\varepsilon_0}{2}\int E^2(\mathbf{r})d^3r - \frac{\varepsilon_0}{2}\int E_{\text{ext}}^2(\mathbf{r})d^3r\,. \tag{1.72}$$

Hence, in the most important case when we are using the potential energy to analyze the statics and dynamics of a system of charges in a fixed external field, i.e. when the second term on the right-hand side of Eq. (72) may be considered as a constant, we may still use for U an expression similar to the familiar Eq. (65), but with the field $\mathbf{E}(\mathbf{r})$ being the sum (69) of the internal and external fields.

Let us see how this works in a very simple situation. A uniform external electric field $\mathbf{E}_{\text{ext}}$ is applied normally to a very broad, plane layer that contains a very large and equal number of free electric charges of both signs – see Fig. 7. What is the equilibrium distribution of the charges over the layer?

[27] This total U (or rather its part dependent on our system of charges) is sometimes called the *Gibbs potential energy* of the system. (I will discuss this notion in detail in Sec. 3.5.)

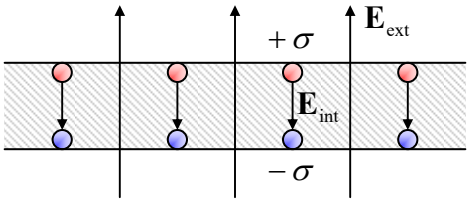


Fig. 1.7. A simple model of the electric field screening in a conductor. Here (and in all figures below) the red and blue colors are used to denote the opposite charge signs.

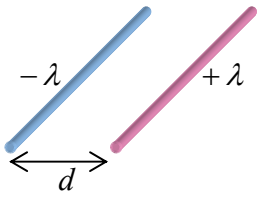
Since any area-uniform distribution of the charge inside the layer does not affect the field (and hence its energy) outside it, and the equilibrium distribution has to minimize the total potential energy of the system, Eq. (72) immediately gives the answer: the distribution should provide $\mathbf{E} \equiv \mathbf{E}_{int} + \mathbf{E}_{ext} = 0$ inside the whole layer – the effect called the electric field *screening*. The only way to ensure this equality is to have enough free charges of opposite signs residing on the layer's surfaces to induce a uniform field $\mathbf{E}_{int} = -\mathbf{E}_{ext}$, exactly compensating the external field at each point inside the layer – see Fig. 7. According to Eq. (24), the areal density of these surface charges should equal $\pm\sigma$, with $\sigma = E_{ext}/\varepsilon_0$. This is a rudimentary but reasonable model of conductors' *polarization* – to be discussed in detail in the next chapter.

1.4. Exercise problems

1.1. Calculate the electric field of a thin, long, straight filament, electrically charged with a constant linear density λ, by using two approaches:

(i) directly from the Coulomb law, and
(ii) from the Gauss law.

1.2. Two thin, straight, parallel filaments separated by distance d carry equal and opposite uniformly distributed charges with linear density λ – see the figure on the right. Calculate the force (per unit length) of the Coulomb interaction of the filaments. Compare its functional dependence on d with the Coulomb law for two point charges, and interpret their difference.



1.3. Calculate the electric field of the following spherically symmetric charge distribution: $\rho(r) = \rho_0 \exp\{-\lambda r\}$.

1.4. A sphere of radius R, whose volume had been charged with a constant density ρ, is split with a very narrow planar gap passing through its center. Calculate the force of the mutual electrostatic repulsion of the resulting two hemispheres.

1.5. A thin spherical shell of radius R, which had been charged with a constant areal density σ, is split into two equal halves with a very thin planar cut passing through the sphere's center. Calculate the force of electrostatic repulsion between the resulting hemispheric shells, and compare the result with that of the previous problem.

1.6. Calculate the spatial distribution of the electrostatic potential created by a straight thin filament of a finite length $2l$, charged with a constant linear density λ, and explore the result in the limits of very small and very large distances from the filament.

1.7. A thin planar sheet, perhaps of an irregular shape, carries an electric charge with a constant areal density σ.

(i) Express the electric field's component normal to the plane, at a certain distance from it, via the solid angle Ω at which the sheet is visible from the observation point.

(ii) Use the result to calculate the field in the center of a cube with one face charged with a constant density σ.

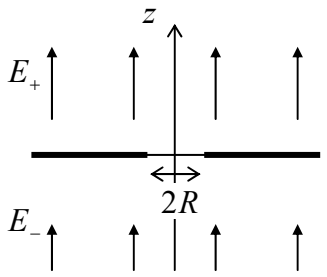
1.8. Can one create, in an extended region of space, electrostatic fields with the Cartesian components proportional to the following products of the Cartesian coordinates $\{x, y, z\}$:

(i) $\{yz, xz, xy\}$,

(ii) $\{xy, xy, yz\}$?

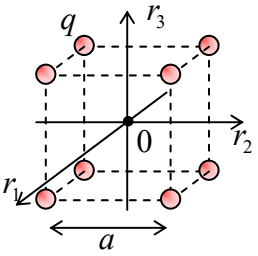
1.9. Distant sources have been used to create different uniform electrostatic fields in two half-spaces:

$$\mathbf{E}(\mathbf{r})\big|_{r>>R} = \mathbf{n}_z \times \begin{cases} E_+, & \text{at } z < 0, \\ E_-, & \text{at } z > 0, \end{cases}$$



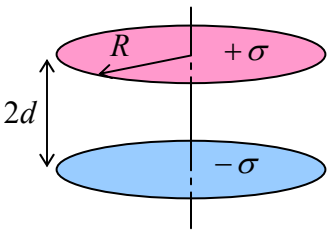
except for a transitional region of scale R near the origin, where the field is perturbed but still axially symmetric. (As will be discussed in the next chapter, this may be done, for example, using a thin conducting membrane with a round hole of radius R in it – see the figure on the right.) Prove that such field may serve as an electrostatic lens for charged particles flying along the z-axis, at distances $\rho << R$ from it, and calculate the focal distance f of this lens. Spell out the conditions of validity of your result.

1.10. Eight equal point charges q are located in the corners of a cube of side a. Calculate all Cartesian components E_j of the electric field, and their spatial derivatives $\partial E_j/\partial r_{j'}$, in the cube's center, where r_j are the Cartesian coordinates oriented along the cube's sides – see the figure on the right. Are all of your results valid for the center of a planar square, with four equal charges at its corners?



1.11. By a direct calculation, find the average electric potential of a spherical surface of radius R, created by a point charge q located at a distance $r > R$ from the sphere's center. Use the result to prove the following general *mean value theorem*: the electric potential at any point is always equal to its average value on any spherical surface with the center at that point while containing no electric charges inside it.

1.12. Two similar thin, circular, coaxial disks of radius R, separated by distance $2d$, are uniformly charged with equal and opposite areal densities $\pm\sigma$ – see the figure on the right. Calculate and sketch the distribution of the electrostatic potential and the electric field of the disks along their common axis.



1.13. The electrostatic potential, created by some electric charge distribution, is

$$\phi(\mathbf{r}) = C\left(\frac{1}{r} + \frac{1}{2r_0}\right)\exp\left\{-\frac{r}{r_0}\right\},$$

where C and r_0 are constants, and $r \equiv |\mathbf{r}|$ is the distance from the origin. Calculate the charge distribution in space.

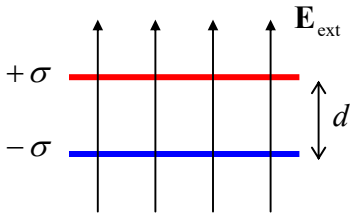
1.14. A thin, flat, rectangular sheet of size $a \times b$ is electrically charged with a constant areal density σ. Without an explicit calculation of the spatial distribution $\phi(\mathbf{r})$ of the electrostatic potential induced by this charge, find the ratio of its values in the center and in the corners of the rectangle.

Hint: Consider partitioning the rectangle into several similar parts and using the linear superposition principle.

1.15. Calculate the electrostatic energy per unit area of the system of two thin, parallel planes with equal and opposite charges of a constant areal density σ, separated by distance d.

1.16. The system analyzed in the previous problem (two thin, parallel, oppositely charged planes) is now placed into an external, uniform, normal electric field $E_{ext} = \sigma/\varepsilon_0$ – see the figure on the right. Find the force (per unit area) acting on each plane, by two methods:

(i) directly from the electric field distribution, and
(ii) from the potential energy of the system.



1.17. Explore the relationship between the Laplace equation (42) and the minimum of the electrostatic field energy (65).

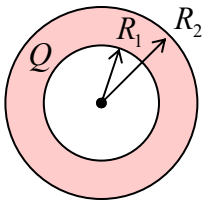
1.18. Prove the following *reciprocity theorem of electrostatics*:[28] if two spatially-confined charge distributions $\rho_1(\mathbf{r})$ and $\rho_2(\mathbf{r})$ create, respectively, electrostatic potentials $\phi_1(\mathbf{r})$ and $\phi_2(\mathbf{r})$, then

$$\int \rho_1(\mathbf{r})\phi_2(\mathbf{r})d^3r = \int \rho_2(\mathbf{r})\phi_1(\mathbf{r})d^3r\,.$$

Hint: Consider the integral $\int \mathbf{E}_1 \cdot \mathbf{E}_2 d^3r$.

1.19. Calculate the energy of the electrostatic interaction of two spheres, of radii R_1 and R_2, each with a spherically symmetric charge distribution, separated by distance $d > R_1 + R_2$.

1.20. Calculate the electrostatic energy U of a (generally, thick) spherical shell, with charge Q uniformly distributed through its volume – see the figure on the right. Interpret the dependence of U on the inner cavity's radius R_1, at fixed Q and R_2.



[28] This is only the simplest one of several reciprocity theorems in electromagnetism – see, e.g., Sec. 6.8 below.

Chapter 2. Charges and Conductors

This chapter starts our discussion of the very common situations when the electric charge distribution in space is not known a priori, but rather should be calculated in a self-consistent way together with the electric field it creates. The simplest situations of this kind involve conductors and lead to the so-called boundary problems in that the partial differential equations describing the field distribution have to be solved with appropriate boundary conditions. Such problems are also typical for other parts of electrodynamics (*and indeed for other fields of physics as well*)*, so following tradition, I will use this chapter's material as a playground for a discussion of various methods of boundary problem solution, and the special functions most frequently encountered on that way.*

2.1. Polarization and screening

The basic principles of electrostatics outlined in Chapter 1 present the conceptually full solution of the problem of finding the electrostatic field (and hence Coulomb forces) induced by electric charges distributed over space with some density $\rho(\mathbf{r})$. However, in most practical situations, this function is not known but should be found self-consistently with the field. For example, if a sample of relatively dense material is placed into an external electric field, it is typically *polarized*, i.e. acquires some local charges of its own, which contribute to the total electric field $\mathbf{E}(\mathbf{r})$ inside, and even outside it – see Fig. 1a.

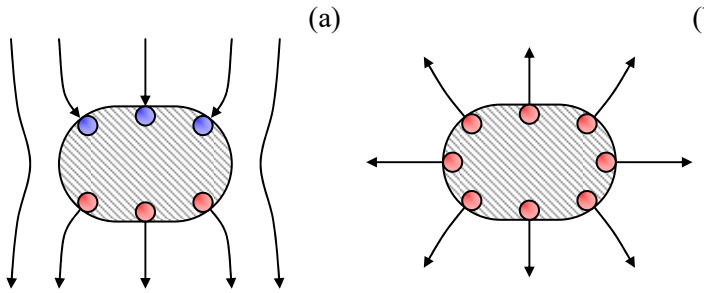


Fig. 2.1. Two typical electrostatic situations involving conductors: (a) polarization by an external field, and (b) re-distribution of the conductor's own charge over its surface – schematically. Here and below, the red and blue points denote charges of opposite signs.

The full solution of such problems should satisfy not only the fundamental Eq. (1.7) but also the so-called *constitutive relations* between the macroscopic variables describing the sample's material. Due to the atomic character of real materials, such relations may be very involved. In this part of my series, I will have time to address these relations, for various materials, only rather superficially,[1] focusing on their simple approximations. Fortunately, in most practical cases such approximations work very well.

In particular, for the polarization of good conductors, a very reasonable approximation is given by the so-called *macroscopic model*, in which the free charges in the conductor are treated as a charged continuum that is free to move under the effect of the force $\mathbf{F} = q\mathbf{E}$ exerted by the *macroscopic* electric field $\mathbf{E}$, i.e. the field averaged over space on the atomic scale – see also the discussion at the end of Sec.

[1] A more detailed discussion of the electrostatic field screening may be found, e.g., in SM Sec. 6.4. (Alternatively, see either Sec. 13.5 of J. Hook and H. Hall, *Solid State Physics*, 2nd ed., Wiley, 1991; or Chapter 17 of N. Ashcroft and N. Mermin, *Solid State Physics*, Brooks Cole, 1976.)

1.1. In electrostatics (which excludes the case dc currents, to be discussed in Chapter 4 below), there should be no such motion, so everywhere inside the conductor the macroscopic electric field should vanish:

$$\mathbf{E} = 0\,. \tag{2.1a}$$

This is the *electric field screening*[2] effect, meaning, in particular, that conductors' polarization in an external electric field has the extreme form shown (rather schematically) in Fig. 1a, with the field of the induced surface charges completely compensating the external field in the conductor's bulk. Note that Eq. (1a) may be rewritten in another, frequently more convenient form:

Conductor: macroscopic model

$$\phi = \text{const}\,, \tag{2.1b}$$

where ϕ is the *macroscopic* electrostatic potential related to the macroscopic field by Eq. (1.33).[3] (If a problem includes several unconnected conductors, the constant in Eq. (1b) may be specific for each of them.)

Now let us examine what we can say about the electric field in free space just *outside* a conductor, within the same macroscopic model. At close proximity, any smooth surface (in our current case, that of a conductor) looks planar. Let us integrate Eq. (1.28) over a narrow ($d << l$) rectangular loop C encircling a part of such plane conductor's surface (see the dashed line in Fig. 2a), and apply it to the electric field vector $\mathbf{E}$ the well-known vector algebra equality – the *Stokes theorem*[4]

$$\int_S (\nabla \times \mathbf{E})_n \, d^2 r = \oint_C \mathbf{E} \cdot d\mathbf{r}\,, \tag{2.2}$$

where S is any surface limited by the contour C.

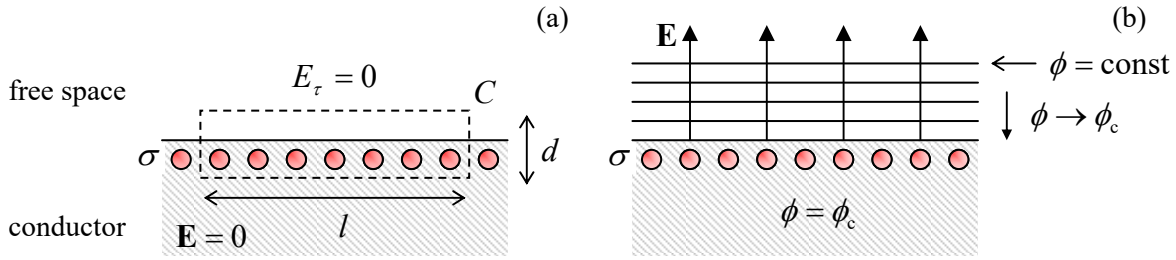


Fig. 2.2. (a) The surface charge layer at a conductor's surface, and (b) the electric field lines and equipotential surfaces near it.

In our current case, the contour is dominated by two straight lines of length l, so if l is much smaller than the characteristic spatial scale of the field's changes but much larger than the interatomic distances, the right-hand side of Eq. (2) may be well approximated as $[(E_\tau)_{\text{in}} - (E_\tau)_{\text{out}}]\, l$, where E_τ is the tangential component of the corresponding macroscopic field, parallel to the surface. On the other hand, according to Eq. (1.28), the left-hand side of Eq. (2) equals zero. Hence, the macroscopic field's

[2] This term, used for the *electric* field, should not be confused with *shielding* – the term used for the description of *magnetic* field's reduction by magnetic materials – see Chapter 5 below.

[3] Since averaging of a function over space is a linear operation, any linear relation between genuine (microscopic) variables, including Eq. (1.33), is also valid for the corresponding macroscopic variables.

[4] See, e.g., MA Eq. (12.1).

component E_τ should be continuous at the surface, and to satisfy Eq. (1a) inside the conductor, the component has to vanish immediately outside it: $(E_\tau)_{\text{out}} = 0$. This means that the electrostatic potential immediately outside of a conducting surface cannot change along it. In other words, the equipotential surfaces outside a conductor should "lean" to the conductor's surface, with their potential values approaching the constant potential of the constant potential of the conductor – see Fig. 2b.

So, the electrostatic field just outside any conductor has to be normal to its surface. To find this normal field, we may apply the universal relation (1.24) to our macroscopic field **E**. Since in our current case $E_n = 0$ inside the conductor, we get

Surface charge density

$$\sigma = \varepsilon_0 \left(E_n\right)_{\text{out}} \equiv -\varepsilon_0 \left(\nabla \phi\right)_n \equiv -\varepsilon_0 \frac{\partial \phi}{\partial n}, \tag{2.3}$$

where σ is the macroscopic areal density of the conductor's surface charge. Note that deriving this universal relation between the normal component of the field and the surface charge density, we have not used any cause-vs-effect arguments, so Eq. (3) is valid regardless of whether the surface charge is induced by an externally applied field (as in the case of conductor's polarization, shown in Fig. 1a), or the electric field is induced by the electric charge placed on the conductor and then self-redistributed over its surface (Fig. 1b), or it is some combination of both effects.

Before starting to use the macroscopic model for the solution of particular problems of electrostatics, let me use the balance of this section to briefly discuss its limitations. (The reader in a rush may skip this discussion and proceed to Sec. 2; however, I believe that every educated physicist has to understand when this model works, and when it does not.)

Since the argumentation which has led us to Eq. (1.24) and hence to Eq. (3) is valid for any thickness d of the Gauss pillbox, within the macroscopic model, the whole surface charge is located within an infinitely thin surface layer. This is of course impossible physically: for one, this would require an infinite volumic density ρ of the charge. In reality, the charged layer (and hence the region of the electric field's crossover from the finite value (3) to zero) has a nonzero thickness λ. At least three effects contribute to λ.

(i) *Atomic structure of matter*. Within each atom, and frequently between the adjacent atoms as well, the genuine ("microscopic") electric field is highly non-uniform. Thus, as was already stated above, Eq. (1) is valid only for the *macroscopic field*, i.e. the field averaged over distances of the order of the atomic size scale $a_0 \sim 10^{-10}$ m,[5] and cannot be applied to the field changes on that scale. As a result, the surface layer of charges cannot be much thinner than a_0.

(ii) *Thermal excitation*. According to Eq. (1.9), in the whole field-free bulk of a conductor, the net charge density, $\rho = e(n - n_e)$, [6] has to vanish, so the numbers of protons in atomic nuclei (n) and electrons (n_e) per unit volume have to be balanced. However, if an external electric field penetrates a conductor, free electrons can shift in or out of its affected part, depending on the field's contribution to their potential energy, $\Delta U = q_e\phi = -e\phi$. (Here the arbitrary constant in ϕ is chosen to give $\phi = 0$ well inside the conductor.) In classical statistics, this change is described by the Boltzmann distribution:[7]

[5] This scale originates from the quantum-mechanical effects of electron motion, characterized by the *Bohr radius* $r_B \equiv \hbar^2/m_e(e^2/4\pi\varepsilon_0) \approx 0.53\times10^{-10}$ m – see, e.g., QM Eq. (1.10). It also defines the scale $E_B = e/4\pi\varepsilon_0 r_B^2 \sim 10^{12}$ SI units (V/m) of the microscopic electric fields inside atoms. (Please note how large these fields are.)

[6] In this series, e denotes the fundamental charge, $e \approx 1.6\times10^{-19}$ C > 0, so that the electron's charge equals $(-e)$.

[7] See, e.g., SM Sec. 3.1.

$$n_e(\mathbf{r}) = n\exp\left\{-\frac{U(\mathbf{r})}{k_B T}\right\}, \tag{2.4}$$

where T is the absolute temperature in kelvins (K), and $k_B \approx 1.38\times10^{-23}$ J/K is the Boltzmann constant. As a result, the net charge density is

$$\rho(\mathbf{r}) = en\left(1-\exp\left\{\frac{e\phi(\mathbf{r})}{k_B T}\right\}\right). \tag{2.5}$$

The penetrating electric field polarizes the atoms as well. As will be discussed in the next chapter, such polarization results in the reduction of the electric field by a material-specific dimensionless factor κ (larger, but typically not too much larger than 1), called the *dielectric constant*. As a result, the Poisson equation (1.41) takes the so-called *Poisson-Boltzmann* form,[8]

$$\frac{d^2\phi}{dz^2} = -\frac{\rho}{\kappa\varepsilon_0} = \frac{en}{\kappa\varepsilon_0}\left(\exp\left\{\frac{e\phi}{k_B T}\right\}-1\right), \tag{2.6}$$

where we have taken advantage of the 1D geometry of the system to simplify the Laplace operator, with the z-axis normal to the surface.

Even with this simplification, Eq. (6) is a nonlinear differential equation allowing an analytical but rather bulky solution. Since our current goal is just to estimate the field penetration depth λ, let us simplify the equation further by considering the low-field limit: $e|\phi| \sim e|E|\lambda << k_B T$. In this limit, we may extend the exponent into the Taylor series, and keep only two leading terms (of which the first one cancels with the following unity). As a result, Eq. (6) becomes linear,

$$\frac{d^2\phi}{dz^2} = \frac{en}{\varepsilon\varepsilon_0}\frac{e\phi}{k_B T}, \qquad \text{i.e. } \frac{d^2\phi}{dz^2} = \frac{1}{\lambda^2}\phi, \tag{2.7}$$

where the constant λ, in this case, is called the *Debye* (or "*Debye-Hückel*") *screening length* λ_D:

$$\lambda_D^2 \equiv \frac{\kappa\varepsilon_0 k_B T}{e^2 n}. \tag{2.8}$$

Debye screening length

As the reader certainly knows, Eq. (7) describes an exponential decrease of the electric potential, with the characteristic length λ_D: $\phi \propto \exp\{-z/\lambda_D\}$, where the z-axis is directed into the conductor. Plugging in the involved fundamental constants into Eq. (8), we get the following estimate: $\lambda_D[\text{m}] \approx 70\times(\kappa\times T[\text{K}]/n[\text{m}^{-3}])^{1/2}$. According to this formula, in semiconductors at room temperature, the Debye length may be rather substantial. For example, in silicon ($\kappa \approx 12$) doped to the free charge carrier concentration $n = 3\times10^{18}$ cm^{-3} (the value typical for modern integrated circuits),[9] $\lambda_D \approx 2$ nm, still well

[8] This equation and/or its straightforward generalization to the case of charged particles (ions) of several kinds is also (especially in the theories of electrolytes and plasmas) called the *Debye-Hückel equation*.

[9] There is a good reason for making an estimate of λ_D for this case: the electric field created by the gate electrode of a field-effect transistor, penetrating into doped silicon by a depth ~λ_D, controls the electric current in this most important electronic device – on whose back all our information technology rides. Because of that, λ_D establishes the possible scale of semiconductor circuit shrinking, which is the basis of the well-known Moore's law. (Practically, the scale is determined by integrated circuit patterning techniques, and Eq. (8) may be used to find the proper charge carrier density n and hence the necessary level of silicon doping – see, e.g., SM Sec. 6.4.)

above the atomic size scale a_0, thus justifying the estimate However, for typical good metals ($n \sim 10^{29}$ m^{-3}, $\kappa \sim 10$) the same formula gives $\lambda_D \sim 10^{-11}$ m, less than a_0. In this case, Eq. (8) should not be taken literally, because it is based on the assumption of a continuous charge distribution.

(iii) *Quantum statistics*. Actually, the last estimate is not valid for good metals (and highly doped semiconductors) for one more reason: their free electrons obey the quantum (*Fermi-Dirac*) statistics rather than the Boltzmann distribution (4).[10] As a result, at all realistic temperatures, the electrons form a degenerate quantum gas, occupying all available energy states below some energy level $\mathscr{E}_F >> k_B T$, called the *Fermi energy*. In these conditions, the screening of a relatively low electric field may be described by replacing Eq. (5) with

$$\rho \equiv e(n - n_e) = eg(\mathscr{E}_F)(-U) = -e^2 g(\mathscr{E}_F)\phi, \tag{2.9}$$

where $g(\mathscr{E})$ is the density of quantum states (per unit volume per unit energy) at the electron's energy $\mathscr{E}$. At the Fermi surface, the density is of the order of $n/\mathscr{E}_F$.[11] As a result, we again get the second of Eqs. (7), but with a different characteristic scale λ, defined by the following relation:

Thomas-Fermi screening length

$$\lambda_{TF}^2 \equiv \frac{\kappa\varepsilon_0}{e^2 g(\mathscr{E}_F)} \sim \frac{\kappa\varepsilon_0 \mathscr{E}_F}{e^2 n}, \tag{2.10}$$

and called the *Thomas-Fermi screening length*. Since for most good metals, n is of the order of 10^{29} m^{-3}, and $\mathscr{E}_F$ is of the order of 10 eV, Eq. (10) typically gives λ_{TF} close to a few a_0, and makes the Thomas-Fermi screening theory valid at least semi-quantitatively.

To summarize, the electric field penetration into good conductors is limited to a depth λ ranging from a fraction of a nanometer to a few nanometers, so for problems with a characteristic linear size much larger than that scale, the macroscopic model (1) gives very good accuracy, and we will use them in the rest of this chapter. However, the reader should remember that in many situations involving semiconductors, as well as at some nanoscale experiments with metals, the electric field penetration should be taken into account.

Another important condition of the macroscopic model's validity is imposed on the electric field's magnitude, which is especially significant for semiconductors. Indeed, as Eq. (6) shows, Eq. (7) is only valid if $e|\phi| << k_B T$, so $|E| \sim |\phi|/\lambda_D$ should be much lower than $k_B T/e\lambda_D$. In the example given above ($\lambda_D \approx 2$ nm, $T = 300$ K), this means $|E| << E_t \sim 10^7$V/m $\equiv 10^5$V/cm – the value readily reachable in the lab. In larger fields, the field penetration becomes nonlinear, leading in particular to the very important effect of *carrier depletion*; it will be discussed in SM Sec. 6.4. For typical metals, such linearity limit, $E_t \sim \mathscr{E}_F/e\lambda_{TF}$ is much higher, $\sim 10^{11}$ V/m, but the model may be violated at lower fields by other effects, such as the impact-ionization leading to *electric breakdown*, which may start at $\sim 10^6$ V/m.

2.2. Capacitance

Let us start using the macroscopic model from systems consisting of charged conductors only, with no so-called *stand-alone* charges in the free space outside them.[12] Our goal here is to calculate the

[10] See, e.g., SM Sec. 2.8. For a more detailed derivation of Eq. (10), see SM Chapter 3.
[11] See, e.g., SM Sec. 3.3.

distributions of the electric field **E** and potential ϕ in space, and the distribution of the surface charge density σ over the conductor surfaces. However, before doing that for particular situations, let us see if there are any integral measures of these distributions, which should be our primary focus.

The simplest case is of course a single conductor in the otherwise free space. According to Eq. (1b), all its volume should have the same electrostatic potential ϕ, evidently providing one convenient global measure of the situation. Another integral measure is provided by the total charge

$$Q \equiv \int_V \rho\, d^3r \equiv \oint_S \sigma\, d^2r\,, \tag{2.11}$$

where the last integral is extended over the whole surface S of the conductor. In the general case, what can we tell about the relation between Q and ϕ? At $Q = 0$, there is no electric field in the system, and it is natural (though not absolutely necessary) to select the arbitrary constant in the electrostatic potential to have $\phi = 0$ everywhere. Then, if the conductor is charged with a non-zero Q, according to the linear Eq. (1.7), the electric field at any point of space has to be proportional to that charge. Hence the electrostatic potential at all points, including its value ϕ inside the conductor, is also proportional to Q:

$$\phi = \wp Q\,. \tag{2.12}$$

The proportionality coefficient $\wp$, which depends on the conductor's size and shape, but on neither ϕ nor Q, is called its *reciprocal capacitance* (or, not too often, "electric elastance"). Usually, Eq. (12) is rewritten in a different form,

$$\boxed{Q = C\phi, \quad \text{with } C \equiv \frac{1}{\wp},} \tag{2.13}$$

Self-capacitance

where C is called *self-capacitance*. (Frequently, C is called just *capacitance*, but as we will see very soon, for more complex situations the latter term may be ambiguous.)

Before calculating C for particular geometries, let us have a look at the electrostatic energy U of a single conductor. To calculate it, of the several relations discussed in Chapter 1, Eq. (1.61) is most convenient, because all elementary charges q_k are now parts of the conductor charge, and hence reside at the same potential ϕ – see Eq. (1b) again. As a result, the equality becomes very simple:

$$U = \frac{1}{2}\phi \sum_k q_k \equiv \frac{1}{2}\phi Q\,. \tag{2.14}$$

Moreover, using the linear relation (13), the same result may be re-written in two more forms:

$$\boxed{U = \frac{Q^2}{2C} = \frac{C}{2}\phi^2.} \tag{2.15}$$

Electro-static energy

We will discuss several ways to calculate C in the next sections, and right now will have a quick look at just the simplest example for that we have calculated everything necessary in the previous chapter: a conducting sphere of radius R. Indeed, we already know the electric field distribution: according to Eq. (1), $E = 0$ inside the sphere, while Eq. (1.19), with $Q(r) = Q$, describes the field distribution outside it, because of the evident spherical symmetry of the surface charge distribution.

[12] In some texts, these charges are called "free". This term is somewhat misleading, because they may well be bound, i.e. unable to move freely.

Moreover, since the latter formula is exactly the same as for the point charge placed in the sphere's center, the potential's distribution in space may be obtained from Eq. (1.35) by replacing q with the sphere's full charge Q. Hence, on the surface of the sphere (and, according to Eq. (1b), through its interior),

$$\phi = \frac{1}{4\pi\varepsilon_0}\frac{Q}{R}. \tag{2.16}$$

Comparing this result with the definition (13), for the sphere's self-capacitance we obtain a very simple formula[13]

C: shere

$$\boxed{C = 4\pi\varepsilon_0 R}. \tag{2.17}$$

This formula, which should be well familiar to the reader, is convenient to get some feeling of how large the SI unit of capacitance (1 *farad*, abbreviated as F) is: the self-capacitance of Earth ($R_E \approx 6.34\times10^6$ m) is below 1 mF! Another important note is that while Eq. (17) is not exactly valid for a conductor of arbitrary shape, it implies an important general estimate

$$C \sim 2\pi\varepsilon_0 a \tag{2.18}$$

where a is the scale of the linear size of any conductor.[14]

Now proceeding to a system of *two* arbitrary conductors, we immediately see why we should be careful with the capacitance definition: one constant C is insufficient to describe all electrostatic properties of such a system. Indeed, here we have two, generally different conductor potentials, ϕ_1 and ϕ_2, that may depend on both conductor charges, Q_1 and Q_2. Using the same arguments as for the single-conductor case, we may conclude that the dependence is always linear:

$$\begin{aligned} \phi_1 &= \wp_{11}Q_1 + \wp_{12}Q_2, \\ \phi_2 &= \wp_{21}Q_1 + \wp_{22}Q_2, \end{aligned} \tag{2.19}$$

but now has to be described by more than one coefficient. Actually, it turns out that there are three rather than four different coefficients in these relations, because

$$\wp_{12} = \wp_{21}. \tag{2.20}$$

This equality may be proved in several ways, for example, using the general *reciprocity theorem of electrostatics* (whose proof was the subject of Problem 1.17):

$$\int \rho_1(\mathbf{r})\phi^{(2)}(\mathbf{r})d^3r = \int \rho_2(\mathbf{r})\phi^{(1)}(\mathbf{r})d^3r, \tag{2.21}$$

[13] In the Gaussian units, using the standard replacement $4\pi\varepsilon_0 \to 1$, this relation takes an even simpler form: $C = R$, very easy to remember. Generally, in the Gaussian units (but not in the SI system!) the capacitance has the dimensionality of length, i.e. is measured in centimeters. Note also that a fractional SI unit, 1 picofarad (10^{-12} F), is very close to the Gaussian unit: 1 pF = $[(1\times10^{-12})/(4\pi\varepsilon_0\times10^{-2})]$ cm ≈ 0.8998 cm. So, 1 pF is close to the capacitance of a metallic ball with a 1-cm radius, making this unit very convenient for human-scale systems.

[14] These arguments are somewhat insufficient to say which size should be used for a in the case of narrow, extended conductors, e.g., a thin, long wire. Very soon we will see that in such cases the electrostatic energy, and hence C, depends mostly on the *larger* size of the conductor.

where $\phi^{(1)}(\mathbf{r})$ and $\phi^{(2)}(\mathbf{r})$ are the potential distributions induced, respectively, by two electric charge distributions, $\rho_1(\mathbf{r})$ and $\rho_2(\mathbf{r})$. In our current case, each of these integrals is limited to the volume (or, more exactly, the surface) of the corresponding conductor, where each potential is constant and may be taken out of the integral. As a result, Eq. (21) is reduced to

$$Q_1\phi^{(2)}(\mathbf{r}_1) = Q_2\phi^{(1)}(\mathbf{r}_2). \tag{2.22}$$

In terms of Eq. (19), $\phi^{(2)}(\mathbf{r}_1)$ is just $\wp_{12}Q_2$, while $\phi^{(1)}(\mathbf{r}_2)$ equals $\wp_{21}Q_1$. Plugging these expressions into Eq. (22), and canceling the product Q_1Q_2, we arrive at Eq. (20).

Hence the 2×2 matrix of coefficients $\wp_{jj'}$ (called the *reciprocal capacitance matrix*) is always symmetric, and using the natural notation $\wp_{11} \equiv \wp_1$, $\wp_{22} \equiv \wp_2$, $\wp_{12} = \wp_{21} \equiv \wp$, we may rewrite it in a simpler form:

$$\begin{pmatrix} \wp_1 & \wp \\ \wp & \wp_2 \end{pmatrix}. \tag{2.23}$$

Plugging the relation (19), in this new notation, into Eq. (1.61), we see that the full electrostatic energy of the system may be expressed as a quadratic form of its charges:

$$U = \frac{\wp_1}{2}Q_1^2 + \wp Q_1Q_2 + \frac{\wp_2}{2}Q_2^2. \tag{2.24}$$

It is evident that the middle term on the right-hand side of this equality describes the electrostatic coupling of the conductors. (Without it, the energy would be just a sum of two independent electrostatic energies of conductors 1 and 2.)[15] Still, even with this simplification, Eqs. (19) and (20) show that in the general case of arbitrary charges Q_1 and Q_2, the system of two conductors should be characterized by three, rather than just one coefficient ("the capacitance"). This is why we may attribute a single capacitance to the system only in some particular cases.

For practice, the most important of them is when the system as the whole is electrically neutral: $Q_1 = -Q_2 \equiv Q$. In this case, the most important function of Q is the difference between the conductors' potentials, called the *voltage*:[16]

$$\boxed{V \equiv \phi_1 - \phi_2,} \tag{2.25}$$

Voltage: definition

For that function, the subtraction of two Eqs. (19) gives

$$\boxed{V = \frac{Q}{C}, \quad \text{with } C \equiv \frac{1}{\wp_1 + \wp_2 - 2\wp},} \tag{2.26}$$

Mutual capacitance

where the coefficient C is called the *mutual capacitance* between the conductors – or, again, just "capacitance" if the term's meaning is absolutely clear from the context. The same coefficient describes

[15] This is why systems with $\wp \ll \wp_1, \wp_2$ are called *weakly coupled*, and may be analyzed using approximate methods – see, e.g., Fig. 4 and its discussion below.

[16] A word of caution: in condensed matter physics and electrical engineering, voltage is most commonly defined as the difference between *electrochemical* rather than *electrostatic* potentials. These two notions coincide if the conductors have equal *workfunctions* – for example, if they are made of the same material. In this course, this condition will be implied, and the difference between the two voltages ignored – to be discussed in detail in SM Sec. 6.3.

the electrostatic energy of the system. Indeed, plugging Eqs. (19) and (20) into Eq. (24), we see that both forms of Eq. (15) are reproduced if ϕ is replaced with V, Q_1 with Q, and with C meaning the mutual capacitance:

Capacitor's energy

$$U = \frac{Q^2}{2C} = \frac{C}{2}V^2. \tag{2.27}$$

The best-known system for which the mutual capacitance C may be readily calculated is the *plane* (or "parallel-plate") *capacitor*: a system of two conductors separated with a narrow plane gap of a constant thickness d and an area $A \sim a^2 >> d^2$ – see Fig. 3.

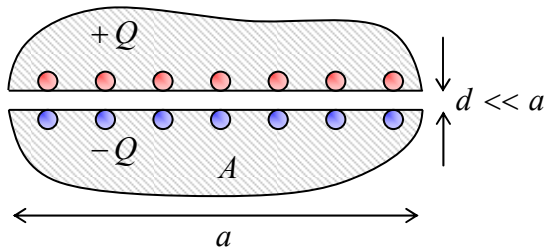


Fig. 2.3. Plane capacitor – schematically.

Since the surface charges that contribute to the opposite charges $\pm Q$ of the conductors of this system, attract each other, in the limit $d << a$ they sit entirely on the opposite surfaces limiting the gap, so there is virtually no electric field outside of the gap, while (according to the discussion in Sec. 1) inside the gap it is normal to the surfaces. According to Eq. (3), the magnitude of this field is $E = \sigma/\varepsilon_0$. Integrating this field across thickness d of the narrow gap, we get $V \equiv \phi_1 - \phi_2 = Ed = \sigma d/\varepsilon_0$, so $\sigma = \varepsilon_0 V/d$. However, due to the constancy of the potential of each electrode, V should not depend on the position in the gap area. As a result, σ should be also constant over all the gap area A, regardless of the external geometry of the conductors (see Fig. 3 again), and hence $Q = \sigma A = \varepsilon_0 V/d$. Thus we may write $V = Q/C$, with

C: Plane capacitor

$$C = \frac{\varepsilon_0 A}{d}. \tag{2.28}$$

Let me offer a few comments on this well-known formula. First, it is valid even if the gap is not quite planar – for example, if it gently curves on a scale much larger than d, but retains its thickness. Second, Eq. (28), which is valid only if $A \sim a^2$ is much larger than d^2, ignores the nonuniform electric fields spreading to distances $\sim d$ beyond the gap edges. Such *fringe fields* result in an additional *stray capacitance* $C' \sim \varepsilon_0 a << C \sim \varepsilon_0 a \times (a/d)$.[17] Finally, the same condition ($A >> d^2$) assures that C is much larger than the self-capacitance C_j of each conductor – see Eq. (18).

The opportunities opened by the last fact for electronic engineering and experimental physics practice are rather astonishing. For example, a very realistic 3-nm layer of high-quality aluminum oxide, which may provide nearly perfect electric insulation between two thin conducting films, with an area of 0.1 m^2 (a typical area of silicon wafers used in the semiconductor industry) provides $C \sim 1$ mF,[18] larger than the self-capacitance of the whole planet Earth!

[17] The exact value of C' depends on the shape of the conductors. In a rare case when it has been calculated analytically, two thin round concentric disks of radius R, $C' = \varepsilon_0 R\,[\ln(16\pi R/d) - 1]$.

[18] Just as in Sec. 1, for the estimate to be realistic, I took into account the additional factor κ (for aluminum oxide, close to 10) which should be included in the numerator of Eq. (28) to make it applicable to dielectrics – see Chapter 3 below.

In a plane capacitor with $d << a$, the electrostatic coupling of the two conductors is evidently very *strong*. As an opposite example of a *weakly* coupled system, let us consider two conducting spheres of the same radius R, separated by a much larger distance d (Fig. 4).

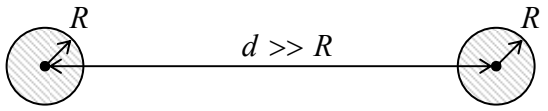


Fig. 2.4. A system of two far-separated, similar conducting spheres.

In this case, the diagonal components of the matrix (23) may be approximately found from Eq. (16), i.e. by neglecting the coupling altogether:

$$\wp_1 = \wp_2 \approx \frac{1}{4\pi\varepsilon_0 R}. \tag{2.29}$$

Now, if we had just one sphere (say, number 1), the electric potential at distance d from its center would be given by Eq. (16): $\phi = Q_1/4\pi\varepsilon_0 d$. If we move to this point a small ($R << d$) sphere without its own charge, we may expect that its potential should not be too far from this result, so $\phi_2 \approx Q_1/4\pi\varepsilon_0 d$. Comparing this expression with the second of Eqs. (19) (taken for $Q_2 = 0$), we get

$$\wp \approx \frac{1}{4\pi\varepsilon_0 d} << \wp_{1,2}. \tag{2.30}$$

From here and Eq. (26), the mutual capacitance

$$C \approx \frac{1}{\wp_1 + \wp_2} \approx 2\pi\varepsilon_0 R. \tag{2.31}$$

We see that (somewhat counter-intuitively), in this limit C does not depend substantially on the distance between the spheres, i.e. does *not* describe their electrostatic coupling. The off-diagonal coefficients of the *reciprocal* capacitance matrix (20) play this role much better – see Eq. (30).

Now let us consider the case when only one conductor of the two is charged, for example, $Q_1 \equiv Q$, while $Q_2 = 0$. Then Eqs. (19)-(20) yield

$$\phi_1 = \wp_1 Q_1. \tag{2.32}$$

Now, we may follow Eq. (13) and define $C_1 \equiv 1/\wp_1$ (and $C_2 \equiv 1/\wp_2$), just to see that such *partial capacitances* of the conductors of the system differ from its mutual capacitance C – cf. Eq. (26). For example, in the case shown in Fig. 4, $C_1 = C_2 \approx 4\pi\varepsilon_0 R \approx 2C$.

Finally, let us consider one more frequent case when one of the conductors carries a certain charge (say, $Q_1 = Q$), but the potential of its counterpart is sustained constant, say $\phi_2 = 0$.[19] (This condition is especially easy to implement if the second conductor is much larger than the first one. Indeed, as the estimate (18) shows, in this case, it would take a much larger charge Q_2 to make the potential ϕ_2 comparable with ϕ_1.) In this case the second of Eqs. (19), with the account of Eq. (20), yields $Q_2 = -(\wp/\wp_2)Q_1$. Plugging this relation into the first of those equations, we get

[19] In electrical engineering, such a constant-potential conductor is called the *ground*. This term stems from the fact that in many cases the electrostatic potential of the (weakly) conducting ground at the Earth's surface is virtually unaffected by laboratory-scale electric charges.

$$Q_1 = C_1^{\text{ef}}\phi_1, \qquad \text{with } C_1^{\text{ef}} \equiv \left(\mathcal{P}_1 - \frac{\mathcal{P}^2}{\mathcal{P}_2}\right)^{-1} \equiv \frac{\mathcal{P}_2}{\mathcal{P}_1\mathcal{P}_2 - \mathcal{P}^2}. \tag{2.33}$$

Thus, this *effective capacitance* of the first conductor is generally different from both its partial capacitance C_1 and the mutual capacitance C of the system, emphasizing again how accurate one should be using the term "capacitance" without a qualifier.

Note also that none of these capacitances is equal to any element of the matrix reciprocal to the matrix (23):

$$\begin{pmatrix} \mathcal{P}_1 & \mathcal{P} \\ \mathcal{P} & \mathcal{P}_2 \end{pmatrix}^{-1} = \frac{1}{\mathcal{P}^2 - \mathcal{P}_1\mathcal{P}_2}\begin{pmatrix} -\mathcal{P}_2 & \mathcal{P} \\ \mathcal{P} & -\mathcal{P}_1 \end{pmatrix}. \tag{2.34}$$

Because of this reason, this *physical capacitance matrix*, which expresses the vector of conductor charges via the vector of their potentials, is less convenient for most applications than the reciprocal capacitance matrix (23). The same conclusion is valid for multi-conductor systems, which are most conveniently characterized by an evident generalization of Eq. (19). Indeed, in this case, even the mutual capacitance between two selected conductors may depend on the electrostatic conditions of other components of the system.

Logically, at this point I would need to discuss the particular, but practically very important case when the regions where the electric field between each pair of conductors is most significant do not overlap – such as in the example shown in Fig. 5a. In this case, the system's properties may be discussed using the *equivalent-circuit* language, representing each such region as a *lumped* (localized) *capacitor*, with a certain mutual capacitance C, and the whole system as some connection of these capacitors by conducting "wires", whose length and geometry are not important – see Fig. 5b.

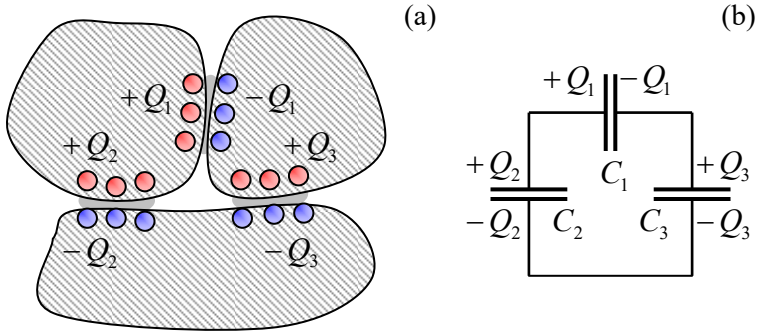


Fig. 2.5. (a) A simple system of conductors, with three well-localized regions of high electric field (and hence surface charge) concentration, and (b) its representation with an equivalent circuit of three lumped capacitors.

Since the analysis of such equivalent circuits is covered in typical introductory physics courses, I will save time by skipping their discussion. However, since such circuits are very frequently met in physical experiment and electrical engineering practice, I would urge the reader to self-test their understanding of this topic by solving a couple of problems offered at the end of this chapter,[20] and if their solution presents any difficulty, review the corresponding section in an undergraduate textbook.

[20] These problems have been selected to emphasize the fact that not every circuit may be reduced to the simplest connections of the component capacitors and/or their groups in parallel and/or in series.

2.3. The simplest boundary problems

In the general case when the electric field distribution in the free space between the conductors cannot be easily found from the Gauss law or a particular symmetry, the best approach is to try to solve the differential Laplace equation (1.42), with the boundary conditions (1b):

$$\nabla^2\phi = 0, \qquad \phi\Big|_{S_k} = \phi_k\,, \tag{2.35}$$

Typical boundary problem

where S_k is the surface of the k^{th} conductor of the system. After this *boundary problem* has been solved, i.e. the spatial distribution $\phi(\mathbf{r})$ has been found at all points outside the conductors, it is straightforward to use Eq. (3) to find the surface charge density, and finally the total charge

$$Q_k = \oint_{S_k} \sigma d^2 r \tag{2.36}$$

of each conductor, and hence any component of the reciprocal capacitance matrix. As an illustration, let us implement this program for three very simple problems.

(i) Plane capacitor (Fig. 3). In this case, the easiest way to solve the Laplace equation is to use the linear (Cartesian) coordinates with one axis (say, z) normal to the conductor surfaces – see Fig. 6.

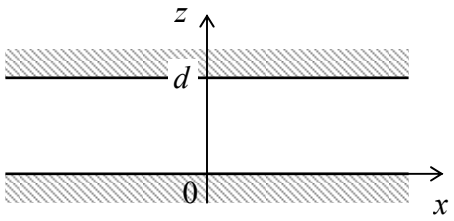


Fig. 2.6. The plane capacitor as the system for the simplest illustration of the boundary problem (35) and its solution.

In these coordinates, the Laplace operator is just the sum of three second derivatives.[21] It is evident that due to the problem's translational symmetry within the [x, y] plane, deep inside the gap (i.e. at any lateral distance from the edges much larger than d) the electrostatic potential may only depend on the coordinate normal to the gap surfaces: $\phi(\mathbf{r}) = \phi(z)$. For such a function, the derivatives over x and y vanish, and the boundary problem (35) is reduced to a very simple ordinary differential equation

$$\frac{d^2\phi}{dz^2}(z) = 0, \tag{2.37}$$

with boundary conditions

$$\phi(0) = 0, \quad \phi(d) = V. \tag{2.38}$$

(For the sake of notation simplicity, I have used the discretion of adding a constant to the potential, to make one of the potentials vanish, and also the definition (25) of the voltage V.) The general solution of Eq. (37) is a linear function: $\phi(z) = c_1 z + c_2$, whose constant coefficients $c_{1,2}$ may be readily found from the boundary conditions (38). The final solution is

$$\phi = V\frac{z}{d}. \tag{2.39}$$

[21] See, e.g. MA Eq. (9.1).

From here the only nonzero component of the electric field is

$$E_z = -\frac{d\phi}{dz} = -\frac{V}{d}, \tag{2.40}$$

and the surface charge of the capacitor plates is

$$\sigma = \varepsilon_0 E_n = \mp\varepsilon_0 E_z = \pm\varepsilon_0 \frac{V}{d}, \tag{2.41}$$

where the upper and lower signs correspond to the upper and lower plates, respectively. Since σ does not depend on x and y, we can get the full charges $Q_1 = -Q_2 \equiv Q$ of the surfaces by its multiplication by the gap area A, giving us again the already obtained result (28) for the mutual capacitance $C \equiv Q/V$. I believe that this calculation, though very easy, may serve as a good illustration of the boundary problem solution approach, which will be used below for more complex cases.

(ii) Coaxial-cable capacitor. *Coaxial cable* is a system of two round cylindrical, coaxial conductors, with the cross-section shown in Fig. 7.

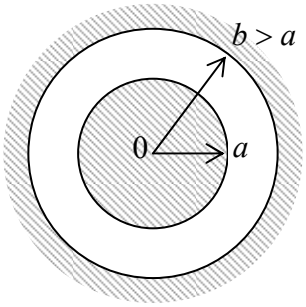


Fig. 2.7. The cross-section of a coaxial cable.

Evidently, in this case, the cylindrical coordinates $\{\rho, \varphi, z\}$, with the z-axis coinciding with the common axis of the cylinders, are most convenient.[22] Due to the axial symmetry of the problem, in these coordinates $\mathbf{E}(\mathbf{r}) = \mathbf{n}_\rho E(\rho)$, $\phi(\mathbf{r}) = \phi(\rho)$, so in the general expression for the Laplace operator[23] we may take $\partial/\partial\varphi = \partial/\partial z = 0$. As a result, only the radial term of the operator survives, and the boundary problem (35) takes the form

$$\frac{1}{\rho}\frac{d}{d\rho}\left(\rho\frac{d\phi}{d\rho}\right) = 0, \qquad \phi(a) = V, \qquad \phi(b) = 0. \tag{2.42}$$

The sequential double integration of this ordinary linear differential equation is elementary (and similar to that of the Poisson equation in spherical coordinates, carried out in Sec. 1.3), giving

$$\rho\frac{d\phi}{d\rho} = c_1, \qquad \phi(\rho) = c_1\int_a^\rho \frac{d\rho''}{\rho''} + c_2 = c_1 \ln\frac{\rho}{a} + c_2. \tag{2.43}$$

The constants $c_{1,2}$ may be found using boundary conditions (42):

[22] I am sorry for using, for the 2D radius, the same letter ρ as for the volumic density of charge. (Both notations are too common to refuse.) I do not believe this may lead to confusion, because the letter will not be used in two different meanings during any particular discussion.

[23] See, e.g., MA Eq. (10.3).

$$c_2 = V, \qquad c_1 \ln\frac{b}{a} + c_2 = 0, \tag{2.44}$$

giving $c_1 = -V/\ln(b/a)$, so Eq. (43) takes the following form:

$$\phi(\rho) = V\left[1 - \frac{\ln(\rho/a)}{\ln(b/a)}\right]. \tag{2.45}$$

Next, for our axial symmetry, the general expression for the gradient of a scalar function is reduced to its radial derivative, so

$$E(\rho) \equiv -\frac{d\phi(\rho)}{d\rho} = \frac{V}{\rho \ln(b/a)}. \tag{2.46}$$

This expression, plugged into Eq. (2), allows us to find the density of the conductors' surface charge. For example, for the inner electrode

$$\sigma_a = \varepsilon_0 E(a) = \frac{\varepsilon_0 V}{a \ln(b/a)}, \tag{2.47}$$

so its full charge (per unit length of the system) is

$$\frac{Q}{l} = 2\pi a \sigma_a = \frac{2\pi\varepsilon_0 V}{\ln(b/a)}. \tag{2.48}$$

(It is straightforward to check that the charge of the outer electrode is equal and opposite.) Hence, by the definition of the mutual capacitance, its value per unit length is

$$\boxed{\frac{C}{l} \equiv \frac{Q}{lV} = \frac{2\pi\varepsilon_0}{\ln(b/a)}}. \tag{2.49}$$

C: Coaxial cable

This expression shows that the total capacitance C is proportional to the systems length l (if $l >> a, b$), while being only logarithmically dependent on is the dimensions of its cross-section. Since the logarithm of a large argument is an extremely slow function (sometimes called a *quasi-constant*), if the external conductor is made very large ($b >> a$), the capacitance diverges, but very weakly. Such *logarithmic divergence* may be cut by any minuscule additional effect, for example by the finite length l of the system. This fact yields the following very useful estimate of the self-capacitance of a *single* round wire of radius a:

$$C \approx \frac{2\pi\varepsilon_0 l}{\ln(l/a)}, \qquad \text{for } l >> a. \tag{2.50}$$

On the other hand, if the gap d between the conductors is very narrow: $d \equiv b - a << a$, then $\ln(b/a) \equiv \ln(1 + d/a)$ may be approximated as d/a, and Eq. (49) is reduced to $C \approx 2\pi\varepsilon_0 al/d$, i.e. to Eq. (28) for the plane capacitor, of the appropriate area $A = 2\pi al$.

(iii) Spherical capacitor. This is a system of two conductors, with a *central* cross-section similar to that of the coaxial cable (Fig. 7), but now with spherical rather than axial symmetry. This symmetry implies that we may be better off using spherical coordinates, so the potential ϕ depends only on one of them: the distance r from the common center of the conductors: $\phi(\mathbf{r}) = \phi(r)$. As we already know from Sec. 1.3, in this case the general expression for the Laplace operator is reduced to its first (radial) term,

so the Laplace equation takes the simple form (1.47). Moreover, we have already found the general solution of this equation – see Eq. (1.50):

$$\phi(r) = \frac{c_1}{r} + c_2, \tag{2.51}$$

Now acting exactly as above, i.e. determining the (only essential) constant c_1 from the boundary condition $\phi(a) - \phi(b) = V$, we get

$$c_1 = V\left(\frac{1}{a} - \frac{1}{b}\right)^{-1}, \qquad \text{so that} \quad \phi(r) = \frac{V}{r}\left(\frac{1}{a} - \frac{1}{b}\right)^{-1} + c_2. \tag{2.52}$$

Next, we can use the spherical symmetry to find the electric field, $\mathbf{E}(\mathbf{r}) = \mathbf{n}_r E(r)$, with

$$E(r) = -\frac{d\phi}{dr} = \frac{V}{r^2}\left(\frac{1}{a} - \frac{1}{b}\right)^{-1}, \tag{2.53}$$

and hence its values on conductors' surfaces, and then the surface charge density σ from Eq. (3). For example, for the inner conductor's surface,

$$\sigma_a = \varepsilon_0 E(a) = \varepsilon_0 \frac{V}{a^2}\left(\frac{1}{a} - \frac{1}{b}\right)^{-1}, \tag{2.54}$$

so, finally, for the full charge of that conductor, we get the following result:

$$Q = 4\pi a^2 \sigma = 4\pi\varepsilon_0\left(\frac{1}{a} - \frac{1}{b}\right)^{-1} V. \tag{2.55}$$

(Again, the charge of the outer conductor is equal and opposite.) Now we can use the definition (26) of the mutual capacitance to get the final result:

C: Spherical capacitor

$$C \equiv \frac{Q}{V} = 4\pi\varepsilon_0\left(\frac{1}{a} - \frac{1}{b}\right)^{-1} \equiv 4\pi\varepsilon_0 \frac{ab}{b-a}. \tag{2.56}$$

For $b >> a$, it coincides with Eq. (17) for the self-capacitance of the inner conductor. On the other hand, if the gap d between two conductors is narrow, $d \equiv b - a << a$, then

$$C = 4\pi\varepsilon_0 \frac{a(a+d)}{d} \approx 4\pi\varepsilon_0 \frac{a^2}{d}, \tag{2.57}$$

i.e. the capacitance approaches that of the planar capacitor of the area $A = 4\pi a^2$ – as it should.

All this seems (and indeed is) very straightforward, but let us contemplate what was the reason for such easy successes. In each of the cases (i)-(iii) we have managed to find such coordinates that both the Laplace equation and the boundary conditions involved only one of them. The necessary condition for the former fact is for the coordinates to be *orthogonal*. This means that the three vector components of the local differential $d\mathbf{r}$, due to small variations of the new coordinates (say, dr, $d\theta$, and $d\varphi$ for the spherical coordinates), are mutually perpendicular.

2.4. Using other orthogonal coordinates

The cylindrical and spherical coordinates used above are only the simplest examples of the *curvilinear orthogonal* (or just "orthogonal") coordinates, and that approach may be extended to other

coordinate systems of this type. As an example, let us calculate the self-capacitance of a thin, round conducting disk. The cylindrical or spherical coordinates would not give much help here, because while they have the appropriate axial symmetry, they would make the boundary condition on the disk too complicated: involving two coordinates, either ρ and z, or r and θ. Help comes from noting that the flat disk, i.e. the area with $z = 0$, $r < R$, may be viewed as the limiting case of an *axially-symmetric ellipsoid* (or "degenerate ellipsoid", or "ellipsoid of rotation", or "spheroid") – the surface formed by rotation of the usual ellipse about one of its major axes – which would be also the symmetry axis of the disk – in Fig. 8, the z-axis.

–

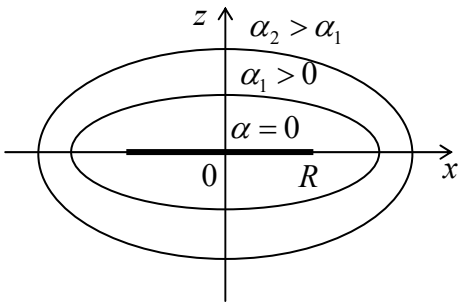


Fig. 2.8. Solving the disk's capacitance problem. (The cross-section of the system by the vertical plane $y = 0$.)

Analytically, this ellipsoid may be described by the following equation:

$$\frac{x^2 + y^2}{a^2} + \frac{z^2}{b^2} = 1, \tag{2.58}$$

where a and b are the so-called *major semi-axes*, whose ratio determines the ellipse's *eccentricity* – the degree of its "squeezing". For our problem, we will only need *oblate* ellipsoids with $a \geq b$; according to Eq. (58), they may be represented as surfaces of constant α in the *oblate spheroidal* (also called "degenerate ellipsoidal") *coordinates* $\{\alpha, \beta, \varphi\}$ that are related to the Cartesian coordinates as follows:[24]

$$\begin{aligned} x &= R\cosh\alpha\sin\beta\cos\varphi, & 0 \leq \alpha < \infty, \\ y &= R\cosh\alpha\sin\beta\sin\varphi, \quad \text{with} & 0 \leq \beta \leq \pi, \\ z &= R\sinh\alpha\cos\beta, & 0 \leq \varphi \leq 2\pi. \end{aligned} \tag{2.59}$$

Such spheroidal coordinates are an evident generalization of the spherical coordinates, which correspond to the limit $\alpha >> 1$ (i.e. $r >> R$). In the opposite limit, the surface of constant $\alpha = 0$ describes our thin disk of radius R, with the coordinate β describing the distance $\rho \equiv (x^2 + y^2)^{1/2} = R\sin\beta$ of its point from the z-axis. It is almost evident (and easy to prove) that the curvilinear coordinates (59) are also orthogonal; the Laplace operator in them is:

$$\nabla^2 = \frac{1}{R^2(\cosh^2\alpha - \sin^2\beta)} \times \left[\begin{aligned} &\frac{1}{\cosh\alpha}\frac{\partial}{\partial\alpha}\left(\cosh\alpha\frac{\partial}{\partial\alpha}\right) \\ &+ \frac{1}{\sin\beta}\frac{\partial}{\partial\beta}\left(\sin\beta\frac{\partial}{\partial\beta}\right) + \left(\frac{1}{\sin^2\beta} - \frac{1}{\cosh^2\alpha}\right)\frac{\partial^2}{\partial\varphi^2} \end{aligned} \right]. \tag{2.60}$$

Though this expression may look a bit intimidating, let us notice that since in our current problem, the boundary conditions depend only on α:[25]

[24] For solution of some problems, it is convenient to use Eqs. (59) with $-\infty < \alpha < +\infty$ and $0 \leq \beta \leq \pi/2$.

[25] I have called the disk's potential V, to distinguish it from the potential ϕ at an arbitrary point of space.

$$\phi\big|_{\alpha=0} = V, \qquad \phi\big|_{\alpha=\infty} = 0, \tag{2.61}$$

there is every reason to assume that the electrostatic potential in all space is a function of α alone; in other words, that all ellipsoids α = const are the equipotential surfaces. Indeed, acting on such a function $\phi(\alpha)$ by the Laplace operator (60), we see that the two last terms in the square brackets vanish, and the Laplace equation (35) is reduced to a simple ordinary differential equation

$$\frac{d}{d\alpha}\left[\cosh\alpha \frac{d\phi}{d\alpha}\right] = 0. \tag{2.62}$$

Integrating it twice, just as we did in the three previous problems, we get

$$\phi(\alpha) = c_1 \int \frac{d\alpha}{\cosh\alpha}. \tag{2.63}$$

This integral may be readily worked out using the substitution $\xi \equiv \sinh\alpha$ (which gives $d\xi \equiv \cosh\alpha\, d\alpha$, i.e. $d\alpha = d\xi/\cos\alpha$, and $\cosh^2\alpha = 1 + \sinh^2\alpha \equiv 1 + \xi^2$):

$$\phi(\alpha) = c_1 \int_0^{\sinh\alpha} \frac{d\xi}{1+\xi^2} + c_2 = c_1 \tan^{-1}(\sinh\alpha) + c_2. \tag{2.64}$$

The integration constants $c_{1,2}$ may be simply found from the boundary conditions (61), and we arrive at the following final expression for the electrostatic potential:

$$\phi(\alpha) = V\left[1 - \frac{2}{\pi}\tan^{-1}(\sinh\alpha)\right] \equiv \frac{2V}{\pi}\tan^{-1}\left(\frac{1}{\sinh\alpha}\right). \tag{2.65}$$

This solution satisfies both the Laplace equation and the boundary conditions. Mathematicians tell us that the solution of any boundary problem of the type (35) is *unique*, so we do not need to look any further.

Now we may use Eq. (3) to find the surface density of electric charge, but in the case of a thin disk, it is more natural to add up such densities on its top and bottom surfaces at the same distance $\rho = (x^2 + y^2)^{1/2}$ from the disk's center. The densities are evidently equal, due to the problem symmetry about the plane $z = 0$, so the total density is $\sigma = 2\varepsilon_0 E_n|_{z=+0}$. According to Eq. (65), and the last of Eqs. (59), the electric field on the upper surface is

$$E_n\big|_{z=+0} = -\frac{\partial\phi}{\partial z}\Big|_{z=+0} = -\frac{\partial\phi(\alpha)}{\partial(R\sinh\alpha\cos\beta)}\Big|_{\alpha=0} = \frac{2}{\pi}V\frac{1}{R\cos\beta} = \frac{2}{\pi}V\frac{1}{(R^2-\rho^2)^{1/2}}, \tag{2.66}$$

and we see that the charge is distributed over the disk very nonuniformly:

$$\sigma = \frac{4}{\pi}\varepsilon_0 V \frac{1}{(R^2-\rho^2)^{1/2}}, \tag{2.67}$$

with a singularity at the disk edge. Below we will see that such singularities are very typical for sharp edges of conductors. Fortunately, in our current case the divergence is integrable, giving a finite disk charge:

$$Q = \int_{\substack{\text{disk}\\ \text{surface}}} \sigma d^2\rho = \int_0^R \sigma(\rho) 2\pi\rho d\rho = \frac{4}{\pi}\varepsilon_0 V \int_0^R \frac{2\pi\rho d\rho}{(R^2-\rho^2)^{1/2}} = 4\varepsilon_0 VR \int_0^1 \frac{d\xi}{(1-\xi)^{1/2}} = 8\varepsilon_0 RV. \qquad (2.68)$$

Thus, for the disk's self-capacitance we get a very simple result,

$$C = 8\varepsilon_0 R \equiv \frac{2}{\pi} 4\pi\varepsilon_0 R, \qquad (2.69)$$

a factor of $\pi/2 \approx 1.57$ lower than that for the conducting sphere of the same radius, but still complying with the general estimate (18).

Can we always find such a "good" system of orthogonal coordinates? Unfortunately, the answer is *no*, even for highly symmetric geometries. This is why the practical value of this approach is limited, and other, more general methods of boundary problem solution are clearly needed. Before proceeding to their discussion, however, let me note that in the case of 2D problems (i.e. cylindrical geometries[26]), the orthogonal coordinate method gets much help from the following *conformal mapping* approach.

Let us consider a pair of Cartesian coordinates $\{x, y\}$ of the cylinder's cross-section plane as a complex variable $\boldsymbol{z} \equiv x + iy$,[27] where i is the imaginary unit ($i^2 = -1$), and let $\boldsymbol{w}(\boldsymbol{z}) = u + iv$ be an *analytic complex function* of $\boldsymbol{z}$.[28] For our current purposes, the most important property of an analytic function is that its real and imaginary parts obey the following *Cauchy-Riemann relations*:[29]

$$\frac{\partial u}{\partial x} = \frac{\partial v}{\partial y}, \qquad \frac{\partial v}{\partial x} = -\frac{\partial u}{\partial y}. \qquad (2.70)$$

For example, for the function

$$\boldsymbol{w} = \boldsymbol{z}^2 \equiv (x+iy)^2 \equiv (x^2 - y^2) + 2ixy, \qquad (2.71)$$

whose real and imaginary parts are

$$u \equiv \mathrm{Re}\,\boldsymbol{w} = x^2 - y^2, \qquad v \equiv \mathrm{Im}\,\boldsymbol{w} = 2xy, \qquad (2.72)$$

we immediately see that $\partial u/\partial x = 2x = \partial v/\partial y$, and $\partial v/\partial x = 2y = -\partial u/\partial y$, in accordance with Eq. (70).

Let us differentiate the first of Eqs. (70) over x again, then change the order of differentiation, and after that use the latter of those equations:

$$\frac{\partial^2 u}{\partial x^2} = \frac{\partial}{\partial x}\frac{\partial u}{\partial x} = \frac{\partial}{\partial x}\frac{\partial v}{\partial y} = \frac{\partial}{\partial y}\frac{\partial v}{\partial x} = -\frac{\partial}{\partial y}\frac{\partial u}{\partial y} = -\frac{\partial^2 u}{\partial y^2}, \qquad (2.73)$$

---

[26] Let me remind the reader that the term *cylindrical* describes any surface formed by a translation, along a straight line, of an *arbitrary* curve, and hence more general than the usual circular cylinder. (In this terminology, for example, a prism is also a cylinder of a particular type, formed by a translation of a polygon.)

[27] The complex variable $\boldsymbol{z}$ should not be confused with the (real) 3rd spatial coordinate z! We are considering 2D problems now, with the potential independent of z.

[28] An analytic (or "holomorphic") function may be defined as one that may be expanded into the Taylor series in its complex argument, i.e. is infinitely differentiable in the given point. (Almost all "regular" functions, such as $\boldsymbol{z}^n$, $\boldsymbol{z}^{1/n}$, $\exp \boldsymbol{z}$, $\ln \boldsymbol{z}$, etc., and their linear combinations are analytic at all $\boldsymbol{z}$, maybe besides certain special points.) If the reader needs to brush up on their background on this subject, I can recommend a popular textbook by M. Spiegel *et al*., *Complex Variables*, 2nd ed., McGraw-Hill, 2009.

[29] These relations may be used, in particular, to prove the Cauchy integral formula – see, e.g., MA Eq. (15.1).

and similarly for v. This means that the sum of second-order partial derivatives of each of the real functions $u(x, y)$ and $v(x, y)$ is zero, i.e. that both functions obey the 2D Laplace equation. This mathematical fact opens a nice way of solving problems of electrostatics for (relatively simple) 2D geometries. Imagine that for a particular boundary problem we have found a function $\boldsymbol{w}(\boldsymbol{z})$ for that either $u(x, y)$ or $v(x, y)$ is constant on all electrode surfaces. Then all lines of constant u (or v) represent equipotential surfaces, i.e. the problem of the potential distribution has been essentially solved.

As a simple example, let us consider a problem important for practice: the *quadrupole electrostatic lens* – a system of four cylindrical electrodes with hyperbolic cross-sections, whose boundaries are described by the following relations:

$$x^2 - y^2 = \begin{cases} +a^2, & \text{for the left and right electrodes,} \\ -a^2, & \text{for the top and bottom electrodes,} \end{cases} \tag{2.74}$$

voltage-biased as shown in Fig. 9a.

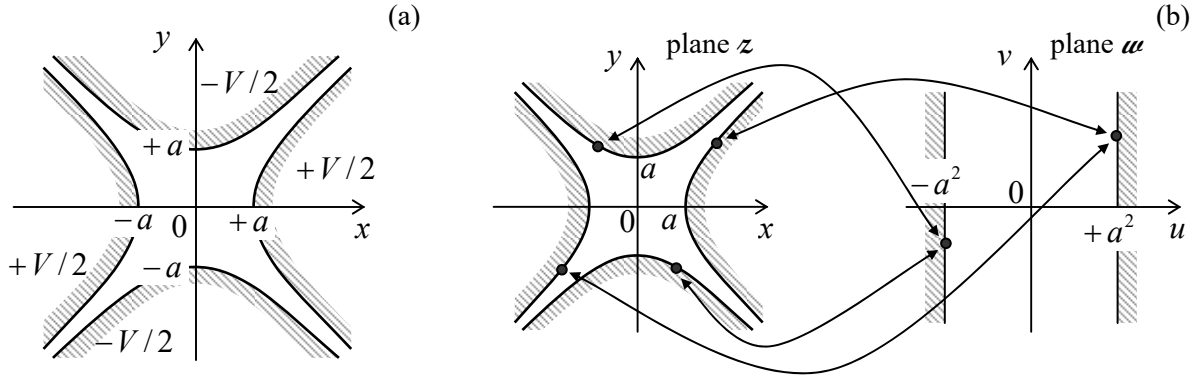


Fig. 2.9. (a) The quadrupole electrostatic lens' cross-section and (b) its conformal mapping.

Comparing these relations with Eqs. (72), we see that each electrode surface corresponds to a constant value of the real part $u(x, y)$ of the function given by Eq. (71): $u = \pm a^2$. Moreover, the potentials of both surfaces with $u = +a^2$ are equal to $+V/2$, while those with $u = -a^2$ are equal to $-V/2$. Hence we may conjecture that the electrostatic potential at each point is a function of u alone; moreover, a simple linear function,

$$\phi = c_1 u + c_2 \equiv c_1 (x^2 - y^2) + c_2 , \tag{2.75}$$

is a valid (and hence the unique) solution of our boundary problem. Indeed, it does satisfy the Laplace equation, while the constants $c_{1,2}$ may be readily selected in a way to satisfy all the boundary conditions shown in Fig. 9a:

$$\phi = \frac{V}{2} \frac{x^2 - y^2}{a^2} . \tag{2.76}$$

so the boundary problem has been solved.

According to Eq. (76), all equipotential surfaces are hyperbolic cylinders, similar to those of the electrode surfaces. What remains is to find the electric field at an arbitrary point inside the system:

$$E_x = -\frac{\partial \phi}{\partial x} = -V \frac{x}{a^2}, \quad E_y = -\frac{\partial \phi}{\partial y} = V \frac{y}{a^2}. \tag{2.77}$$

These formulas show, in particular, that if charged particles (e.g., electrons in an electron-optics system) are launched to fly ballistically through such a lens, along the z-axis, they experience a force pushing them toward the symmetry axis and proportional to the particle's deviation from the axis (and thus equivalent in action to an optical lens with a positive refraction power) in one direction, and a force pushing them out (negative refractive power) in the perpendicular direction. One can show that letting the particles fly through several such lenses, with alternating voltage polarities, in series, enables beam focusing.[30]

Hence, we have reduced the 2D Laplace boundary problem to that of finding the proper analytic function $\boldsymbol{w}(\boldsymbol{z})$. This task may be also understood as that of finding a *conformal map*, i.e. a correspondence between components of any point pair, $\{x, y\}$ and $\{u, v\}$, residing, respectively, on the initial Cartesian plane $\boldsymbol{z}$ and the plane $\boldsymbol{w}$ of the new variables. For example, Eq. (71) maps the real electrode configuration onto a plane capacitor of an infinite area (Fig. 9b), and the simplicity of Eq. (75) is due to the fact that for the latter system the equipotential surfaces are just parallel planes u = const.

For more complex geometries, the suitable analytic function $\boldsymbol{w}(\boldsymbol{z})$ may be hard to find. However, for conductors with piece-linear cross-section boundaries, substantial help may be obtained from the following *Schwarz-Christoffel integral*

$$\boldsymbol{w}(\boldsymbol{z}) = \text{const} \times \int \frac{d\boldsymbol{z}}{(\boldsymbol{z}-x_1)^{k_1}(\boldsymbol{z}-x_2)^{k_2}...(\boldsymbol{z}-x_{N-1})^{k_{N-1}}}. \qquad (2.78)$$

that provides a conformal mapping of the interior of an *arbitrary* N-sided polygon onto the plane $\boldsymbol{w} = u + iv$, onto the upper half ($y > 0$) of the plane $\boldsymbol{z} = x + iy$. In Eq. (78), x_j ($j = 1, 2, N - 1$) are the points of the $y = 0$ axis (i.e., of the boundary of the mapped region on plane $\boldsymbol{z}$) to which the corresponding polygon vertices are mapped, while k_j are the exterior angles at the polygon vertices, measured in the units of π, with $-1 \le k_j \le +1$ – see Fig. 10.[31] Of the points x_j, two may be selected arbitrarily (because their effects may be compensated by the multiplicative constant in Eq. (78), and the additive constant of integration), while all the others have to be adjusted to provide the correct mapping.

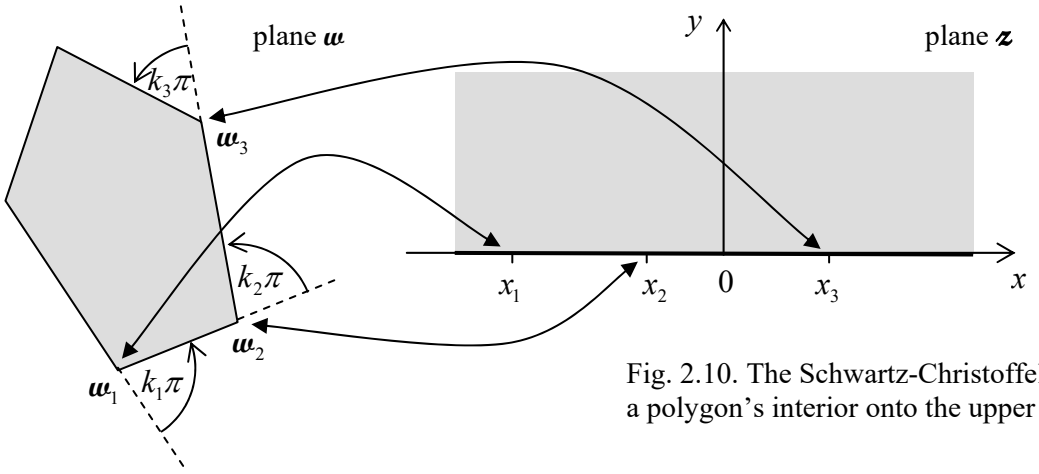


Fig. 2.10. The Schwartz-Christoffel mapping of a polygon's interior onto the upper half-plane.

[30] See, e.g., textbook by P. Grivet, *Electron Optics*, 2nd ed., Pergamon, 1972.

[31] The integral (78) includes only $(N - 1)$ rather than N poles because a polygon's shape is fully determined by $(N - 1)$ positions $\boldsymbol{w}_j$ of its vertices and $(N - 1)$ angles πk_j. In particular, since the algebraic sum of all external angles of a polygon equals 2π, the last angle parameter $k_j = k_N$ is uniquely determined by the set of the previous ones.

In the general case, the complex integral (78) may be hard to tackle. However, in some important cases, in particular those with right angles ($k_j = \pm ½$) and/or with some points $\boldsymbol{w}_j$ at infinity, the integrals may be readily worked out, giving explicit analytical expressions for the mapping functions $\boldsymbol{w}(\boldsymbol{z})$. For example, let us consider a semi-infinite strip defined by restrictions $-1 \le u \le +1$ and $0 \le v$, on the $\boldsymbol{w}$-plane – see the left panel of Fig. 11.

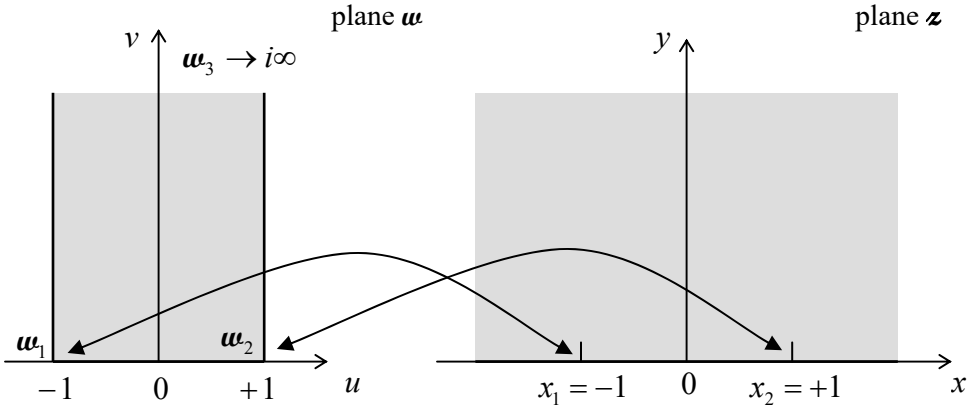


Fig. 2.11. A semi-infinite strip mapped onto the upper half-plane.

The strip may be considered as a triangle, with one vertex at the infinitely distant vertical point $\boldsymbol{w}_3 = 0 + i\infty$. Let us map the polygon onto the upper half of plane $\boldsymbol{z}$, shown on the right panel of Fig. 11, with the vertex $\boldsymbol{w}_1 = -1 + i\,0$ mapped onto the point $\boldsymbol{z}_1 = -1 + i\,0$, and the vertex $\boldsymbol{w}_2 = +1 + i\,0$ mapped onto the point $\boldsymbol{z}_2 = +1 + i\,0$. Since the external angles at these vertices are equal to $+\pi/2$, and hence $k_1 = k_2 = +½$, Eq. (78) is reduced to

$$\boldsymbol{w}(\boldsymbol{z}) = \text{const} \times \int \frac{d\boldsymbol{z}}{(\boldsymbol{z}+1)^{1/2}(\boldsymbol{z}-1)^{1/2}} \equiv \text{const} \times \int \frac{d\boldsymbol{z}}{(\boldsymbol{z}^2-1)^{1/2}} \equiv \text{const} \times i \int \frac{d\boldsymbol{z}}{(1-\boldsymbol{z}^2)^{1/2}}. \qquad (2.79)$$

This complex integral may be worked out, just as for real $\boldsymbol{z}$, with the substitution $\boldsymbol{z} = \sin\xi$, giving

$$w(\boldsymbol{z}) = \text{const}' \times \int^{\sin^{-1}\boldsymbol{z}} d\xi = c_1 \sin^{-1}\boldsymbol{z} + c_2. \qquad (2.80)$$

Determining the constants $c_{1,2}$ from the required mapping, i.e. from the conditions $\boldsymbol{w}(-1 + i\,0) = -1 + i\,0$ and $\boldsymbol{w}(+1+ i\,0) = +1 + i\,0$ (see the arrows in Fig. 11), we finally get[32]

$$\boldsymbol{w}(\boldsymbol{z}) = \frac{2}{\pi}\sin^{-1}\boldsymbol{z}, \quad \text{i.e. } \boldsymbol{z} = \sin\frac{\pi\boldsymbol{w}}{2}. \qquad (2.81a)$$

Using the well-known expression for the sine of a complex argument,[33] we may rewrite this elegant result in either of the following two forms for the real and imaginary components of $\boldsymbol{z}$ and $\boldsymbol{w}$:

$$u = \frac{2}{\pi}\sin^{-1}\frac{2x}{\left[(x+1)^2+y^2\right]^{1/2}+\left[(x-1)^2+y^2\right]^{1/2}}, \quad v = \frac{2}{\pi}\cosh^{-1}\frac{\left[(x+1)^2+y^2\right]^{1/2}+\left[(x-1)^2+y^2\right]^{1/2}}{2},$$

[32] Note that this function differs only by a linear transformation of variables from the function $\boldsymbol{z} = c\cosh\boldsymbol{w}$, which is the canonical form of the definition of the so-called *elliptic* (not ellipsoidal!) orthogonal coordinates.

[33] See, e.g., MA Eq. (3.5).

$$x = \sin\frac{\pi u}{2}\cosh\frac{\pi v}{2}, \qquad y = \cos\frac{\pi u}{2}\sinh\frac{\pi v}{2}. \tag{2.81b}$$

It is amazing how perfectly the last formula manages to keep $y \equiv 0$ at the different borders of our $\boldsymbol{w}$-region (Fig. 11): at its side borders ($u = \pm 1$, $0 \le v < \infty$), this is performed by the first multiplier, while at the bottom border ($-1 \le u \le +1$, $v = 0$), the equality is enforced by the second multiplier.

This mapping may be used to solve several electrostatics problems with the geometry shown in Fig. 11a; probably the most surprising of them is the following one. A straight gap of width $2t$ is cut in a very thin conducting plane, and voltage V is applied between the resulting half-planes – see the bold straight lines in Fig. 12.

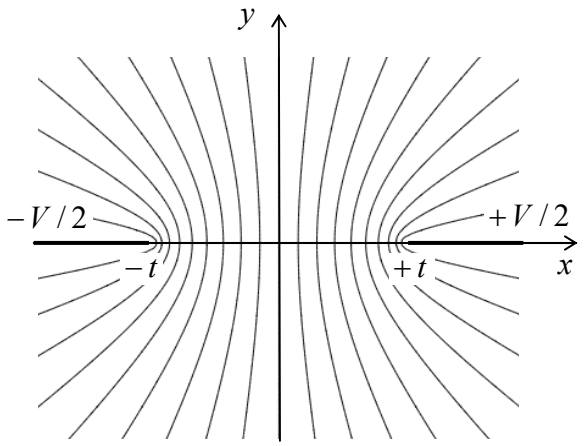


Fig. 2.12. The equipotential surfaces of the electric field between two thin conducting semi-planes (or rather their cross-sections by the plane z = const).

Selecting a Cartesian coordinate system with the z-axis directed along the cut, the y-axis normal to the plane, and the origin in the middle of the cut (Fig. 12), we can write the boundary conditions of this Laplace problem as

$$\phi = \begin{cases} +V/2, & \text{for } x > +t,\ y = 0, \\ -V/2, & \text{for } x < -t,\ y = 0. \end{cases} \tag{2.82}$$

(Due to the problem's symmetry, we may expect that in the middle of the gap, i.e. at $-t < x < +t$ and $y = 0$, the electric field is parallel to the plane and hence $\partial\phi/\partial y = 0$.) The comparison of Figs. 11 and 12 shows that if we normalize our coordinates $\{x, y\}$ to t, Eqs. (81) provide the conformal mapping of our system onto the plane $\boldsymbol{z}$ to a plane capacitor on the plane $\boldsymbol{w}$, with the voltage V between two conducting planes located at $u = \pm 1$. Since we already know that in that case $\phi = (V/2)u$, we may immediately use the first of Eqs. (81b) to write the final solution of the problem:[34]

$$\phi = \frac{V}{2}u = \frac{V}{\pi}\sin^{-1}\frac{2x}{\left[(x+t)^2 + y^2\right]^{1/2} + \left[(x-t)^2 + y^2\right]^{1/2}}. \tag{2.83}$$

The thin lines in Fig. 12 show the corresponding equipotential surfaces;[35] it is evident that the electric field concentrates at the gap edges, just as it did at the edge of the thin disk (Fig. 8). Let me

[34] This result may be also obtained by the Green's function method, to be discussed in Sec. 10 below.

[35] Another graphical representation of the electric field distribution, by *field lines*, is less convenient. (It is more useful for the magnetic field, which may be represented by a scalar potential only in particular cases, so there is no surprise that the field lines were introduced only by Michael Faraday in the 1830s.) As a reminder, the field

leave the remaining calculation of the surface charge distribution and the mutual capacitance between the half-planes (per unit length of the system in the z-direction) for the reader's exercise.

2.5. Variable separation – Cartesian coordinates

The general approach of the methods discussed in the last two sections was to satisfy the Laplace equation by a function of a single variable that also satisfies the boundary conditions. Unfortunately, in many cases this cannot be done – at least, using reasonably simple functions. In this case, a very powerful method called the *variable separation*,[36] may work, typically producing "semi-analytical" results in the form of series (infinite sums) of either elementary or well-studied special functions. Its main idea is to look for the solution of the boundary problem (35) as the sum of partial solutions,

$$\phi = \sum_k c_k \phi_k \ , \tag{2.84}$$

where each function ϕ_k satisfies the Laplace equation, and then select the set of coefficients c_k to satisfy the boundary conditions. More specifically, in the variable separation method, the partial solutions ϕ_k are looked for in the form of a product of functions, each depending on just one spatial coordinate.

Let us discuss this approach on the classical example of a rectangular box with conducting walls (Fig. 13), with the same potential (that I will take for zero) at all its sidewalls and the lower lid, but a different potential V at the top lid ($z = c$). Moreover, to demonstrate the power of the variable separation method, let us carry out all the calculations for a more general case when the top lid's potential is an arbitrary 2D function $V(x, y)$.[37]

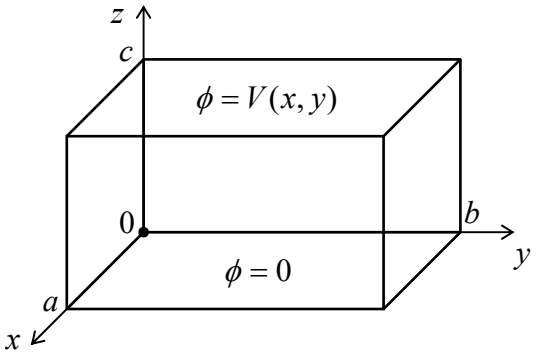


Fig. 2.13. The standard playground for the variable separation discussion: a rectangular box with five conducting, grounded walls and a fixed potential distribution $V(x, y)$ on the top lid.

For this geometry, it is natural to use the Cartesian coordinates $\{x, y, z\}$, representing each of the partial solutions in Eq. (84) as the following product

---

line is the curve to which the field vectors are tangential at each point. Hence the electric field lines are always normal to the equipotential surfaces, so it is always straightforward to sketch them, if desirable, from the equipotential surface pattern – like the one shown in Fig. 12.

[36] This method was already discussed in CM Sec. 6.5 and then used also in Secs. 6.6 and 8.4 of that course. However, it is so important that I need to repeat its discussion in this part of my series, for the benefit of the readers who have skipped the Classical Mechanics course for whatever reason.

[37] Such voltage distributions may be implemented in practice, for example, using the so-called *mosaic electrodes* consisting of many electrically-insulated and individually-biased panels.

$$\phi_k = X(x)Y(y)Z(z). \tag{2.85}$$

Plugging it into the Laplace equation expressed in the Cartesian coordinates,

$$\frac{\partial^2\phi_k}{\partial x^2} + \frac{\partial^2\phi_k}{\partial y^2} + \frac{\partial^2\phi_k}{\partial z^2} = 0, \tag{2.86}$$

and dividing the result by *XYZ*, we get

$$\frac{1}{X}\frac{d^2X}{dx^2} + \frac{1}{Y}\frac{d^2Y}{dy^2} + \frac{1}{Z}\frac{d^2Z}{dz^2} = 0. \tag{2.87}$$

Here comes the punch line of the variable separation method: since the first term of this sum may depend only on *x*, the second one only of *y*, etc., Eq. (87) may be satisfied everywhere in the volume only if each of these terms equals a constant. In a minute we will see that for our current problem (Fig. 13), these constant *x*- and *y*-terms have to be negative; hence let us denote these *variable separation constants* as $(-\alpha^2)$ and $(-\beta^2)$, respectively. Now Eq. (87) shows that the constant *z*-term has to be positive; denoting it as γ^2 we get the following relation:

$$\alpha^2 + \beta^2 = \gamma^2. \tag{2.88}$$

Now the variables are separated in the sense that for the functions $X(x)$, $Y(y)$, and $Z(z)$ we got separate ordinary differential equations,

$$\frac{d^2X}{dx^2} + \alpha^2 X = 0, \qquad \frac{d^2Y}{dy^2} + \beta^2 Y = 0, \qquad \frac{d^2Z}{dz^2} - \gamma^2 Z = 0, \tag{2.89}$$

which are related only by Eq. (88) for their constant parameters.

Let us start with the equation for $X(x)$. Its general solution is the sum of functions $\sin\alpha x$ and $\cos\alpha x$, multiplied by arbitrary coefficients. Let us select these coefficients to satisfy our boundary conditions. First, since $\phi \propto X$ should vanish at the back vertical wall of the box (i.e., with the coordinate origin choice shown in Fig. 13, at $x = 0$ for any *y* and *z*), the coefficient at $\cos\alpha x$ should be zero. The remaining coefficient (at $\sin\alpha x$) may be included in the general factor c_k in Eq. (84), so we may take *X* in the form

$$X = \sin\alpha x. \tag{2.90}$$

This solution satisfies the boundary condition at the opposite wall ($x = a$) only if the product αa is a multiple of π, i.e. if α is equal to any of the following numbers (commonly called *eigenvalues*):[38]

$$\alpha_n = \frac{\pi}{a}n, \qquad \text{with } n = 1, 2, ... \tag{2.91}$$

(Terms with negative values of *n* would not be linearly-independent from those with positive *n* and may be dropped from the sum (84). The value $n = 0$ is formally possible but would give $X = 0$, i.e. $\phi_k = 0$, at

[38] Note that according to Eqs. (91)-(92), as the spatial dimensions *a* and *b* of the system are increased, the distances between the adjacent eigenvalues tend to zero. This fact implies that for spatially infinite systems, the eigenvalue spectra are continuous, so the sums of the type (84) become integrals; however, the general approach remains the same. A few problems of this type are provided in Sec. 9 for the reader's exercise.

any x, i.e. no contribution to sum (84), so it may be dropped as well.) Now we see that we indeed had to take α real, i.e. α^2 positive – otherwise, instead of the oscillating function (90), we would have a sum of two exponential functions, which cannot equal zero at two independent points of the x-axis.

Since the equation (89) for function $Y(y)$ is similar to that for $X(x)$, and the boundary conditions on the walls perpendicular to axis y ($y = 0$ and $y = b$) are similar to those for x-walls, the absolutely similar reasoning gives

$$Y = \sin \beta y, \qquad \beta_m = \frac{\pi}{b} m, \qquad \text{with } m = 1, 2, \ldots, \tag{2.92}$$

where the integer m may be selected independently of n. Now we see that according to Eq. (88), the separation constant γ depends on two integers n and m, so the relationship may be rewritten as

$$\gamma_{nm} = \left[\alpha_n^2 + \beta_m^2\right]^{1/2} = \pi \left[\left(\frac{n}{a}\right)^2 + \left(\frac{m}{b}\right)^2\right]^{1/2}. \tag{2.93}$$

The corresponding solution of the differential equation for Z may be represented as a linear combination of two exponents $\exp\{\pm\gamma_{nm} z\}$, or alternatively of two hyperbolic functions, $\sinh\gamma_{nm} z$ and $\cosh\gamma_{nm} z$, with arbitrary coefficients. At our choice of coordinate origin, the latter option is preferable because $\cosh\gamma_{nm} z$ cannot satisfy the zero boundary condition at the bottom lid of the box ($z = 0$). Hence we may take Z in the form

$$Z = \sinh \gamma_{nm} z, \tag{2.94}$$

which automatically satisfies that condition.

Now it is the right time to merge Eqs. (84)-(85) and (90)-(94), replacing the temporary index k with the full set of possible eigenvalues, in our current case of two integer indices n and m:

Variable separation in Cartesian coordinates (example)

$$\phi(x, y, z) = \sum_{n,m=1}^{\infty} c_{nm} \sin\frac{\pi n x}{a} \sin\frac{\pi m y}{b} \sinh \gamma_{nm} z, \tag{2.95}$$

where γ_{nm} is given by Eq. (93). This solution satisfies not only the Laplace equation but also the boundary conditions on all walls of the box, besides the top lid, for arbitrary coefficients c_{nm}. The only job left is to choose these coefficients from the top-lid requirement:

$$\phi(x, y, c) \equiv \sum_{n,m=1}^{\infty} c_{nm} \sin\frac{\pi n x}{a} \sin\frac{\pi m y}{b} \sinh \gamma_{nm} c = V(x, y). \tag{2.96}$$

It may look bad to have just one equation for the infinite set of coefficients c_{nm}. However, the decisive help comes from the fact that the functions of x and y that participate in Eq. (96), form *full*, *orthogonal* sets of 1D functions. The last term means that the integrals of the products of the functions with different integer indices over the region of interest equal zero. Indeed, direct integration gives

$$\int_0^a \sin\frac{\pi n x}{a} \sin\frac{\pi n' x}{a} dx = \frac{a}{2} \delta_{nn'}, \tag{2.97}$$

where $\delta_{nn'}$ is the Kronecker symbol, and similarly for y (with the evident replacements $a \to b$, and $n \to m$). Hence, a fruitful way to proceed is to multiply both sides of Eq. (96) by the product of the basis functions, with arbitrary indices n' and m', and integrate the result over x and y:

$$\sum_{n,m=1}^{\infty} c_{nm} \sinh \gamma_{nm} c \int_0^a \sin\frac{\pi n x}{a} \sin\frac{\pi n' x}{a} dx \int_0^b \sin\frac{\pi m y}{b} \sin\frac{\pi m' y}{b} dy = \int_0^a dx \int_0^b dy\, V(x,y) \sin\frac{\pi n' x}{a} \sin\frac{\pi m' y}{b} . \quad (2.98)$$

Due to Eq. (97), all terms on the left-hand side of the last equation, besides those with $n = n'$ and $m = m'$, vanish, and (replacing n' with n, and m' with m, for notation brevity) we finally get

$$c_{nm} = \frac{4}{ab \sinh \gamma_{nm} c} \int_0^a dx \int_0^b dy\, V(x,y) \sin\frac{\pi n x}{a} \sin\frac{\pi m y}{b} . \quad (2.99)$$

The relations (93), (95), and (99) give the complete solution of the posed boundary problem; we can see both good and bad news here. The first bit of bad news is that in the general case, we still need to work out the integrals (99) – formally, the infinite number of them. In some cases, it is possible to do this analytically, in one shot. For example, if the top lid in our problem is a single conductor, i.e. has a constant potential V_0, we may take $V(x,y) = V_0 = \text{const}$, and both 1D integrations are elementary; for example

$$\int_0^a \sin\frac{\pi n x}{a} dx = \frac{a}{\pi n} \int_0^{\pi n} \sin\xi d\xi = \frac{a}{\pi n} \times \begin{cases} 2, & \text{for } n \text{ odd}, \\ 0, & \text{for } n \text{ even}, \end{cases} \quad (2.100)$$

and similarly for the integral over y, so

$$c_{nm} = \frac{16 V_0}{\pi^2 nm \sinh \gamma_{nm} c} \times \begin{cases} 1, & \text{if both } n \text{ and } m \text{ are odd}, \\ 0, & \text{otherwise}. \end{cases} \quad (2.101)$$

The second bad news is that even on such a happy occasion, we still have to sum up the series (95), so our result may only be called analytical with some reservations because in most cases we need to perform numerical summation to get the final numbers or plots.

Now the first *good* news. Computers are very efficient for both operations (95) and (99), i.e. for the summation and integration. (As was discussed in Sec. 1.2, random errors are averaged out at these operations.) As an example, Fig. 14 shows the plots of the electrostatic potential in a cubic box ($a = b = c$), with an equipotential top lid ($V = V_0 = \text{const}$), obtained by a numerical summation of the series (95), using the analytical expression (101). The remarkable feature of this calculation is a very fast convergence of the series; for the middle cross-section of the cubic box ($z/c = 0.5$), already the first term (with $n = m = 1$) gives an accuracy of about 6%, while the sum of four leading terms (with $n, m = 1, 3$) reduces the error to just 0.2%. (For a longer box, $c > a, b$, the convergence is even faster – see the discussion below.) Only very close to the corners between the top lid and the sidewalls, where the potential changes rapidly, several more terms are necessary to get a reasonable accuracy.

The related piece of good news is that our "semi-analytical" result allows its ultimate limits to be explored analytically. For example, Eq. (93) shows that for a very flat box (with $c << a, b$), $\gamma_{n,m} z \le \gamma_{n,m} c << 1$ at least for the lowest terms of series (95), with $n, m << c/a, c/b$. In this case, the sinh functions in Eqs. (96) and (99) may be well approximated with their arguments, and their ratio by z/c. So if we limit the summation to these terms, Eq. (95) gives a very simple result

$$\phi(x,y) \approx \frac{z}{c} V(x,y) , \quad (2.102)$$

which means that each elementary segment of the flat box behaves just as a plane capacitor. Only near the sidewalls, the higher terms in the series (95) are important, producing some deviations from Eq. (102). (For the general problem with an arbitrary function $V(x,y)$, this is also true in all regions where this function changes sharply.)

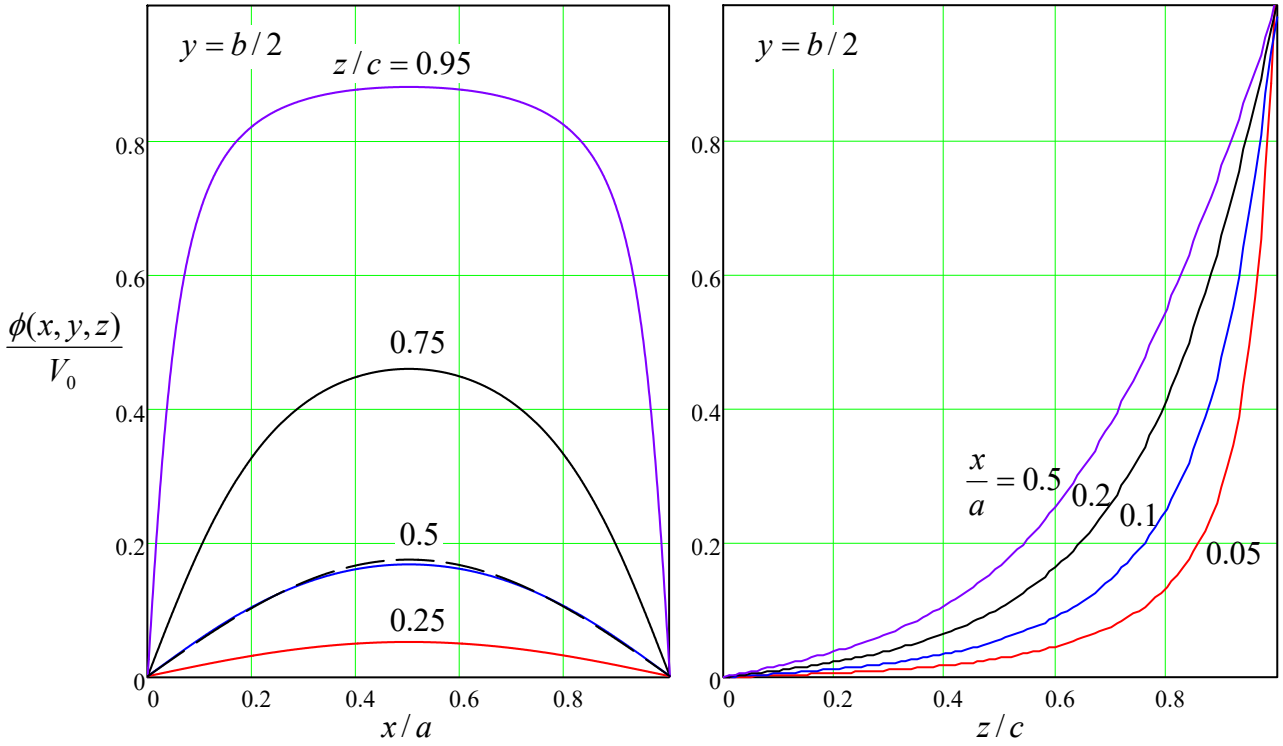


Fig. 2.14. The electrostatic potential's distribution inside a cubic box ($a = b = c$) with a constant voltage V_0 on the top lid (Fig. 13), calculated numerically from Eqs. (93), (95), and (101). The dashed line on the left panel shows the contribution of the main term of the series (with $n = m = 1$) to the full result, for $z/c = 0.5$.

In the opposite limit ($a, b << c$), Eq. (93) shows that on the contrary, $\gamma_{n,m}c >> 1$ for all n and m. Moreover, the ratio $\sinh\gamma_{n,m}z/\sinh\gamma_{n,m}c$ drops sharply if either n or m is increased, provided that z is not too close to c. Hence in this case a very good approximation may be obtained by keeping just the leading term, with $n = m = 1$, in Eq. (95), so the challenge of summation disappears. (As was discussed above, this approximation works reasonably well even for a cubic box.) In particular, for the constant potential of the upper lid, we can use Eq. (101) and the exponential asymptotic for both sinh functions, to get a very simple formula:

$$\phi = \frac{16}{\pi^2}\sin\frac{\pi x}{a}\sin\frac{\pi y}{b}\exp\left\{-\pi\frac{\left(a^2+b^2\right)^{1/2}}{ab}(c-z)\right\}. \qquad (2.103)$$

These results may be readily generalized to some other problems. For example, if all walls of the box shown in Fig. 13 have an arbitrary potential distribution, we may use the linear superposition principle to represent the electrostatic potential distribution as the sum of six partial solutions of the type of Eq. (95), each with one wall biased by the corresponding voltage, and all other grounded ($\phi = 0$).

To summarize, the results given by the variable separation method in the Cartesian coordinates are closer to what we could call a genuinely analytical solution than to a purely numerical solution.

Now, let us explore the issues that arise when this method is applied in other orthogonal coordinate systems.

2.6. Variable separation – polar coordinates

If a system of conductors is cylindrical, the potential distribution is independent of the z-coordinate along the cylinder axis: $\partial\phi/\partial z = 0$, and the Laplace equation becomes two-dimensional. If the conductor's cross-section is rectangular, the variable separation method works best in Cartesian coordinates $\{x, y\}$, and is just a particular case of the 3D solution discussed above. However, if the cross-section is circular, much more compact results may be obtained by using the polar coordinates $\{\rho, \varphi\}$. As we already know from Sec. 3(ii), these 2D coordinates are orthogonal, so the two-dimensional Laplace operator is a sum of two separable terms.[39] Requiring, just as we have done above, each component of the sum (84) to satisfy the Laplace equation, we get

$$\frac{1}{\rho}\frac{\partial}{\partial\rho}\left(\rho\frac{\partial\phi_k}{\partial\rho}\right)+\frac{1}{\rho^2}\frac{\partial^2\phi_k}{\partial\varphi^2}=0. \tag{2.104}$$

In a full analogy with Eq. (85), let us represent each particular solution ϕ_k as a product $\mathcal{R}(\rho)\mathcal{F}(\varphi)$. Plugging this expression into Eq. (104) and then dividing all its parts by $\mathcal{R}\mathcal{F}/\rho^2$, we get

$$\frac{\rho}{\mathcal{R}}\frac{d}{d\rho}\left(\rho\frac{d\mathcal{R}}{d\rho}\right)+\frac{1}{\mathcal{F}}\frac{d^2\mathcal{F}}{d\varphi^2}=0. \tag{2.105}$$

Following the same reasoning as for the Cartesian coordinates, we get two separated ordinary differential equations

$$\rho\frac{d}{d\rho}\left(\rho\frac{d\mathcal{R}}{d\rho}\right)=\nu^2\mathcal{R}, \tag{2.106}$$

$$\frac{d^2\mathcal{F}}{d\varphi^2}+\nu^2\mathcal{F}=0, \tag{2.107}$$

where ν^2 is the variable separation constant.

Let us start their analysis from Eq. (106), plugging into it a probe solution $\mathcal{R} = c\rho^\alpha$ where c and α are some constants. The elementary differentiation shows that if $\alpha \neq 0$, the equation is indeed satisfied for any c, with just one requirement imposed on the constant α, namely $\alpha^2 = \nu^2$. This means that the following linear superposition

$$\mathcal{R}=a_\nu\rho^{+\nu}+b_\nu\rho^{-\nu}, \qquad \text{for } \nu\neq 0, \tag{2.108}$$

with any constant coefficients a_ν and b_ν, is also a solution of Eq. (106). Moreover, the general theory of linear ordinary differential equations tells us that the solution of a second-order equation like Eq. (106) may only depend on just two constant factors that scale two linearly independent functions. Hence, for all values $\nu^2 \neq 0$, Eq. (108) presents the *general* solution of that equation. The case when $\nu = 0$, in which the functions $\rho^{+\nu}$ and $\rho^{-\nu}$ are just constants and hence are *not* linearly independent, is special, but in this case, the integration of Eq. (106) is straightforward,[40] giving

[39] See, e.g., MA Eq. (10.3) with $\partial/\partial z = 0$.

[40] Actually, we have already performed it in Sec. 3 – see Eq. (43).

$$\mathcal{R} = a_0 + b_0 \ln \rho, \qquad \text{for } \nu = 0. \tag{2.109}$$

In order to specify the separation constant, let us explore Eq. (107), whose general solution is

$$\mathcal{F} = \begin{cases} c_\nu \cos \nu\varphi + s_\nu \sin \nu\varphi, & \text{for } \nu \neq 0, \\ c_0 + s_0 \varphi, & \text{for } \nu = 0. \end{cases} \tag{2.110}$$

There are two possible cases here. In many boundary problems solvable in cylindrical coordinates, the free-space region, in which the Laplace equation is valid, extends continuously around the origin point $\rho = 0$. In this region, the potential has to be continuous and uniquely defined, so $\mathcal{F}$ has to be a 2π-periodic function of φ. For that, one needs the product $\nu(\varphi + 2\pi)$ to equal $\nu\varphi + 2\pi n$, with n being an integer, immediately giving us a discrete spectrum of possible values of the variable separation constant:

$$\nu = n = 0, \pm 1, \pm 2, \ldots \tag{2.111}$$

In this case, both functions $\mathcal{R}$ and $\mathcal{F}$ may be labeled with the integer index n. Taking into account that the terms with negative values of n may be summed up with those with positive n, and that s_0 has to equal zero (otherwise the 2π-periodicity of function $\mathcal{F}$ would be violated), we see that the general solution of the 2D Laplace equation for such geometries may be represented as

Variable separation in polar coordinates

$$\phi(\rho,\varphi) = a_0 + b_0 \ln \rho + \sum_{n=1}^{\infty} \left(a_n \rho^n + \frac{b_n}{\rho^n} \right) \left(c_n \cos n\varphi + s_n \sin n\varphi \right). \tag{2.112}$$

Let us see how all this machinery works on the famous problem of a round cylindrical conductor placed into an electric field that is uniform and perpendicular to the cylinder's axis at large distances (see Fig. 15a), as if it is created by a large plane capacitor. First of all, let us explore the effect of the system's symmetries on the coefficients in Eq. (112). Selecting the coordinate system as shown in Fig. 15a, and taking the cylinder's potential for zero, we immediately get $a_0 = 0$.

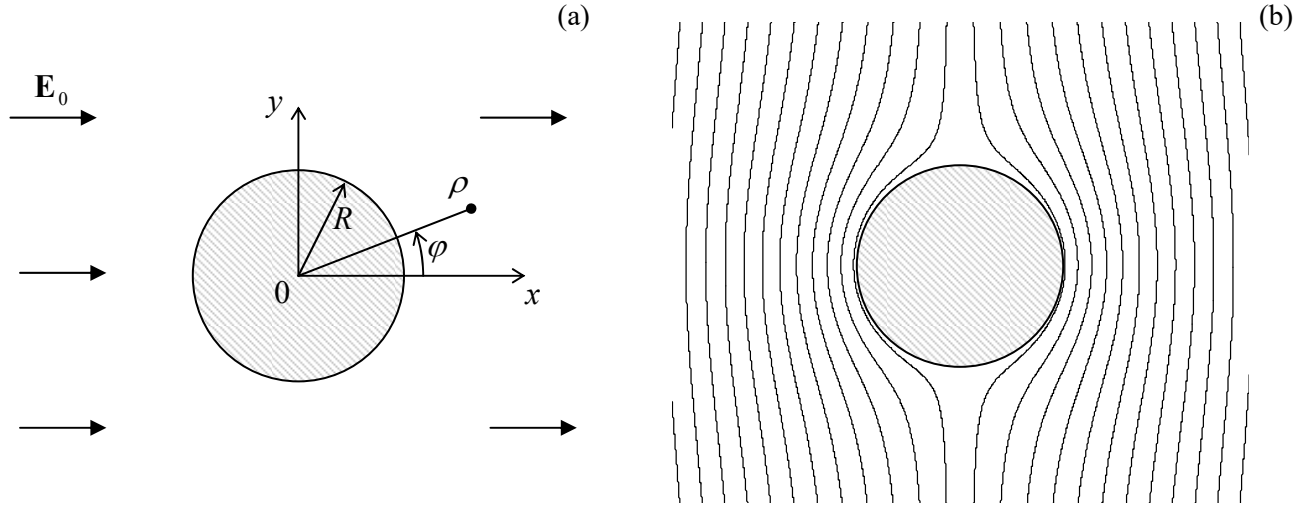


Fig. 2.15. A conducting cylinder inserted into an initially uniform electric field perpendicular to its axis: (a) the problem's geometry, and (b) the equipotential surfaces given by Eq. (117).

Moreover, due to the mirror symmetry about the plane [x, z], the solution has to be an even function of the angle φ, and hence all coefficients s_n should also equal zero. Also, at large distances ($\rho \gg R$) from the cylinder, its effect on the electric field should vanish, and the potential should approach that of the uniform external field $\mathbf{E} = E_0\mathbf{n}_x$:

$$\phi \to -E_0 x \equiv -E_0 \rho \cos\varphi, \quad \text{for } \rho \to \infty . \tag{2.113}$$

This is only possible if in Eq. (112), $b_0 = 0$, and also all coefficients a_n with $n \neq 1$ vanish, while the product a_1c_1 should be equal to $(-E_0)$. Thus the solution is reduced to the following form

$$\phi(\rho,\varphi) = -E_0 \rho \cos\varphi + \sum_{n=1}^{\infty} \frac{B_n}{\rho^n} \cos n\varphi , \tag{2.114}$$

in which the coefficients $B_n \equiv b_n c_n$ should be found from the boundary condition at $\rho = R$:

$$\phi(R,\varphi) = 0 . \tag{2.115}$$

This requirement yields the following equation,

$$\left(-E_0 R + \frac{B_1}{R} \right) \cos\varphi + \sum_{n=2}^{\infty} \frac{B_n}{R^n} \cos n\varphi = 0, \tag{2.116}$$

which should be satisfied for all φ. This equality, read backward, may be considered as an expansion of a function identically equal to zero into a series over mutually orthogonal functions $\cos n\varphi$. It is evidently valid if all coefficients of the expansion, including $(-E_0R + B_1/R)$, and all B_n for $n \geq 2$ are equal to zero. Moreover, mathematics tells us that such expansions are unique, so this is the only possible solution of Eq. (116). So, $B_1 = E_0R^2$, and our final answer (valid only outside of the cylinder, i.e. for $\rho \geq R$), is

$$\phi(\rho,\varphi) = -E_0 \left(\rho - \frac{R^2}{\rho} \right) \cos\varphi \equiv -E_0 \left(1 - \frac{R^2}{x^2 + y^2} \right) x . \tag{2.117}$$

This result, which may be graphically represented with the equipotential surfaces shown in Fig. 15b, shows a smooth transition between the uniform field (113) far from the cylinder, to the equipotential surface of the cylinder (with $\phi = 0$). Such smoothening is very typical for Laplace equation solutions. Indeed, as we know from Chapter 1, these solutions correspond to the lowest integral of the potential gradient's square, i.e. to the lowest potential energy (1.65) possible at the given boundary conditions.

To complete the problem, let us use Eq. (3) to calculate the distribution of the surface charge density over the cylinder's cross-section:

$$\sigma = \varepsilon_0 E_n \big|_{\text{surface}} \equiv -\varepsilon_0 \frac{\partial \phi}{\partial \rho}\Big|_{\rho = R} = \varepsilon_0 E_0 \cos\varphi \frac{\partial}{\partial \rho} \left(\rho - \frac{R^2}{\rho} \right)_{\rho = R} = 2\varepsilon_0 E_0 \cos\varphi. \tag{2.118}$$

This very simple formula shows that with the field direction shown in Fig. 15a ($E_0 > 0$), the surface charge is positive on the right-hand side of the cylinder and negative on its left-hand side, thus creating a field directed from the right to the left, which exactly compensates the external field inside the conductor, where the net field is zero. (Please take one more look at the schematic Fig. 1a.) Note also that the net electric charge of the cylinder is zero, in correspondence with the problem symmetry.

Another useful by-product of the calculation (118) is that the surface electric field equals $2E_0\cos\varphi$, and hence its largest magnitude is twice the field far from the cylinder. Such electric field concentration is very typical for all convex conducting surfaces.

The last observation gets additional confirmation from the second possible topology when Eq. (110) is used to describe problems with no angular periodicity. A typical example of this situation is a cylindrical conductor with a cross-section that features a corner limited by two straight-line segments (Fig. 16). Indeed, we may argue that at $\rho < R$ (where R is the radial extension of the planar sides of the corner, see Fig. 16), the Laplace equation may be satisfied by a sum of partial solutions $\mathcal{R}(\rho)\mathcal{F}(\varphi)$, if the angular components of the products satisfy the boundary conditions on the corner sides. Taking (just for the simplicity of notation) the conductor's potential to be zero, and one of the corner's sides as the x-axis ($\varphi = 0$), these boundary conditions are

$$\mathcal{F}(0) = \mathcal{F}(\beta) = 0\,, \tag{2.119}$$

where the angle β may be anywhere between 0 and 2π – see Fig. 16.

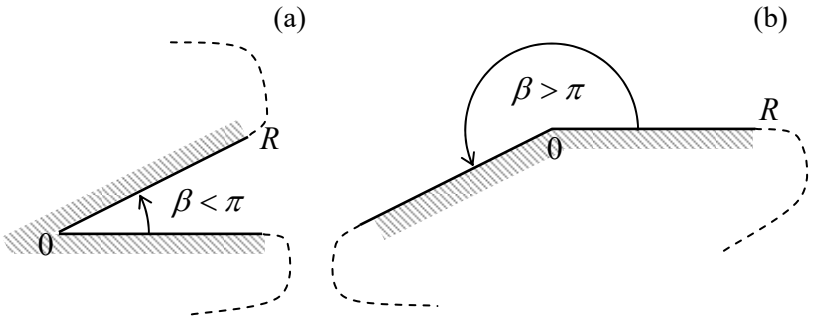


Fig. 2.16. The cross-sections of cylindrical conductors with (a) a corner and (b) a wedge.

Comparing this condition with Eq. (110), we see that it requires s_0 and all c_ν to vanish, and ν to take one of the values of the following discrete spectrum:

$$\nu_m\beta = \pi m, \qquad \text{with } m = 1, 2, \ldots. \tag{2.120}$$

Hence the full solution of the Laplace equation for this geometry takes the form

$$\phi = \sum_{m=1}^{\infty} a_m \rho^{\pi m/\beta} \sin\frac{\pi m\varphi}{\beta}, \qquad \text{for } \rho < R, \quad 0 \le \varphi \le \beta\,, \tag{2.121}$$

where the constants s_ν have been incorporated into a_m. The set of coefficients a_m cannot be universally determined, because it depends on the exact shape of the conductor outside the corner, and the externally applied electric field. However, whatever the set is, in the limit $\rho \to 0$, the solution (121) is almost[41] always dominated by the term with the lowest $m = 1$:

$$\phi \to a_1 \rho^{\pi/\beta} \sin\frac{\pi}{\beta}\varphi\,, \tag{2.122}$$

because the higher terms tend to zero faster. This potential distribution corresponds to the surface charge density

[41] Exceptions are possible only for highly symmetric configurations when the external field is specially crafted to make $a_1 = 0$. In this case, the solution at $\rho \to 0$ is dominated by the first nonzero term of the series (121).

$$\sigma = \varepsilon_0 E_n\big|_{\text{surface}} = -\varepsilon_0 \left.\frac{\partial \phi}{\partial(\rho\varphi)}\right|_{\rho=\text{const},\ \varphi\to+0} = -\varepsilon_0 \frac{\pi a_1}{\beta}\rho^{(\pi/\beta-1)}. \tag{2.123}$$

(It is similar, with the opposite sign, on the opposite face of the angle.)

The result (123) shows that if we are dealing with a concave corner ($\beta < \pi$, see Fig. 16a), the charge density (and the surface electric field) tends to zero. On the other hand, at a "convex corner" with $\beta > \pi$ (actually, a wedge – see Fig. 16b), both the charge and the field's strength concentrate, formally diverging at $\rho \to 0$. (So, do not sit on a roof's ridge during a thunderstorm; rather hide in a ditch!) We have already seen qualitatively similar effects for the thin round disk and the split plane.

2. 7. Variable separation – cylindrical coordinates

Now, let us discuss how to generalize the approach discussed in the previous section to problems whose geometry is still axially symmetric, but where the electrostatic potential depends not only on the radial and angular coordinates but also on the axial coordinate: $\partial\phi/\partial z \neq 0$. The classical example of such a problem is shown in Fig. 17. Here the sidewall and the bottom lid of a hollow round cylinder are kept at a fixed potential (say, $\phi = 0$), but the potential V fixed at the top lid is different. Evidently, this problem is qualitatively similar to the rectangular box problem solved above (Fig. 13), and we will also try to solve it first for the case of arbitrary voltage distribution over the top lid: $V = V(\rho, \varphi)$.

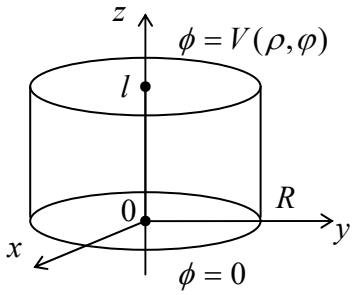


Fig. 2.17. A cylindrical volume with conducting walls.

Following the main idea of the variable separation method, let us require that each partial function ϕ_k in Eq. (84) satisfies the Laplace equation, now in the full cylindrical coordinates $\{\rho, \varphi, z\}$:[42]

$$\frac{1}{\rho}\frac{\partial}{\partial\rho}\left(\rho\frac{\partial\phi_k}{\partial\rho}\right) + \frac{1}{\rho^2}\frac{\partial^2\phi_k}{\partial\varphi^2} + \frac{\partial^2\phi_k}{\partial z^2} = 0. \tag{2.124}$$

Plugging ϕ_k in the form of the product $\mathcal{R}(\rho)\mathcal{F}(\varphi)\mathcal{Z}(z)$ into Eq. (124) and then dividing all resulting terms by this product, we get

$$\frac{1}{\rho\mathcal{R}}\frac{d}{d\rho}\left(\rho\frac{d\mathcal{R}}{d\rho}\right) + \frac{1}{\rho^2\mathcal{F}}\frac{d^2\mathcal{F}}{d\varphi^2} + \frac{1}{\mathcal{Z}}\frac{d^2\mathcal{Z}}{dz^2} = 0. \tag{2.125}$$

Since the first two terms of Eq. (125) can only depend on the polar variables ρ and φ, while the third term, only on z, at least that term should equal a constant. Denoting it (just like we did in the rectangular box problem) by γ^2, we get the following set of two equations:

[42] See, e.g., MA Eq. (10.3).

$$\frac{d^2\mathcal{Z}}{dz^2} = \gamma^2 \mathcal{Z}, \tag{2.126}$$

$$\frac{1}{\rho\mathcal{R}}\frac{d}{d\rho}\left(\rho\frac{d\mathcal{R}}{d\rho}\right) + \gamma^2 + \frac{1}{\rho^2\mathcal{F}}\frac{d^2\mathcal{F}}{d\varphi^2} = 0. \tag{2.127}$$

Now, multiplying all the terms of Eq. (127) by ρ^2, we see that the last term of the result, $(d^2\mathcal{F}/d\varphi^2)/\mathcal{F}$, may depend only on φ, and thus should equal a constant. Calling that constant ν^2 (just as in Sec. 6 above), we separate Eq. (127) into an angular equation,

$$\frac{d^2\mathcal{F}}{d\varphi^2} + \nu^2\mathcal{F} = 0, \tag{2.128}$$

and a radial equation:

$$\frac{d^2\mathcal{R}}{d\rho^2} + \frac{1}{\rho}\frac{d\mathcal{R}}{d\rho} + \left(\gamma^2 - \frac{\nu^2}{\rho^2}\right)\mathcal{R} = 0. \tag{2.129}$$

We see that the ordinary differential equations for the functions $\mathcal{Z}(z)$ and $\mathcal{F}(\varphi)$ (and hence their solutions) are identical to those discussed earlier in this chapter. However, Eq. (129) for the radial function $\mathcal{R}(\rho)$ (called the *Bessel equation*) is more complex than in the 2D case and depends on two independent constant parameters, γ and ν. The latter challenge may be readily overcome if we notice that any change of γ may be reduced to the corresponding re-scaling of the radial coordinate ρ. Indeed, introducing a dimensionless variable $\xi \equiv \gamma\rho$,[43] Eq. (129) may be reduced to an equation with just one parameter, ν:

Bessel equation

$$\boxed{\frac{d^2\mathcal{R}}{d\xi^2} + \frac{1}{\xi}\frac{d\mathcal{R}}{d\xi} + \left(1 - \frac{\nu^2}{\xi^2}\right)\mathcal{R} = 0.} \tag{2.130}$$

Moreover, we already know that for angle-periodic problems, the spectrum of eigenvalues of Eq. (128) is discrete: $\nu = n$, with integer n.

Unfortunately, even in this case, Eq. (130), which is the canonical form of the Bessel equation, cannot be satisfied by a single "elementary" function. The solutions that we need for our current problem are called the *Bessel function of the first kind of order* ν, commonly denoted as $J_\nu(\xi)$. Let me review in brief those properties of these functions that are most relevant to our problem – and many other problems discussed in this series.[44]

First of all, the Bessel function of a negative integer order is very simply related to that of the positive order:

$$J_{-n}(\xi) = (-1)^n J_n(\xi), \tag{2.131}$$

enabling us to limit our discussion to the functions with $n \geq 0$. Figure 18 shows four of these functions with the lowest positive n.

[43] Note that this normalization is specific for each value of the variable separation parameter γ. Also, please notice that the normalization is meaningless for $\gamma = 0$, i.e. for the case $Z(z) = \text{const}$. However, if we need partial solutions for this particular value of γ, we can always use Eqs. (108)-(109).

[44] For a more complete discussion of these functions, see the literature listed in MA Sec. 16, for example, Chapter 6 (written by F. Olver) in the famous collection compiled and edited by Abramowitz and Stegun, available online.

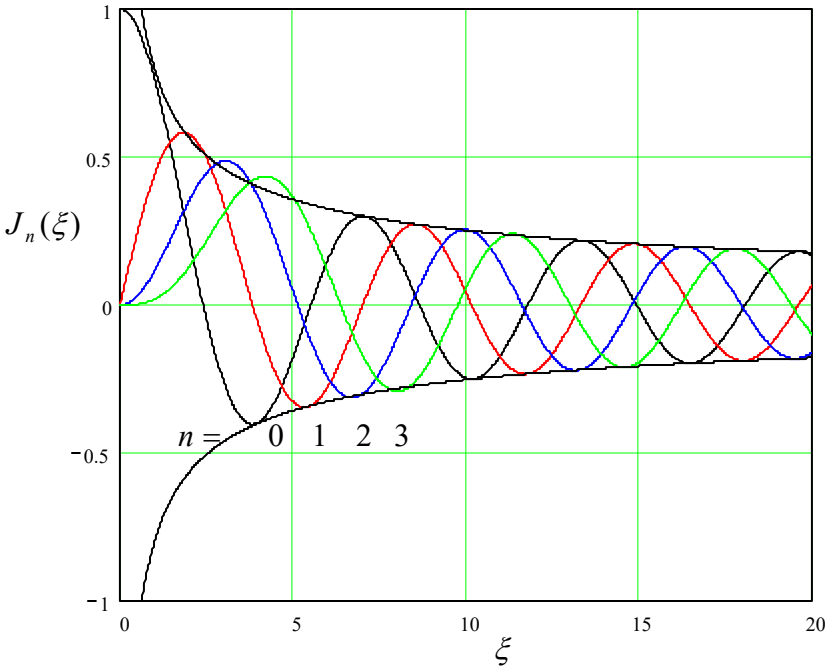


Fig. 2.18. Several Bessel functions $J_n(\xi)$ of integer order. The dashed lines show the envelope of the asymptotes (135).

As its argument is increased, each function is initially close to a power law: $J_0(\xi) \approx 1$, $J_1(\xi) \approx \xi/2$, $J_2(\xi) \approx \xi^2/8$, etc. This behavior follows from the Taylor series

$$J_n(\xi) = \left(\frac{\xi}{2}\right)^n \sum_{k=0}^{\infty} \frac{(-1)^k}{k!(n+k)!}\left(\frac{\xi}{2}\right)^{2k}, \tag{2.132}$$

which is formally valid for any ξ, and may even serve as an alternative definition of the functions $J_n(\xi)$. However, the series is converging fast only at small arguments, $\xi << n$, where its leading term is

$$J_n(\xi)\Big|_{\xi\to 0} \to \frac{1}{n!}\left(\frac{\xi}{2}\right)^n. \tag{2.133}$$

At $\xi \approx n + 1.86n^{1/3}$, the Bessel function reaches its maximum[45]

$$\max_{\xi}\left[J_n(\xi)\right] \approx \frac{0.675}{n^{1/3}}, \tag{2.134}$$

and then starts to oscillate with a period gradually approaching 2π, a phase shift that increases by $\pi/2$ with each unit increment of n, and an amplitude that decreases as $\xi^{-1/2}$. All these features are described by the following asymptotic formula:

$$J_n(\xi)\Big|_{\xi\to\infty} \to \left(\frac{2}{\pi\xi}\right)^{1/2} \cos\left(\xi - \frac{\pi}{4} - \frac{n\pi}{2}\right), \tag{2.135}$$

which starts to give a reasonable approximation soon after the function peaks – see Fig. 18.[46]

[45] These two approximations for the Bessel function peak are strictly valid for $n >> 1$, but may be used for reasonable estimates starting already from $n = 1$. For example, $\max_{\xi}[J_1(\xi)]$ is close to 0.58 and is reached at $\xi \approx 2.4$, just about 30% away from the values given by the asymptotic formulas.

Now we are ready for our case study (Fig. 17). Since the functions the $Z(z)$ have to satisfy not only Eq. (126) but also the bottom-lid boundary condition $Z(0) = 0$, they are proportional to $\sinh\gamma z$ – cf. Eq. (94). Then Eq. (84) becomes

$$\phi = \sum_{n=0}^{\infty}\sum_{\gamma} J_n(\gamma\rho)\left(c_{n\gamma}\cos n\varphi + s_{n\gamma}\sin n\varphi\right)\sinh\gamma z\,. \tag{2.136}$$

Next, we need to satisfy the zero boundary condition at the cylinder's side wall ($\rho = R$). This may be ensured by taking

$$J_n(\gamma R) = 0\,. \tag{2.137}$$

Since each function $J_n(x)$ has an infinite number of positive zeros (see Fig. 18 again), which may be numbered by an integer index $m = 1, 2, \ldots$, Eq. (137) may be satisfied with an infinite number of discrete values of the parameter γ:

$$\gamma_{nm} = \frac{\xi_{nm}}{R}, \tag{2.138}$$

where ξ_{nm} is the m-th zero of the function $J_n(x)$ – see the top numbers in the cells of Table 1. (Very soon we will see what we need the bottom numbers for.)

Table 2.1. Approximate values of a few first zeros, ξ_{nm}, of a few lowest-order Bessel functions $J_n(\xi)$ (the top number in each cell), and the values of $dJ_n(\xi)/d\xi$ at these points (the bottom number).

	$m = 1$	2	3	4	5	6
$n = 0$	2.40482 -0.51914	5.52008 +0.34026	8.65372 -0.27145	11.79215 +0.23245	14.93091 -0.20654	18.07106 +0.18773
1	3.83171 -0.40276	7.01559 +0.30012	10.17347 -0.24970	13.32369 +0.21836	16.47063 -0.19647	19.61586 +0.18006
2	5.13562 -0.33967	8.41724 +0.27138	11.61984 -0.23244	14.79595 +0.20654	17.95982 -0.18773	21.11700 +0.17326
3	6.38016 -0.29827	9.76102 +0.24942	13.01520 -0.21828	16.22347 +0.19644	19.40942 -0.18005	22.58273 +0.16718
4	7.58834 -0.26836	11.06471 +0.23188	14.37254 -0.20636	17.61597 +0.18766	20.82693 -0.17323	24.01902 +0.16168
5	8.77148 -0.24543	12.33860 +0.21743	15.70017 -0.19615	18.98013 +0.17993	22.21780 -0.16712	25.43034 +0.15669

Hence, Eq. (136) may be represented in a more explicit form:

Variable separation in cylindrical coordinates (example)

$$\phi(\rho,\varphi,z) = \sum_{n=0}^{\infty}\sum_{m=1}^{\infty} J_n\left(\xi_{nm}\frac{\rho}{R}\right)\left(c_{nm}\cos n\varphi + s_{nm}\sin n\varphi\right)\sinh\left(\xi_{nm}\frac{z}{R}\right). \tag{2.139}$$

[46] Eq. (135) and Fig. 18 clearly show the close analogy between the Bessel functions and the usual trigonometric functions, sine and cosine. To emphasize this similarity, and help the reader to develop more gut feeling of the Bessel functions, let me mention one result of the elasticity theory: while the sinusoidal functions describe, in particular, transverse standing waves on a guitar string, the functions $J_n(\xi)$ describe, in particular, transverse standing waves on an elastic round membrane (say, a round drum), with $J_0(\xi)$ describing their lowest (fundamental) mode – the only mode with a nonzero amplitude of the membrane center's oscillations.

Here the coefficients c_{nm} and s_{nm} have to be selected to satisfy the only remaining boundary condition – that on the top lid:

$$\phi(\rho,\varphi,l) \equiv \sum_{n=0}^{\infty}\sum_{m=1}^{\infty} J_n\left(\xi_{nm}\frac{\rho}{R}\right)\left(c_{nm}\cos n\varphi + s_{nm}\sin n\varphi\right)\sinh\left(\xi_{nm}\frac{l}{R}\right) = V(\rho,\varphi). \tag{2.140}$$

To use it, let us multiply both sides of Eq. (140) by the product $J_n(\xi_{nm'}\rho/R)\cos n'\varphi$, integrate the result over the lid area, and use the following property of the Bessel functions:

$$\int_0^1 J_n(\xi_{nm}s)J_n(\xi_{nm'}s)sds = \frac{1}{2}\left[J_{n+1}(\xi_{nm})\right]^2\delta_{mm'}. \tag{2.141}$$

As a small but important detour, the last relation expresses a very specific ("2D") orthogonality of the Bessel functions with different indices m – do not confuse them with the function order indices n, please![47] Since it relates two Bessel functions of the same order n, it is natural to ask why its right-hand side contains the function with a different order $(n + 1)$. Some gut feeling of that may come from one more very important property of the Bessel functions, the so-called *recurrence relations*:[48]

$$J_{n-1}(\xi) + J_{n+1}(\xi) = \frac{2nJ_n(\xi)}{\xi}, \tag{2.142a}$$

$$J_{n-1}(\xi) - J_{n+1}(\xi) = 2\frac{dJ_n(\xi)}{d\xi}, \tag{2.142b}$$

which in particular yield the following formula (convenient for working out some Bessel function integrals):

$$\frac{d}{d\xi}\left[\xi^n J_n(\xi)\right] = \xi^n J_{n-1}(\xi). \tag{2.143}$$

Let us apply the recurrence relations at the special points ξ_{nm}. At these points, J_n vanishes, and the system of two equations (142) may be readily solved to get, in particular,

$$J_{n+1}(\xi_{nm}) = -\frac{dJ_n}{d\xi}(\xi_{nm}), \tag{2.144}$$

so the square bracket on the right-hand side of Eq. (141) is just $(dJ_n/d\xi)^2$ at $\xi = \xi_{nm}$. Thus the values of the Bessel function derivatives at the zero points of the function, given by the lower numbers in the cells of Table 1, are as important for boundary problem solutions as the zeros themselves.

Now returning to our problem: since the angular functions $\cos n\varphi$ are also orthogonal – both to each other,

---

[47] The Bessel functions of the *same argument* but *different orders* are also orthogonal, but differently:

$$\int_0^{\infty} J_n(\xi)J_{n'}(\xi)\frac{d\xi}{\xi} = \frac{1}{n+n'}\delta_{nn'}.$$

[48] These relations provide, in particular, a convenient way for numerical computation of all $J_n(\xi)$ – after $J_0(\xi)$ has been computed. (The latter task is usually performed using Eq. (132) for smaller ξ and an extension of Eq. (135) for larger ξ.) Note that most mathematical software packages, including all those listed in MA Sec. 16(iv), include ready subroutines for calculation of the functions $J_n(\xi)$ and other special functions used in this lecture series. In this sense, the conditional line separating these "special functions" from "elementary functions" is rather fine.

$$\int_0^{2\pi} \cos(n\varphi)\cos(n'\varphi)\, d\varphi = \pi\delta_{nn'}, \tag{2.145}$$

and to all functions $\sin n\varphi$, the integration over the lid area kills all terms of both series in Eq. (140), besides just one term proportional to $c_{n'm'}$, and hence gives an explicit expression for that coefficient. The counterpart coefficients $s_{n'm'}$ may be found by repeating the same procedure with the replacement of $\cos n'\varphi$ by $\sin n'\varphi$. This evaluation (left for the reader's exercise) completes the solution of our problem for an arbitrary lid potential $V(\rho,\varphi)$.

Still, before leaving the Bessel functions behind (for a while only :-), let me address two important issues. First, we have seen that in our cylinder problem (Fig. 17), the set of functions $J_n(\xi_{nm}\rho/R)$ with different indices m (which characterize the degree of Bessel function's stretch along axis ρ) play a role similar to that of functions $\sin(\pi mx/a)$ in the rectangular box problem shown in Fig. 13. In this context, what is the analog of functions $\cos(\pi mx/a)$ – which may be important for some boundary problems? In a more formal language, are there any functions of the same argument $\xi \equiv \xi_{nm}\rho/R$, that would be linearly independent of the Bessel functions of the first kind, while satisfying the same Bessel equation (130)?

The answer is *yes*. For the definition of such functions, we first need to generalize our prior formulas for $J_n(\xi)$, and in particular Eq. (132), to the case of arbitrary, not necessarily real order ν. Mathematics says that the generalization may be performed in the following way:

$$J_\nu(\xi) = \left(\frac{\xi}{2}\right)^\nu \sum_{k=0}^{\infty} \frac{(-1)^k}{k!\Gamma(\nu+k+1)}\left(\frac{\xi}{2}\right)^{2k}, \tag{2.146}$$

where $\Gamma(s)$ is the so-called *gamma function* that may be defined as[49]

$$\Gamma(s) \equiv \int_0^{\infty} \xi^{s-1} e^{-\xi}\, d\xi. \tag{2.147}$$

The simplest, and the most important property of the gamma function is that for integer values of its argument, it gives the factorial of the number smaller by one:

$$\Gamma(n+1) = n! \equiv 1\cdot 2\cdot \ldots \cdot n, \tag{2.148}$$

so it is essentially a generalization of the notion of the factorial to all real numbers.

The Bessel functions defined by Eq. (146) satisfy, after the replacements $n \to \nu$ and $n! \to \Gamma(n+1)$, virtually all the relations discussed above, including the Bessel equation (130), the asymptotic formula (135), the orthogonality condition (141), and the recurrence relations (142). Moreover, it may be shown that $\nu \neq n$, functions $J_\nu(\xi)$ and $J_{-\nu}(\xi)$ are linearly independent of each other, and hence their linear combination may be used to represent the general solution of the Bessel equation. Unfortunately, as Eq. (131) shows, for $\nu = n$ this is not true, and a solution linearly independent of $J_n(\xi)$ has to be formed differently. The most common way to do that is first to define, for all $\nu \neq n$, the following functions:

$$Y_\nu(\xi) \equiv \frac{J_\nu(\xi)\cos\nu\pi - J_{-\nu}(\xi)}{\sin\nu\pi}, \tag{2.149}$$

[49] See, e.g., MA Eq. (6.7a). Note that $\Gamma(s) \to \infty$ at $s \to 0, -1, -2, \ldots$

called the *Bessel functions of the second kind*, or more often the *Weber functions*,[50] and then to follow the limit $\nu \to n$. At this, both the numerator and denominator of the right-hand side of Eq. (149) tend to zero, but their ratio tends to a finite value called $Y_n(x)$. It may be shown that the resulting functions are still the solutions of the Bessel equation and are linearly independent of $J_n(x)$, though are related just as those functions if the sign of n changes:

$$Y_{-n}(\xi) = (-1)^n Y_n(\xi) . \tag{2.150}$$

Figure 19 shows a few Weber functions of the lowest integer orders.

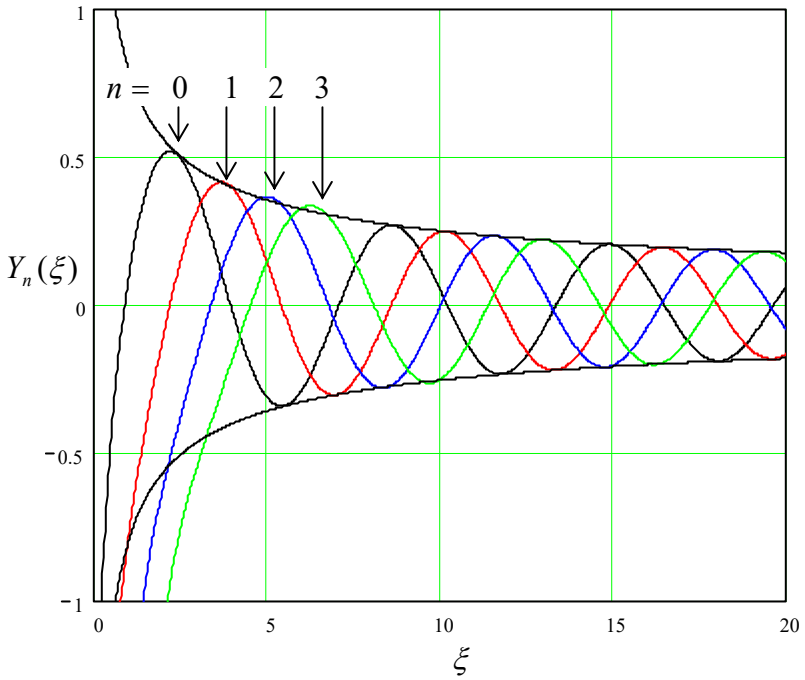


Fig. 2.19. A few Bessel functions of the second kind (a.k.a. the Weber functions, a.k.a. the Neumann functions).

The plots show that their asymptotic behavior is very similar to that of the functions $J_n(\xi)$:

$$Y_n(\xi) \to \left(\frac{2}{\pi\xi}\right)^{1/2} \sin\left(\xi - \frac{\pi}{4} - \frac{n\pi}{2}\right), \qquad \text{for } \xi \to \infty, \tag{2.151}$$

but with the phase shift necessary to make these Bessel functions orthogonal to those of the first order – cf. Eq. (135). However, for small values of argument ξ, the Bessel functions of the second kind behave completely differently from those of the first kind:

$$Y_n(\xi) \to \begin{cases} \dfrac{2}{\pi}\left(\ln\dfrac{\xi}{2} + \gamma\right), & \text{for } n = 0, \\ -\dfrac{(n-1)!}{\pi}\left(\dfrac{\xi}{2}\right)^{-n}, & \text{for } n \neq 0, \end{cases} \tag{2.152}$$

where γ is the so-called *Euler constant*, defined as follows:

[50] Sometimes, they are called the *Neumann* functions and denoted as $N_\nu(\xi)$.

$$\gamma \equiv \lim_{n\to\infty}\left(1+\frac{1}{2}+\frac{1}{3}+...+\frac{1}{n}-\ln n\right)\approx 0.577157\ldots \tag{2.153}$$

As Eqs. (152) and Fig. 19 show, the functions $Y_n(\xi)$ diverge at $\xi \to 0$ and hence cannot describe the behavior of any physical variable, in particular the electrostatic potential.

One may wonder: if this is true, when do we need these functions in physics? Figure 20 shows an example of a simple boundary problem of electrostatics, whose solution by the variable separation method involves both functions $J_n(\xi)$ and $Y_n(\xi)$.

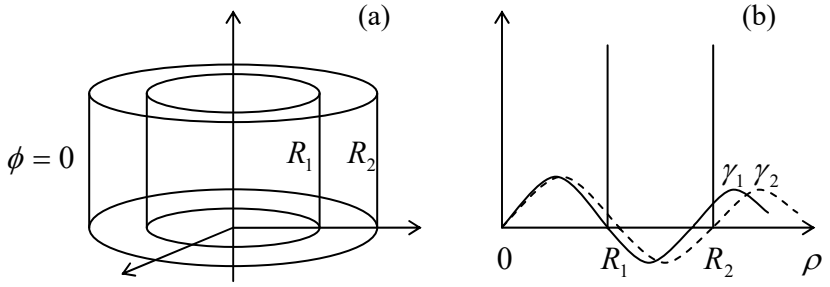


Fig. 2.20. A simple boundary problem that cannot be solved using just one kind of Bessel functions.

Here two round, conducting coaxial cylindrical tubes are kept at the same (say, zero) potential, but at least one of two lids has a different potential. The problem is almost completely similar to that discussed above (Fig. 17), but now we need to find the potential distribution in the free space between the tubes, i.e. for $R_1 < \rho < R_2$. If we use the same variable separation as in the simpler counterpart problem, we need the radial functions $\mathcal{R}(\rho)$ to satisfy two zero boundary conditions: at $\rho = R_1$ and $\rho = R_2$. With the Bessel functions of just the first kind, $J_n(\gamma\rho)$, it is impossible to do, because the two boundaries would impose two independent (and generally incompatible) conditions, $J_n(\gamma R_1) = 0$, and $J_n(\gamma R_2) = 0$, on one "stretching parameter" γ. The existence of the Bessel functions of the second kind immediately saves the day, because if the radial function solution is represented as a linear combination,

$$\mathcal{R} = c_J J_n(\gamma\rho) + c_Y Y_n(\gamma\rho), \tag{2.154}$$

two zero boundary conditions give two equations for γ and the ratio $c \equiv c_Y/c_J$.[51] (Due to the oscillating character of both Bessel functions, these conditions would be typically satisfied by an infinite set of discrete pairs $\{\gamma, c\}$.) Note, however, that generally none of these pairs would correspond to zeros of either J_n or Y_n, so having an analog of Table 1 for the latter function would not help much. Hence, even the simplest problems of this kind (like the one shown in Fig. 20) typically require the numerical solution of transcendental algebraic equations.

[51] A pair of independent linear functions, used for the representation of the general solution of the Bessel equation, may be also chosen differently, using the so-called *Hankel functions*

$$H_n^{(1,2)}(\xi) \equiv J_n(\xi) \pm iY_n(\xi).$$

For representing the general solution of Eq. (130), this alternative is completely similar, for example, to using the pair of complex functions $\exp\{\pm i\alpha x\} \equiv \cos\alpha x \pm i\sin\alpha x$ instead of the pair of real functions $\{\cos\alpha x, \sin\alpha x\}$ for the representation of the general solution of Eq. (89) for $X(x)$.

In order to complete the discussion of variable separation in the cylindrical coordinates, one more issue to address is the so-called *modified Bessel functions*: of the *first kind*, $I_\nu(\xi)$, and of the *second kind*, $K_\nu(\xi)$. They are two linearly independent solutions of the *modified Bessel equation*,

$$\frac{d^2\mathcal{R}}{d\xi^2}+\frac{1}{\xi}\frac{d\mathcal{R}}{d\xi}-\left(1+\frac{\nu^2}{\xi^2}\right)\mathcal{R}=0\,, \tag{2.155}$$

Modified Bessel equation

which differs from Eq. (130) "only" by the sign of one of its terms. Figure 21 shows a simple problem that leads (among many others) to this equation: a round thin conducting cylindrical pipe is sliced, perpendicular to its axis, to rings of equal height h, which are kept at equal but sign-alternating potentials.

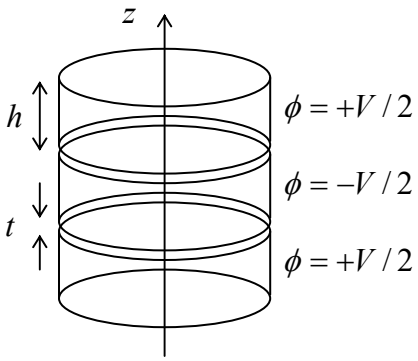


Fig. 2.21. A typical boundary problem whose solution may be conveniently described in terms of the modified Bessel functions.

If the system is very long (formally, infinite) in the z-direction, we may use the variable separation method for the solution of this problem, but now evidently need periodic (rather than exponential) solutions along the z-axis, i.e. linear combinations of $\sin kz$ and $\cos kz$ with various real values of the constant k. Separating the variables, we arrive at a differential equation similar to Eq. (129), but with the negative sign before the separation constant:

$$\frac{d^2\mathcal{R}}{d\rho^2}+\frac{1}{\rho}\frac{d\mathcal{R}}{d\rho}-(k^2+\frac{\nu^2}{\rho^2})\mathcal{R}=0\,. \tag{2.156}$$

The same radial coordinate's normalization, $\xi \equiv k\rho$, immediately leads us to Eq. (155), and hence (for $\nu = n$) to the modified Bessel functions $I_n(\xi)$ and $K_n(\xi)$.

Figure 22 shows the behavior of such functions, of a few lowest orders. One can see that at $\xi \to 0$ the behavior is virtually similar to that of the "usual" Bessel functions – cf. Eqs. (132) and (152), with $K_n(\xi)$ multiplied (by purely historical reasons) by an additional coefficient, $\pi/2$:

$$I_n(\xi)\to\frac{1}{n!}\left(\frac{\xi}{2}\right)^n, \qquad K_n(\xi)\to\begin{cases}-\left[\ln\left(\dfrac{\xi}{2}\right)+\gamma\right], & \text{for } n=0,\\ \dfrac{(n-1)!}{2}\left(\dfrac{\xi}{2}\right)^{-n}, & \text{for } n\neq 0,\end{cases} \tag{2.157}$$

However, the asymptotic behavior of the modified functions is very much different, with $I_n(x)$ exponentially growing, and $K_n(\xi)$ exponentially dropping at $\xi \to \infty$:

$$I_n(\xi) \to \left(\frac{1}{2\pi\xi}\right)^{1/2} e^{\xi}, \qquad K_n(\xi) \to \left(\frac{\pi}{2\xi}\right)^{1/2} e^{-\xi}. \tag{2.158}$$

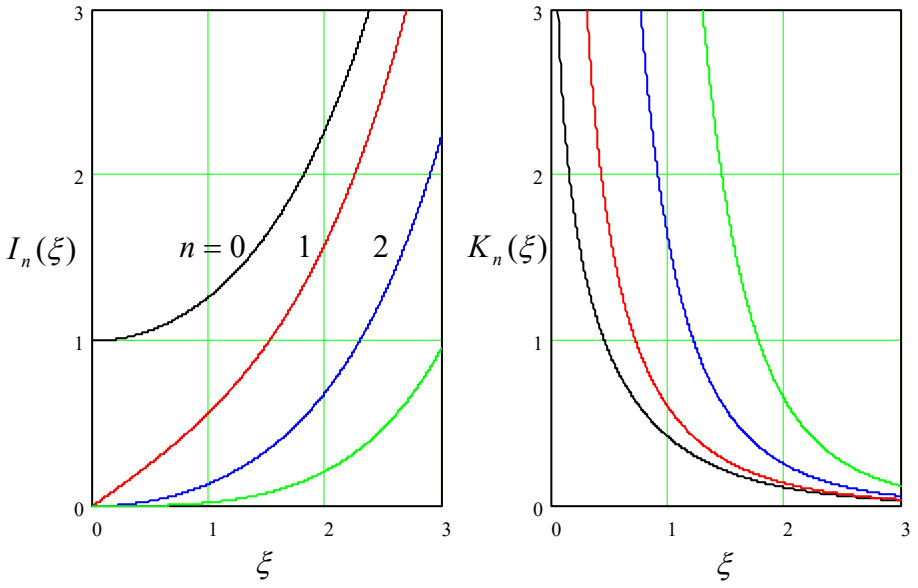


Fig. 2.22. The modified Bessel functions of the first kind (left panel) and the second kind (right panel).

This behavior is completely natural in the context of the problem shown in Fig. 21, in which the electrostatic potential may be represented as a sum of terms proportional to $I_n(\gamma\rho)$ inside the thin pipe, and of terms proportional to $K_n(\gamma\rho)$ outside it.

To complete our brief survey of the Bessel functions, let me note that all of them discussed so far may be considered as particular cases of *Bessel functions of the complex argument*, say $J_n(z)$ and $Y_n(z)$, or, alternatively, $H_n^{(1,2)}(z) \equiv J_n(z) \pm iY_n(z)$.[52] At that, the "usual" Bessel functions $J_n(\xi)$ and $Y_n(\xi)$ may be considered as the sets of values of these generalized functions on the real axis ($z = \xi$), while the modified functions as their particular case on the imaginary axis, i.e. at $z = i\xi$, also with real ξ:

$$I_\nu(\xi) = i^{-\nu} J_\nu(i\xi), \qquad K_\nu(\xi) = \frac{\pi}{2} i^{\nu+1} H_\nu^{(1)}(i\xi). \tag{2.159}$$

Moreover, this generalization of the Bessel functions to the whole complex plane z enables the use of their values along other directions on that plane, for example under angles $\pi/4 \pm \pi/2$. As a result, one arrives at the so-called *Kelvin functions*:

$$\begin{aligned} &\mathrm{ber}_\nu \xi + i\,\mathrm{bei}_\nu \xi \equiv J_\nu(\xi e^{-i\pi/4}), \\ &\mathrm{ker}_\nu\, \xi + i\,\mathrm{kei}_\nu \xi \equiv i\frac{\pi}{2} H_\nu^{(1)}(\xi e^{-i3\pi/4}), \end{aligned} \tag{2.160}$$

which are also useful for some important problems in physics and engineering. Unfortunately, I do not have time/space to discuss these problems in this course.[53]

[52] These complex functions still obey the general relations (143) and (146), with ξ replaced with z.

[53] In the QM part of this series we will run into the so-called *spherical Bessel functions* $j_n(\xi)$ and $y_n(\xi)$, which may be expressed via the Bessel functions of *semi-integer* orders. Surprisingly enough, these functions turn out to be simpler than $J_n(\xi)$ and $Y_n(\xi)$.

2. 8. Variable separation – spherical coordinates

The spherical coordinates are very important in physics, because of the (at least approximate) spherical symmetry of many physical objects – from nuclei and atoms, to water drops in clouds, to planets and stars. Let us again require each component ϕ_k of Eq. (84) to satisfy the Laplace equation. Using the full expression for the Laplace operator in spherical coordinates,[54] we get

$$\frac{1}{r^2}\frac{\partial}{\partial r}\left(r^2\frac{\partial \phi_k}{\partial r}\right)+\frac{1}{r^2\sin\theta}\frac{\partial}{\partial\theta}\left(\sin\theta\frac{\partial\phi_k}{\partial\theta}\right)+\frac{1}{r^2\sin^2\theta}\frac{\partial^2\phi_k}{\partial\varphi^2}=0\,. \tag{2.161}$$

Let us look for a solution of this equation in the following variable-separated form:

$$\phi_k=\frac{\mathcal{R}(r)}{r}\mathcal{P}(\cos\theta)\mathcal{F}(\varphi)\,, \tag{2.162}$$

Separating the variables one by one, starting from φ, just like this has been done in cylindrical coordinates, we get the following equations for the partial functions participating in this solution:

$$\frac{d^2\mathcal{R}}{dr^2}-\frac{l(l+1)}{r^2}\mathcal{R}=0\,, \tag{2.163}$$

$$\frac{d}{d\xi}\left[(1-\xi^2)\frac{d\mathcal{P}}{d\xi}\right]+\left[l(l+1)-\frac{\nu^2}{1-\xi^2}\right]\mathcal{P}=0\,, \tag{2.164}$$

$$\frac{d^2\mathcal{F}}{d\varphi^2}+\nu^2\mathcal{F}=0\,, \tag{2.165}$$

where $\xi\equiv\cos\theta$ is a new variable used in lieu of θ (so $-1\le\xi\le+1$), while ν^2 and $l(l+1)$ are the separation constants. (The reason for the selection of the latter one in this form will be clear in a minute.)

One can see that Eq. (165) is very simple, and is absolutely similar to the Eq. (107) we have got for the polar and cylindrical coordinates. Moreover, the equation for the radial functions is *simpler* than in the cylindrical coordinates. Indeed, let us look for its partial solution in the form cr^α – just as we have done with Eq. (106). Plugging this solution into Eq. (163), we immediately get the following condition on the parameter α:

$$\alpha(\alpha-1)=l(l+1)\,. \tag{2.166}$$

This quadratic equation has two roots, $\alpha=l+1$ and $\alpha=-l$, so the general solution of Eq. (163) is

$$\mathcal{R}=a_l r^{l+1}+\frac{b_l}{r^l}\,. \tag{2.167}$$

However, the general solution of Eq. (164) (called either the *general* or *associated Legendre equation*) cannot be expressed via what is usually called elementary functions.[55] Let us start its discussion from the axially-symmetric case when $\partial\phi/\partial\varphi=0$. This means $\mathcal{F}(\varphi)$ = const, and thus $\nu=0$, so Eq. (164) is reduced to the so-called *Legendre differential equation*:

[54] See, e.g., MA Eq. (10.9).

[55] Again, there is no generally accepted line between the "elementary" and "special" functions.

Legendre equation

$$\frac{d}{d\xi}\left[(1-\xi^2)\frac{d\mathcal{P}}{d\xi}\right]+l(l+1)\mathcal{P}=0\,. \tag{2.168}$$

One can readily verify that the solutions of this equation for integer values of l are specific (*Legendre*) polynomials[56] that may be described by the following *Rodrigues' formula*:

Legendre polynomials

$$\mathcal{P}_l(\xi)=\frac{1}{2^l\,l!}\frac{d^l}{d\xi^l}(\xi^2-1)^l,\quad \text{with } l=0,1,2,\ldots. \tag{2.169}$$

According to this formula, the first few Legendre polynomials are pretty simple:

$$\begin{aligned}
\mathcal{P}_0(\xi)&=1,\\
\mathcal{P}_1(\xi)&=\xi,\\
\mathcal{P}_2(\xi)&=\frac{1}{2}\left(3\xi^2-1\right),\\
\mathcal{P}_3(\xi)&=\frac{1}{2}\left(5\xi^3-3\xi\right),\\
\mathcal{P}_4(\xi)&=\frac{1}{8}\left(35\xi^4-30\xi^2+3\right),\ldots,
\end{aligned} \tag{2.170}$$

though such explicit expressions become more and more bulky as l is increased. As Fig. 23 shows, all these polynomials, which are defined on the [-1, +1] segment, end at the same point: $\mathcal{P}_l(+1) = +1$, while starting either at the same point or at the opposite point: $\mathcal{P}_l(-1) = (-1)^l$. Between these two endpoints, the l^{th} Legendre polynomial has l zeros. It is straightforward to use Eq. (169) to prove that these polynomials form a full, orthogonal set of functions, with the following normalization rule:

$$\int_{-1}^{+1}\mathcal{P}_l(\xi)\mathcal{P}_{l'}(\xi)d\xi=\frac{2}{2l+1}\delta_{ll'}, \tag{2.171}$$

so any function $f(\xi)$ defined on the segment [-1, +1] may be represented as a unique series over the polynomials.[57]

Thus, taking into account the additional division by r in Eq. (162), the general solution of any axially symmetric Laplace problem may be represented as

Variable separation in spherical coordinates (for axial symmetry)

$$\phi(r,\theta)=\sum_{l=0}^{\infty}\left(a_l r^l+\frac{b_l}{r^{l+1}}\right)\mathcal{P}_l(\cos\theta)\,. \tag{2.172}$$

Note a strong similarity between this solution and Eq. (112) for the 2D Laplace problem in the polar coordinates. However, besides the difference in the angular functions, there is also a difference (by one) in the power of the second radial function, and this difference immediately shows up in problem solutions.

[56] Just for reference: if l is not an integer, the general solution of Eq. (2.168) may be represented as a linear combination of the so-called *Legendre functions* (not polynomials!) *of the first and second kind*, $\mathcal{P}_l(\xi)$ and $\mathcal{Q}_l(\xi)$.

[57] This is why, at least for the purposes of this course, there is no good reason for pursuing (more complicated) solutions to Eq. (168) for non-integer values of l, mentioned in the previous footnote.

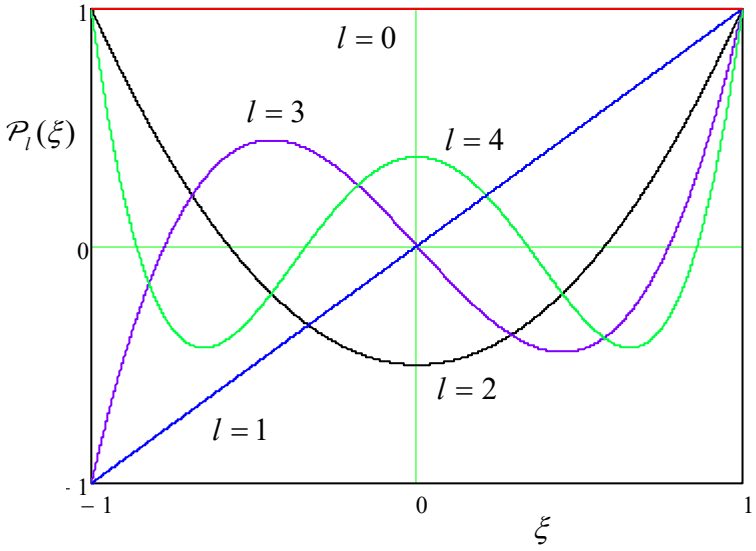


Fig. 2.23. A few lowest Legendre polynomials $\mathcal{P}_l(\xi)$.

Indeed, let us solve a problem similar to that shown in Fig. 15: find the electric field around a conducting sphere of radius R, placed into an initially uniform external field $\mathbf{E}_0$ (whose direction I will now take for the z-axis) – see Fig. 24a.

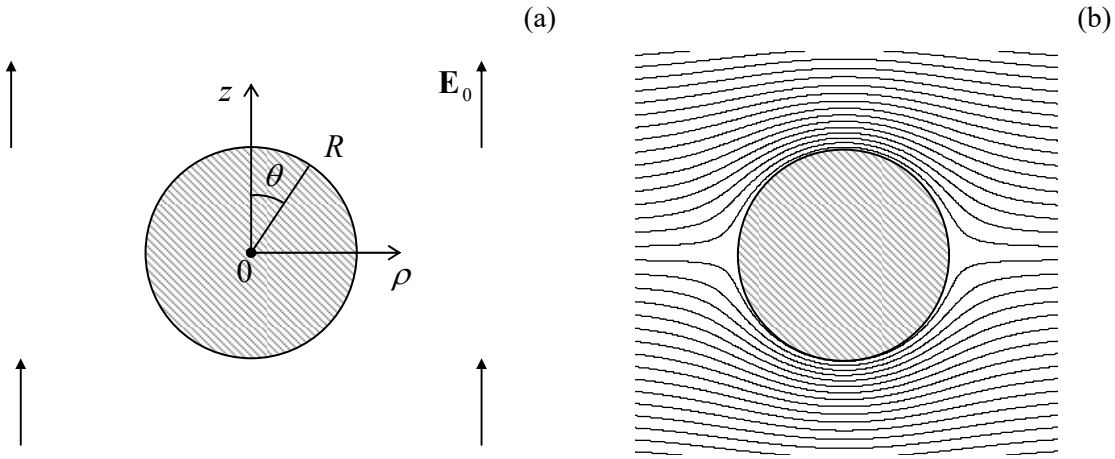


Fig. 2.24. Conducting sphere in a uniform electric field: (a) the problem's geometry, and (b) the equipotential surface pattern given by Eq. (176). The pattern is qualitatively similar but quantitatively different from that for the conducting cylinder in a perpendicular field – cf. Fig. 15.

If we select the arbitrary constant in the electrostatic potential so that $\phi|_{z=0} = 0$, then in Eq. (172) we should take $a_0 = b_0 = 0$. Now, just as has been argued for the cylindrical case, at $r >> R$ the potential should approach that of the uniform field:

$$\phi \to -E_0 z = -E_0 r \cos\theta , \tag{2.173}$$

so in Eq. (172), only one of the coefficients a_l survives: $a_l = -E_0\delta_{l,1}$. As a result, from the boundary condition on the surface, $\phi(R,\theta) = 0$, we get the following equation for the coefficients b_l:

$$\left(-E_0 R + \frac{b_1}{R^2}\right)\cos\theta + \sum_{l\geq 2}\frac{b_l}{R^{l+1}}\mathcal{P}_l(\cos\theta) = 0 . \tag{2.174}$$

Now repeating the argumentation that led to Eq. (117), we may conclude that Eq. (174) is satisfied if

$$b_l = E_0 R^3 \delta_{l,1}, \tag{2.175}$$

so, finally, Eq. (172) is reduced to

$$\phi = -E_0\left(r - \frac{R^3}{r^2}\right)\cos\theta . \tag{2.176}$$

This distribution, shown in Fig. 24b, is very similar to Eq. (117) for the cylindrical case (cf. Fig. 15b, with the account for a different plot orientation), but with a different power of the radius in the second term. This difference leads to a quantitatively different distribution of the surface electric field:

$$E_n = -\frac{\partial\phi}{\partial r}\Big|_{r=R} = 3E_0\cos\theta , \tag{2.177}$$

so its maximal value is a factor of 3 (rather than 2) larger than the external field.

Now let me briefly (mostly just for the reader's reference) mention the Laplace equation solutions in the general case – with no axial symmetry. If the conductor-free space surrounds the origin from all sides, the solutions to Eq. (165) have to be 2π-periodic, and hence $\nu = n = 0, \pm1, \pm2,\ldots$ Mathematics says that Eq. (164) with integer $\nu = n$ and a fixed integer l has a solution only for a limited range of n:[58]

$$-l \le n \le +l . \tag{2.178}$$

These solutions are called a*ssociated Legendre functions* (generally, they are *not* polynomials). For $n \ge 0$, these functions may be defined via the Legendre polynomials, using the following formula:[59]

$$\mathcal{P}_l^n(\xi) = (-1)^n (1-\xi^2)^{n/2}\frac{d^n}{d\xi^n}\mathcal{P}_l(\xi) . \tag{2.179}$$

On the segment $\xi \in [-1, +1]$, each set of the associated Legendre functions with a fixed index n and non-negative values of l form a full, orthogonal set, with the normalization relation,

$$\int_{-1}^{+1}\mathcal{P}_l^n(\xi)\mathcal{P}_{l'}^n(\xi)d\xi = \frac{2}{2l+1}\frac{(l+n)!}{(l-n)!}\delta_{ll'}, \tag{2.180}$$

that is evidently a generalization of Eq. (171).

Since these relations may seem a bit intimidating, let me write down explicit expressions for a few $\mathcal{P}_l^n(\cos\theta)$ with the three lowest values of l and $n \ge 0$, which are most important for applications.

$$l = 0: \qquad \mathcal{P}_0^0(\cos\theta) = 1; \tag{2.181}$$

[58] In quantum mechanics, the letter n is typically reserved for the "principal quantum number", while the azimuthal functions are numbered by index m. However, here I will keep using n as their index because, for this course's purposes, this seems more logical, in view of the similarity of the spherical and cylindrical functions.

[59] Note that some texts use different choices for the front factor (called the *Condon-Shortley phase*) in the functions $\mathcal{P}_l^m$, which do not affect the final results for the spherical harmonics Y_l^m.

$$l=1: \quad \begin{cases} \mathcal{P}_1^0(\cos\theta)=\cos\theta, \\ \mathcal{P}_1^1(\cos\theta)=-\sin\theta; \end{cases} \tag{2.182}$$

$$l=2: \begin{cases} \mathcal{P}_2^0(\cos\theta)=\left(3\cos^2\theta-1\right)/2, \\ \mathcal{P}_2^1(\cos\theta)=-2\sin\theta\cos\theta, \\ \mathcal{P}_2^2(\cos\theta)=-3\cos^2\theta. \end{cases} \tag{2.183}$$

The reader should agree there is not much to fear in these functions – they are just certain sums of products of $\cos\theta \equiv \xi$ and $\sin\theta \equiv (1-\xi^2)^{1/2}$. Fig. 25 shows the plots of a few lowest functions $\mathcal{P}_l^n(\xi)$.

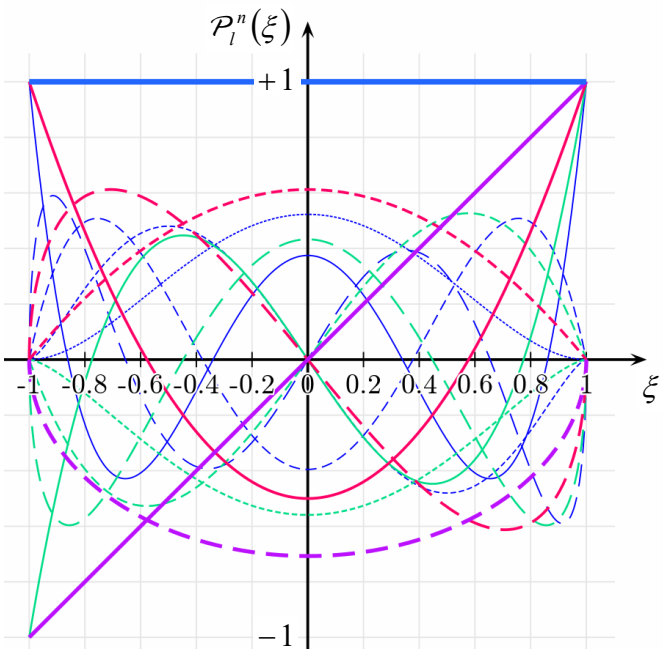


Fig. 2.25. A few lowest associated Legendre functions. (Adapted from an original by *Geek3*, available at https://en.wikipedia.org/wiki/Associated_Legendre_polynomials, under the GNU Free Documentation License.)

Using the associated Legendre functions, the general solution (162) to the Laplace equation in the spherical coordinates may be expressed as

Variable separation in spherical coordinates (general case)

$$\phi(r,\theta,\varphi)=\sum_{l=0}^{\infty}\left(a_l r^l+\frac{b_l}{r^{l+1}}\right)\sum_{n=0}^{l}\mathcal{P}_l^n(\cos\theta)\mathcal{T}_n(\varphi), \qquad \mathcal{T}_n(\varphi)=c_n\cos n\varphi+s_n\sin n\varphi\,. \tag{2.184}$$

Since the difference between the angles θ and φ is somewhat artificial, physicists prefer to think not in terms of the functions $\mathcal{P}$ and $\mathcal{T}$ in separation, but directly about their products that participate in this solution.[60]

[60] In quantum mechanics, it is more convenient to use a slightly different alternative set of basic functions of the same problem, namely the following complex functions called the *spherical harmonics*:

$$Y_l^n(\theta,\varphi)\equiv\left[\frac{2l+1}{4\pi}\frac{(l-n)!}{(l+n)!}\right]^{1/2}\mathcal{P}_l^n(\cos\theta)e^{in\varphi},$$

which are defined for both positive and negative n (within the limits $-l \le n \le +l$) – see, e.g., QM Secs. 3.6 and 5.6. (Note again that in that field, our index n is traditionally denoted as m, and called the *magnetic quantum number*.)

As a rare exception for my courses, to save time I will skip giving an example of using the associated Legendre functions in electrostatics, because quite a few examples of these functions' applications will be given in the quantum mechanics part of this series.

2.9. Charge images

So far, we have discussed various methods of solution of the *Laplace* boundary problem (35). Let us now move on to the discussion of its generalization, the *Poisson* equation (1.41). We need it when besides conductors, we also have stand-alone charges with a known spatial distribution $\rho(\mathbf{r})$. (Its discussion will also allow us, better equipped, to revisit the Laplace problem in the next section.)

Let us start with a somewhat limited, but very useful *charge image* (or "image charge") *method*. Consider a very simple problem: a single point charge near a conducting half-space – see Fig. 26.

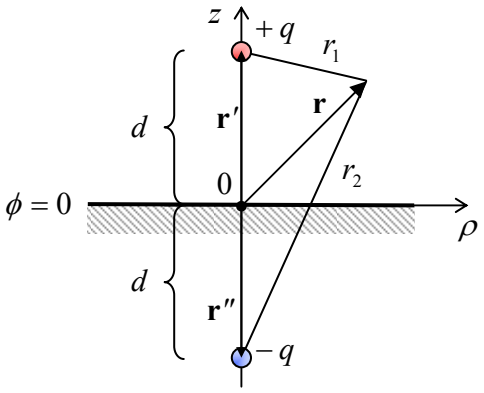


Fig. 2.26. The simplest problem readily solvable by the charge image method. The points' colors are used, as before, to denote the charges of the original (red) and opposite (blue) sign.

Let us prove that its solution, above the conductor's surface ($z \geq 0$), may be represented as:

$$\phi(\mathbf{r}) = \frac{1}{4\pi\varepsilon_0}\left(\frac{q}{r_1} - \frac{q}{r_2}\right) \equiv \frac{q}{4\pi\varepsilon_0}\left(\frac{1}{|\mathbf{r}-\mathbf{r}'|} - \frac{1}{|\mathbf{r}-\mathbf{r}''|}\right), \tag{2.185}$$

or in a more explicit form, using the cylindrical coordinates shown in Fig. 26:

$$\phi(\mathbf{r}) = \frac{q}{4\pi\varepsilon_0}\left\{\frac{1}{\left[\rho^2 + (z-d)^2\right]^{1/2}} - \frac{1}{\left[\rho^2 + (z+d)^2\right]^{1/2}}\right\}, \tag{2.186}$$

where ρ is the distance of the field observation point $\mathbf{r}$ from the "vertical" line on which the charge is located. Indeed, this solution satisfies both the boundary condition $\phi = 0$ at the surface of the conductor ($z = 0$), and the Poisson equation (1.41), with the single δ-functional source at point $\mathbf{r}' = \{0, 0, +d\}$ on its right-hand side, because the second singularity of the solution, at point $\mathbf{r}'' = \{0, 0, -d\}$, is outside the region of the solution's validity ($z \geq 0$). Physically, this solution may be interpreted as the sum of the fields of the actual charge ($+q$) at point $\mathbf{r}'$, and an equal but opposite charge ($-q$) at the "mirror image" point $\mathbf{r}''$ (Fig. 26). This is the basic idea of the charge image method.

Before moving on to more complex problems, let us discuss the situation shown in Fig. 26 in a little bit more detail, due to its fundamental importance. First, we can use Eqs. (3) and (186) to calculate the surface charge density:

$$\sigma = -\varepsilon_0 \frac{\partial \phi}{\partial z}\Big|_{z=0} = -\frac{q}{4\pi}\frac{\partial}{\partial z}\left\{\frac{1}{\left[\rho^2+(z-d)^2\right]^{1/2}} - \frac{1}{\left[\rho^2+(z+d)^2\right]^{1/2}}\right\}_{z=0} = -\frac{q}{4\pi}\frac{2d}{\left(\rho^2+d^2\right)^{3/2}}. \quad (2.187)$$

From this, the total surface charge is

$$Q = \int_S \sigma d^2 r = 2\pi \int_0^\infty \sigma(\rho)\rho d\rho = -\frac{q}{2}\int_0^\infty \frac{2d}{\left(\rho^2+d^2\right)^{3/2}}\rho d\rho. \quad (2.188)$$

This integral may be easily worked out using the substitution $\xi \equiv \rho^2/d^2$ (giving $d\xi = 2\rho d\rho/d^2$):

$$Q = -\frac{q}{2}\int_0^\infty \frac{d\xi}{(\xi+1)^{3/2}} = -q. \quad (2.189)$$

This result is very natural: the conductor brings as much surface charge from its interior to the surface as necessary to fully compensate for the initial charge ($+q$) and hence kill the electric field at large distances as efficiently as possible, hence reducing the total electrostatic energy (1.65) to the lowest possible value.

For a better feeling of this *polarization charge* of the surface, let us take our calculations to the extreme – to the q equal to one elementary change e, and place a particle with this charge (for example, a proton) at a macroscopic distance – say 1 m – from the conductor's surface. Then, according to Eq. (189), the total polarization charge of the surface equals that of an electron, and according to Eq. (187), its spatial extent is of the order of $d^2 = 1$ m^2. This means that if we consider a much smaller part of the surface, $\Delta A << d^2$, its polarization charge magnitude $\Delta Q = \sigma \Delta A$ is much *less than one electron*! For example, Eq. (187) shows that the polarization charge of quite a macroscopic area $\Delta A = 1$ cm^2 right under the initial charge ($\rho = 0$) is $e\Delta A/2\pi d^2 \approx 1.6\times10^{-5}\ e$. Can this be true, or our theory is somehow limited to the charges q much larger than e? (After all, the theory is substantially based on the approximate macroscopic model (1); maybe it is the culprit?)

Surprisingly enough, the answer to this question has become clear (at least to some physicists :-) only as late as in the mid-1980s when several experiments demonstrated, and theorists accepted (some of them rather grudgingly) that the usual polarization charge formulas are valid for elementary charges as well, i.e., such the polarization charge ΔQ of a macroscopic surface area may differ from a multiple of e. The underlying reason for this paradox is the physical nature of the polarization charge of a conductor's surface: as was discussed in Sec. 1, it is due not to new charged particles brought into the conductor (such charge would be in fact a multiple of e), but to a small *shift* of the free charges of a conductor by a very small distance from their equilibrium positions that they had in the absence of the external field induced by charge q. This shift is not quantized, at least on the scale relevant to our problem, and hence neither is ΔQ.

This understanding has paved the way for the invention and experimental demonstration of several new devices including so-called *single-electron transistors*,[61] which may be used, in particular, for ultrasensitive measurement of polarization charges as small as $\sim10^{-6}\ e$. Another important class of single-electron devices is the dc and ac current standards based on the fundamental relation $I = -ef$,

[61] Actually, this term (for which the author of these notes may be blamed :-) is misleading: the operation of the "single-electron transistor" is based on the interplay of discrete charges (multiples of e) transferred between conductors, and *sub*-single-electron polarization charges – see, e.g., K. Likharev, *Proc. IEEE* **87**, 606 (1999).

where I is the dc current carried by electrons transferred with the frequency f. The experimentally achieved[62] relative accuracy of such standards is of the order of 10^{-7}, and is not too far from that provided by the competing approach based on a combination of the Josephson effect and the quantum Hall effect.[63]

Second, let us find the potential energy U of the charge-to-surface interaction. For that, we may use the value of the electrostatic potential (185) at the point of the charge itself ($\mathbf{r} = \mathbf{r}'$), of course ignoring the infinite potential created by the real charge, so the remaining potential is that of the image charge

$$\phi_{\text{image}}(\mathbf{r}') = -\frac{1}{4\pi\varepsilon_0}\frac{q}{2d}. \tag{2.190}$$

Looking at the electrostatic potential's definition given by Eq. (1.31), it may be tempting to immediately write $U = q\phi_{\text{image}} = -(1/4\pi\varepsilon_0)(q^2/2d)$ [WRONG!], but this would be incorrect. The reason is that the potential ϕ_{image} is not independent of q, but is actually induced by this charge. This is why the correct approach is to calculate U from Eq. (1.61), with just one term:

$$U = \frac{1}{2}q\phi_{\text{image}} = -\frac{1}{4\pi\varepsilon_0}\frac{q^2}{4d}, \tag{2.191}$$

giving twice lower energy than the wrong result cited above. To double-check Eq. (191), and also get a better feeling of the factor ½ that distinguishes it from the wrong guess, we can calculate U as the integral of the force exerted on the charge by the conductor's surface charge (i.e., in our formalism, by the image charge):

$$U = -\int_{\infty}^{d} F(z)dz = \frac{1}{4\pi\varepsilon_0}\int_{\infty}^{d}\frac{q^2}{(2z)^2}dz = -\frac{1}{4\pi\varepsilon_0}\frac{q^2}{4d}. \tag{2.192}$$

This calculation clearly accounts for the gradual build-up of the force F, as the real charge is being brought from afar (where we have opted for $U = 0$) toward the surface.

This result has several important applications. For example, let us plot the electrostatic energy U of an electron, i.e. a particle with charge $q = -e$, near a metallic surface, as a function of d. For that, we may use Eq. (191) until our macroscopic model (1) becomes invalid, and U transitions to some negative constant value ($-\psi$) inside the conductor – see Fig. 27a. Since our calculation was for an electron with zero potential energy at infinity, at relatively low temperatures, $k_B T << \psi$, electrons in metals may occupy only the states with energies below $-\psi$ (the so-called *Fermi level*[64]). The positive constant ψ is called the *workfunction* because it describes the smallest work needed to remove the electron from a metal. As was discussed in Sec. 1, in good metals the electric field screening takes place at interatomic distances $a_0 \sim 10^{-10}$ m. Plugging $d = 1\times10^{-10}$ m and $q = -e \approx -1.6\times10^{-19}$ C into Eq. (191), we get $\psi \approx 6\times10^{-19}$ J ≈ 3.5 eV. This crude estimate is in surprisingly good agreement with the experimental values of the workfunction, ranging between 4 and 5 eV for most metals.[65]

[62] See, e.g., M. Keller *et al.*, *Appl. Phys. Lett.* **69**, 1804 (1996) ; F. Stein *et al.*, Metrologia **54**, 1 (2017).
[63] J. Brun-Pickard *et al.*, *Phys. Rev. X* **6**, 041051 (2016).
[64] More discussion of these states may be found in SM Secs. 3.3 and 6.3.
[65] More discussion of the workfunction, and its effect on the electrons' kinetics, is given in SM Sec. 6.3.

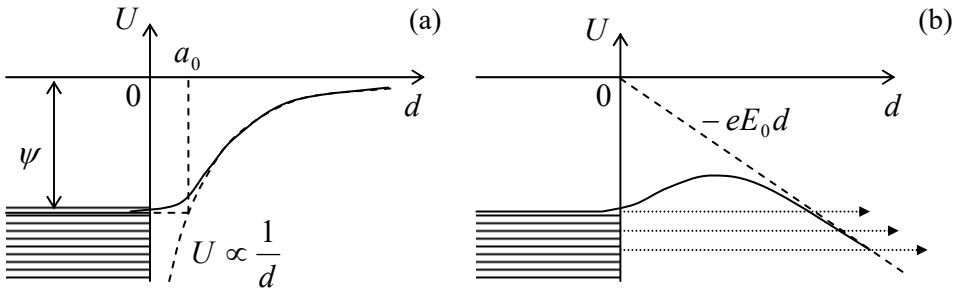


Fig. 2.27. (a) The origin of the workfunction, and (b) the field emission of electrons (schematically).

Next, let us consider the effect of an additional uniform external electric field $\mathbf{E}_0$ applied normally to a metallic surface, on this potential profile. For that, we may the potential energy that the field gives to the electron at distance d from the surface, $U_{\text{ext}} = -eE_0 d$, to that created by the image charge. (As we know from Eq. (1.53), since the field $\mathbf{E}_0$ is independent of the electron's position, its recalculation into the potential energy does not require the coefficient ½.) As a result, the potential energy of an electron near the surface becomes

$$U(d) = -eE_0 d - \frac{1}{4\pi\varepsilon_0}\frac{e^2}{4d}, \qquad \text{for } d >> a_0, \tag{2.193}$$

with a similar crossover to $U = -\psi$ inside the conductor – see Fig. 27b. One can see that at the appropriate sign, and a sufficient magnitude of the applied field, it lowers the potential barrier that prevents electrons from leaving the conductor. At $E_0 \sim \psi/a_0$ (for metals, $\sim 10^{10}$ V/m), this suppression becomes so strong that electrons with energies at, and just below the Fermi level start quantum-mechanical tunneling through the remaining thin barrier. This is the *field electron emission* (or just "field emission") effect, which is used in vacuum electronics to provide efficient cathodes that do not require heating to high temperatures.[66]

Returning to the basic electrostatics, let us find some other conductor geometries where the method of charge images may be effectively applied. First, let us consider a right-angle corner (Fig. 28a). Reflecting the initial charge in the vertical plane, we get the image shown in the top left corner of that panel. This image makes the boundary condition ϕ = const satisfied on the vertical surface of the corner. However, for the same to be true on the horizontal surface, we have to reflect *both* the initial charge *and* the image charge in the horizontal plane, flipping their signs. The final configuration of four charges, shown in Fig. 28a, satisfies all boundary conditions. The resulting potential distribution may be readily written as an evident generalization of Eq. (185). From it, the electric field and electric charge distributions, and the potential energy and forces acting on the charge may be calculated exactly as above – an easy exercise left for the reader.

Next, consider a corner with the angle $\pi/4$ (Fig. 28b). Here we need to repeat the reflection operation not two but four times before we arrive at the final pattern of eight positive and negative charges. (Any attempt to continue this process would lead to overlap with the already existing charges.)

[66] The practical use of such "cold" cathodes is affected by the fact that, as it follows from our discussion in Sec. 4, any nanoscale irregularity of a conducting surface (a protrusion, an atomic cluster, or even a single "adatom" stuck to it) may cause a strong increase of the local field well above the applied uniform field E_0, making the electron emission reproducibility and stability in time significant challenges. In addition, the impact-ionization effects may lead to avalanche-type electric breakdown at dc fields as low as $\sim 3\times 10^6$ V/m.

This reasoning may be readily extended to corners of angles $\beta = \pi/n$, with any integer n, which require $2n$ charges (including the initial one) to satisfy all the boundary conditions.

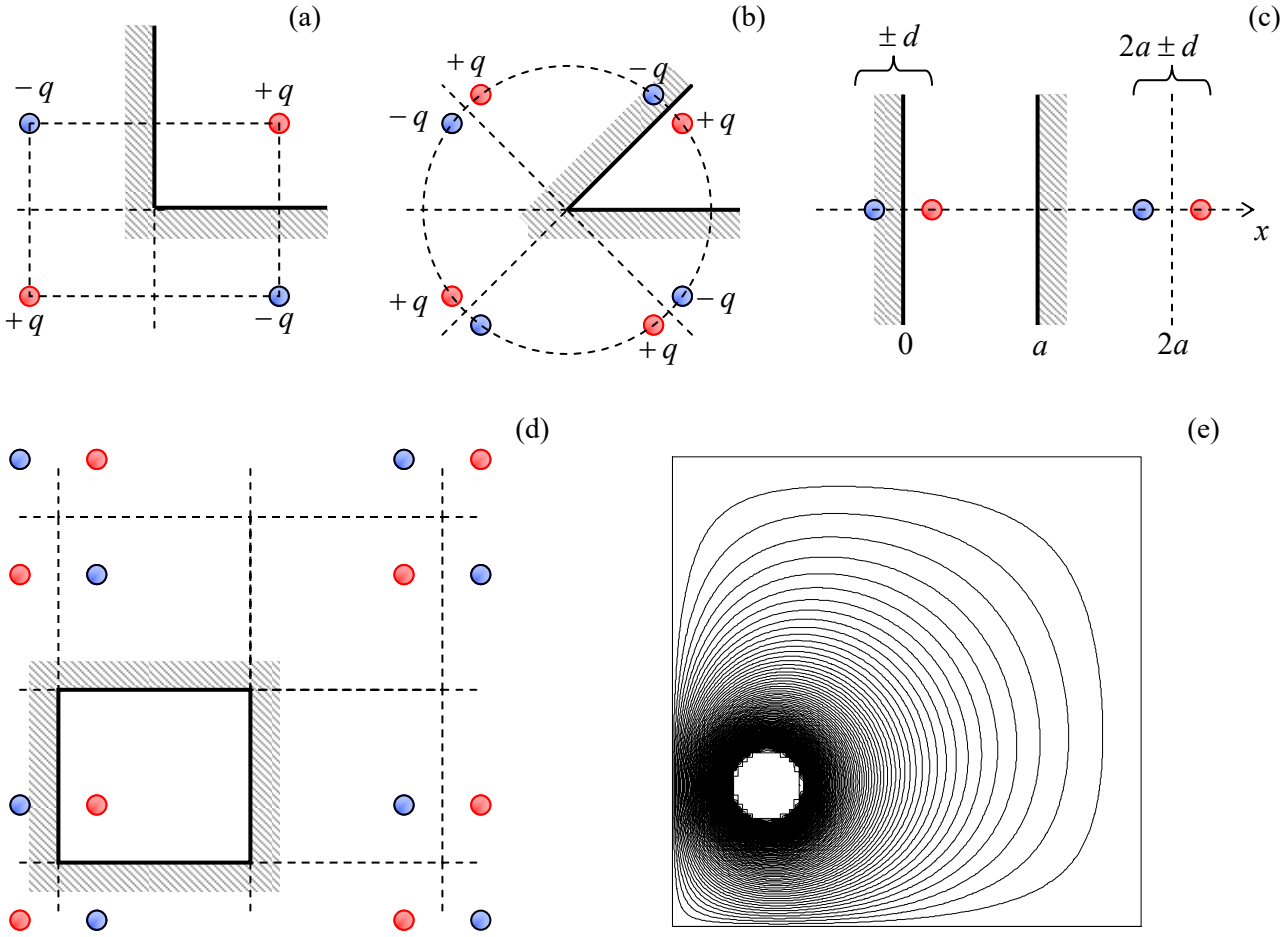


Fig. 2.28. The charge images for (a, b) the corners with angles $\pi/2$ and $\pi/4$, (c) a plane capacitor, and (d) a rectangular box; (e) typical equipotential surfaces for the last system.

Some configurations require an infinite number of images but are still tractable. The most important of them is a system of two parallel conducting surfaces, i.e. an unbiased plane capacitor of infinite area (Fig. 28c). Here the repeated reflection leads to an infinite system of charges $\pm q$ at points

$$x_j^{\pm} = 2aj \pm d\,, \tag{2.194}$$

where d (with $0 < d < a$) is the position of the initial charge, and j is an arbitrary integer. The resulting infinite sum for the potential of the real charge q, created by the field of its images,

$$\phi(d) = \frac{1}{4\pi\varepsilon_0}\left[-\frac{q}{2d} + \sum_{j\neq 0}\sum_{\pm}\frac{\pm q}{\left|d - x_j^{\pm}\right|}\right] \equiv -\frac{q}{4\pi\varepsilon_0}\left\{\frac{1}{2d} + \frac{d^2}{a^3}\sum_{j=1}^{\infty}\frac{1}{j\left[j^2 - (d/a)^2\right]}\right\}, \tag{2.195}$$

is converging (in its last form) very fast. For example, the exact value, $\phi(a/2) = -2\ln 2$ $(q/4\pi\varepsilon_0 a)$, differs by less than 5% from the approximation using just the first term of the sum.

The same method may be applied to 2D (cylindrical) and 3D rectangular conducting boxes that require, respectively, 2D or 3D infinite rectangular lattices of images; for example in a 3D box with sides a, b, and c, charges $\pm q$ are located at points (Fig. 28d)

$$\mathbf{r}_{jkl}^{\pm} = 2ja + 2kb + 2lc \pm \mathbf{r}', \tag{2.196}$$

where $\mathbf{r}'$ is the location of the initial (real) charge, and j, k, and l are arbitrary integers. Figure 28e shows a typical result of the summation of the potentials of this charge set, including the real one, in a 2D box (within the plane of the real charge). One can see that the equipotential surfaces, concentric near the charge, are naturally leaning along the conducting walls of the box, which have to be equipotential.

Even more surprisingly, the image charge method works very efficiently not only for rectilinear geometries but also for spherical ones. Indeed, let us consider a point charge q at distance d from the center of a conducting, grounded sphere of radius R (Fig. 29a), and try to satisfy the boundary condition $\phi = 0$ for the electrostatic potential on the sphere's surface using an imaginary charge q' located at some point beyond the surface, i.e. inside the sphere.

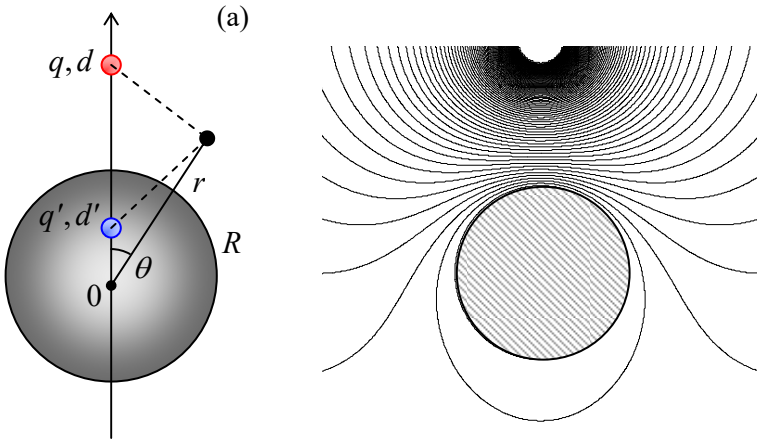


Fig. 2.29. Method of charge images for a conducting sphere: (a) the idea, and (b) the resulting potential distribution in the central plane containing the charge, for the particular case $d = 2R$.

From the problem's symmetry, it is clear that the point should be at the line passing through the real charge and the sphere's center, at some distance d' from the center. Then the total potential created by the two charges at an arbitrary point of free space, i.e. at $r \geq R$ (Fig. 29a) is

$$\phi(r,\theta) = \frac{1}{4\pi\varepsilon_0}\left[\frac{q}{\left(r^2 + d^2 - 2rd\cos\theta\right)^{1/2}} + \frac{q'}{\left(r^2 + d'^2 - 2rd'\cos\theta\right)^{1/2}}\right]. \tag{2.197}$$

This expression shows that we can make the two involved fractions equal and opposite at all points on the sphere's surface (i.e. for any θ at $r = R$) if we take[67]

$$d' = \frac{R^2}{d}, \qquad q' = -\frac{R}{d}q. \tag{2.198}$$

Since the solution of any Poisson boundary problem is unique, Eqs. (197) and (198) give us the final solution for this problem. Fig. 29b shows a typical equipotential pattern following from this solution. It may be surprising how formulas that simple may describe such an elaborate field distribution.

[67] In geometry, such points with $dd' = R^2$, are referred to as the result of mutual *inversion* in a sphere of radius R.

Now we can calculate the total charge Q on the grounded sphere's surface, induced by the external charge q. We could do this, as we have done for the conducting plane, by the brute-force integration of the surface charge density $\sigma = -\varepsilon_0 \partial\phi/\partial r|_{r=R}$. It is more elegant, however, to use the following Gauss law argument. Equality (197) is valid (at $r \geq R$) regardless of whether we are dealing with our real problem (charge q and the conducting sphere) or with the equivalent charge configuration – with the point charges q and q', but no sphere at all. Hence, according to Eq. (1.16), the Gaussian integral over a surface with radius $r = R + 0$, and the total charge inside the sphere should be also the same. Hence we immediately get

$$Q = q' = -\frac{R}{d} q \,. \tag{2.199}$$

A similar argumentation may be used to calculate the charge-to-sphere interaction force:

$$F = qE_{\text{image}}(d) = q\frac{q'}{4\pi\varepsilon_0 (d-d')^2} = -\frac{q^2}{4\pi\varepsilon_0}\frac{R}{d}\frac{1}{(d-R^2/d)^2} = -\frac{q^2}{4\pi\varepsilon_0}\frac{Rd}{(d^2-R^2)^2}\,. \tag{2.200}$$

(Note that this expression is legitimate only at $d > R$.) At large distances, $d >> R$, this attractive force decreases as $1/d^3$. This unusual dependence arises because, as Eq. (199) specifies, the induced charge of the sphere, responsible for the force, is not constant but decreases as $1/d$. In the next chapter, we will see that such force is also typical for the interaction between a point charge and a *dipole*.

All previous formulas were for a sphere that is grounded to keep its potential equal to zero. But what if we keep the sphere galvanically insulated, so its *net charge* is fixed, for example, equals zero? Instead of solving this problem from scratch, let us use (again!) the almighty linear superposition principle. For that, we may add to the previous problem an additional charge, equal to $-Q = -q'$, to the sphere, and argue that this addition gives, at all points, an additional, spherically symmetric potential that does not depend on the potential induced by the external charge q, and was calculated in Sec. 1.2 – see Eq. (1.19). For the interaction force, such addition yields

$$F = \frac{qq'}{4\pi\varepsilon_0 (d-d')^2} + \frac{qq'}{4\pi\varepsilon_0 d^2} = -\frac{q^2}{4\pi\varepsilon_0}\left[\frac{Rd}{(d^2-R^2)^2} - \frac{R}{d^3}\right]. \tag{2.201}$$

At large distances, the two terms proportional to $1/d^3$ cancel each other, giving $F \propto 1/d^5$, so the potential energy of such interaction behaves as $U \propto 1/d^4$. Such a rapid force decay is due to the fact that the field of the uncharged sphere is equivalent to that of two (equal and opposite) induced charges $+q'$ and $-q'$, and the distance between them ($d - d' = d - R^2/d$) tends to zero at $d \to \infty$.

2.10. Green's functions

I have spent so much time/space discussing potential distributions created by a single point charge in various conductor geometries because for any of the geometries, the generalization of these results to the arbitrary distribution $\rho(\mathbf{r})$ of free charges is straightforward. Namely, if a single charge q, located at some point $\mathbf{r}'$, creates at point $\mathbf{r}$ the electrostatic potential

$$\phi(\mathbf{r}) = \frac{1}{4\pi\varepsilon_0} qG(\mathbf{r},\mathbf{r}')\,, \tag{2.202}$$

then, due to the linear superposition principle, an arbitrary charge distribution creates the potential

$$\phi(\mathbf{r}) = \frac{1}{4\pi\varepsilon_0}\sum_j q_j G(\mathbf{r},\mathbf{r}_j) = \frac{1}{4\pi\varepsilon_0}\int \rho(\mathbf{r}')G(\mathbf{r},\mathbf{r}')d^3r'. \quad (2.203)$$

Spatial Green's function

The function $G(\mathbf{r}, \mathbf{r}')$ is called the (spatial) *Green's function* – the notion very fruitful and hence popular in all fields of physics.[68] Evidently, as Eq. (1.35) shows, in the unlimited free space

$$G(\mathbf{r},\mathbf{r}') = \frac{1}{|\mathbf{r}-\mathbf{r}'|}, \quad (2.204)$$

i.e. the Green's function depends only on one scalar argument – the distance between the field-observation point $\mathbf{r}$ and the field-source (charge) point $\mathbf{r}'$. However, as soon as there are conductors around, the situation changes. In this course, I will only discuss Green's functions defined to vanish as soon as the radius-vector $\mathbf{r}$ points to the surface (S) of any conductor:[69]

$$G(\mathbf{r},\mathbf{r}')\Big|_{\mathbf{r}\in S} = 0. \quad (2.205)$$

With this definition, it is straightforward to deduce the Green's functions for the solutions of the last section's problems in which conductors were grounded, i.e. had potential $\phi = 0$. For example, for a semi-space $z \geq 0$ limited by a grounded conducting plane $z = 0$ (Fig. 26), Eq. (185) yields

$$G = \frac{1}{|\mathbf{r}-\mathbf{r}'|} - \frac{1}{|\mathbf{r}-\mathbf{r}''|}, \qquad \text{with } \boldsymbol{\rho}'' = \boldsymbol{\rho}' \text{ and } z'' = -z', \quad (2.206)$$

where $\boldsymbol{\rho}$ is the 2D radius vector. We see that in the presence of conductors (and, as we will see later, any other polarizable media), Green's function may depend not only on the difference $\mathbf{r} - \mathbf{r}'$, but on each of these two arguments in a specific way.

So far, this is just re-naming our old results. The really non-trivial result of the Green's function formalism in electrostatics is that, somewhat counter-intuitively, the knowledge of this function for a system with *grounded* conductors (Fig. 30a) enables the calculation of the field created by *voltage-biased* conductors (Fig. 30b), with the same geometry. To show this, let us use the so-called *Green's theorem* of the vector calculus.[70] The theorem states that for any two scalar, differentiable functions $f(\mathbf{r})$ and $g(\mathbf{r})$, and any volume V,

$$\int_V \left(f\nabla^2 g - g\nabla^2 f\right)d^3r = \oint_S \left(f\nabla g - g\nabla f\right)_n d^2r, \quad (2.207)$$

where S is the surface limiting the volume. Applying the theorem to the electrostatic potential $\phi(\mathbf{r})$ and the Green's function G (also considered as a function of $\mathbf{r}$), let us use the Poisson equation (1.41) to replace $\nabla^2\phi$ with $(-\rho/\varepsilon_0)$, and notice that G, considered as a function of $\mathbf{r}$, obeys the Poisson equation with the δ-functional source:

$$\nabla^2 G(\mathbf{r},\mathbf{r}') = -4\pi\delta(\mathbf{r}-\mathbf{r}'). \quad (2.208)$$

[68] See, e.g., CM Sec. 5.1, QM Secs. 2.2 and 7.4, and SM Sec. 5.5. Note that the numerical coefficient in Eq. (202) (and hence all resulting formulas) is the matter of convention; this choice does not affect the final results.

[69] G so defined is sometimes called the *Dirichlet function*.

[70] See, e.g., MA Eq. (12.3). Actually, this theorem is a ready corollary of the better-known divergence ("Gauss") theorem, MA Eq. (12.2).

(Indeed, according to its definition (202), this function may be formally considered as the field of a point charge $q = 4\pi\varepsilon_0$.) Now swapping the notation of the radius-vectors, $\mathbf{r} \leftrightarrow \mathbf{r}'$, and using the Green's function symmetry, $G(\mathbf{r}, \mathbf{r}') = G(\mathbf{r}', \mathbf{r})$,[71] we get

$$-4\pi\phi(\mathbf{r}) - \int_V \left(-\frac{\rho(\mathbf{r}')}{\varepsilon_0}\right) G(\mathbf{r},\mathbf{r}')d^3r' = \oint_S \left[\phi(\mathbf{r}')\frac{\partial G(\mathbf{r},\mathbf{r}')}{\partial n'} - G(\mathbf{r},\mathbf{r}')\frac{\partial \phi(\mathbf{r}')}{\partial n'}\right] d^2r' . \qquad (2.209)$$

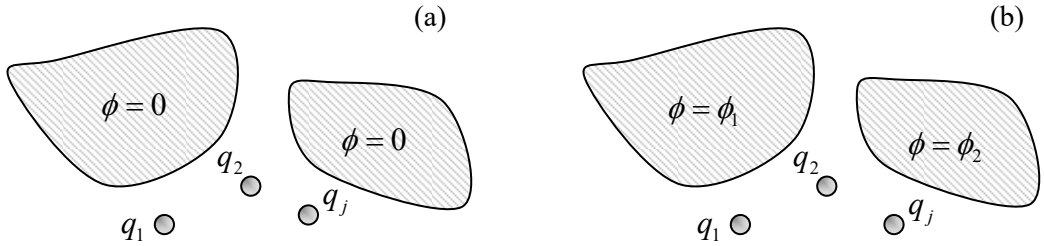


Fig. 2.30. Green's function method allows the solution of a simpler boundary problem (a) to be used for the solution of a more complex problem (b), for the same conductor geometry.

Let us apply this relation to the volume V of *free space* between the conductors, and the boundary S drawn immediately outside of their surfaces. In this case, by its definition, Green's function $G(\mathbf{r}, \mathbf{r}')$ vanishes at the conductor surface, i.e. at $\mathbf{r} \in S$ – see Eq. (205). Now changing the sign of $\partial n'$ (so it would be the outer normal for *conductors*, rather than free space volume V), dividing all terms by 4π, and partitioning the total surface S into the parts (numbered by index j) corresponding to different conductors (possibly, kept at different potentials ϕ_k), we finally arrive at the famous result:[72]

Potential via Green's function

$$\phi(\mathbf{r}) = \frac{1}{4\pi\varepsilon_0}\int_V \rho(\mathbf{r}')G(\mathbf{r},\mathbf{r}')d^3r' + \frac{1}{4\pi}\sum_k \phi_k \oint_{S_k} \frac{\partial G(\mathbf{r},\mathbf{r}')}{\partial n'} d^2r' . \qquad (2.210)$$

While the first term on the right-hand side of this relation is a direct and evident expression of the superposition principle, given by Eq. (203), the second term is highly non-trivial: it describes the effect of conductors with *arbitrary* potentials ϕ_k (Fig. 30b), using the Green's function calculated for the similar system with *grounded* conductors, i.e. with all $\phi_k = 0$ (Fig. 30a). Let me emphasize that since our volume V excludes conductors, the first term on the right-hand side of Eq. (210) includes only the stand-alone charges in the system (in Fig. 30, marked q_1, q_2, etc.), but not the surface charges of the conductors – which are taken into account, indirectly, by the second term.

In order to illustrate what a powerful tool Eq. (210) is, let us use to calculate the electrostatic field in two systems. In the first of them, a plane, circular, conducting disk of radius R, separated with a very thin cut from the remaining conducting plane, is biased with potential $\phi = V$, while the rest of the plane is grounded – see Fig. 31.

[71] This symmetry, evident for the particular cases (204) and (206), may be readily proved for the general case by applying Eq. (207) to the functions $f(\mathbf{r}) \equiv G(\mathbf{r}, \mathbf{r}')$ and $g(\mathbf{r}) \equiv G(\mathbf{r}, \mathbf{r}'')$. With this substitution, the left-hand side of that equality becomes equal to $-4\pi[G(\mathbf{r}'', \mathbf{r}') - G(\mathbf{r}', \mathbf{r}'')]$, while the right-hand side is zero, due to Eq. (205).

[72] In some textbooks, the sign before the surface integral is negative, because their authors use the outer normal to the *free-space* region V rather than those *occupied by conductors* – as I do.

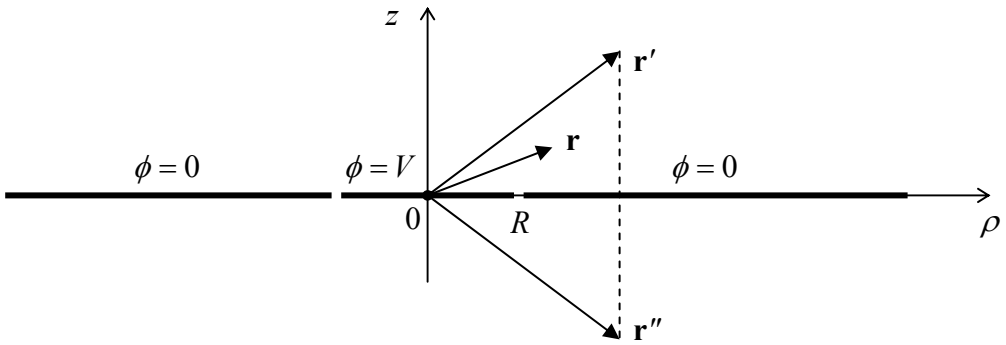


Fig. 2.31. A voltage-biased conducting disk separated from the rest of a conducting plane.

If the width of the gap between the disk and the rest of the plane is negligible, we may apply Eq. (210) without stand-alone charges, $\rho(\mathbf{r}') = 0$, and the Green's function for the uncut plane – see Eq. (206).[73] In the cylindrical coordinates, with the origin at the disk's center (Fig. 31), the function is

$$G(\mathbf{r},\mathbf{r}') = \frac{1}{\left[\rho^2 + \rho'^2 - 2\rho\rho'\cos(\varphi-\varphi') + (z-z')^2\right]^{1/2}} - \frac{1}{\left[\rho^2 + \rho'^2 - 2\rho\rho'\cos(\varphi-\varphi') + (z+z')^2\right]^{1/2}}. \tag{2.211}$$

(The sum of the first three terms under each square root in Eq. (211) is just the squared distance between the horizontal projections $\boldsymbol{\rho}$ and $\boldsymbol{\rho}'$ of the vectors $\mathbf{r}$ and $\mathbf{r}'$ (or $\mathbf{r}''$) correspondingly, while the last terms are the squares of their vertical displacements.)

Now we can readily calculate the derivative participating in Eq. (210), for $z \geq 0$:

$$\frac{\partial G}{\partial n'}\Big|_S = \frac{\partial G}{\partial z'}\Big|_{z'=+0} = \frac{2z}{\left(\rho^2 + \rho'^2 - 2\rho\rho'\cos(\varphi-\varphi') + z^2\right)^{3/2}}. \tag{2.212}$$

Due to the axial symmetry of the system, we may take φ for zero. With this, Eqs. (210) and (212) yield

$$\phi = \frac{V}{4\pi}\oint_S \frac{\partial G(\mathbf{r},\mathbf{r}')}{\partial n'} d^2r' = \frac{Vz}{2\pi}\int_0^{2\pi} d\varphi' \int_0^R \frac{\rho' d\rho'}{\left(\rho^2 + \rho'^2 - 2\rho\rho'\cos\varphi' + z^2\right)^{3/2}}. \tag{2.213}$$

This integral is not overly pleasing, but may be readily worked out at least for points on the symmetry axis ($\rho = 0$, $z \geq 0$):[74]

$$\phi = Vz\int_0^R \frac{\rho' d\rho'}{\left(\rho'^2 + z^2\right)^{3/2}} = \frac{V}{2}\int_0^{R^2/z^2} \frac{d\xi}{(\xi+1)^{3/2}} = V\left[1 - \frac{z}{\left(R^2 + z^2\right)^{1/2}}\right], \tag{2.214}$$

This result shows that if $z \to 0$, the potential tends to V (as it should), while at $z >> R$,

$$\phi \to V\frac{R^2}{2z^2}. \tag{2.215}$$

[73] Indeed, if all parts of the cut plane are grounded, a narrow cut does not change the field distribution, and hence the Green's function, significantly.

[74] There is no need to repeat the calculation for $z \leq 0$: from the symmetry of the problem, $\phi(-z) = \phi(z)$.

Now let us use the same Eq. (210) to solve the (in :-)famous problem of the cut sphere (Fig. 32). Again, if the gap between the two conducting semi-spheres is very thin ($t << R$), we may use Green's function for the grounded (and uncut) sphere. For a particular case $\mathbf{r}' = d\mathbf{n}_z$, this function follows from Eqs. (197)-(198); generalizing the former relation for an arbitrary direction of vector $\mathbf{r}'$, we get

$$G = \frac{1}{\left[r^2 + r'^2 - 2rr'\cos\gamma\right]^{1/2}} - \frac{R/r'}{\left[r^2 + (R^2/r')^2 - 2r(R^2/r')\cos\gamma\right]^{1/2}}, \qquad \text{for } r, r' \geq R, \quad (2.216)$$

where γ is the angle between the vectors $\mathbf{r}$ and $\mathbf{r}'$, and hence $\mathbf{r}''$ – see Fig. 32.

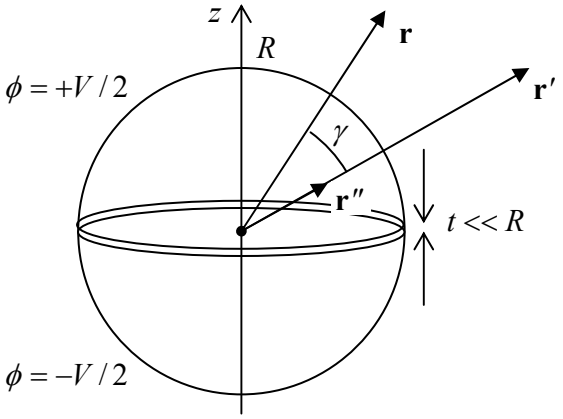


Fig. 2.32. A system of two separated, oppositely biased semi-spheres.

Now, calculating the Green's function's derivative,

$$\frac{\partial G}{\partial r'}\Big|_{r'=R+0} = -\frac{(r^2 - R^2)}{R\left[r^2 + R^2 - 2Rr\cos\gamma\right]^{3/2}}, \quad (2.217)$$

and plugging it into Eq. (210), we see that the integration is again easy only for the field on the symmetry axis (where $\mathbf{r} = z\mathbf{n}_z$, and $\gamma = \theta'$), giving:

$$\phi = \frac{V}{2}\left[1 - \frac{z^2 - R^2}{z\left(z^2 + R^2\right)^{1/2}}\right], \quad \text{for } \theta = 0. \quad (2.218)$$

For $z \to R$, this relation yields $\phi \to V/2$ (as it should), while for $z/R \to \infty$,

$$\phi \to \frac{3R^2}{4z^2}V. \quad (2.219)$$

As will be discussed in the next chapter, such a field is typical for an electric dipole.

2.11. Numerical methods

Despite the richness of analytical methods, for many boundary problems (especially in geometries without a high degree of symmetry), the numerical approach is the only way to the solution.

Though software packages offering their automatic numerical solution are abundant nowadays,[75] it is important for every educated physicist to understand "what is under the hood", at least because most universal programs exhibit mediocre performance in comparison with custom codes written for particular problems, and sometimes do not converge at all, especially for fast-changing (say, exponential) functions. The very brief discussion presented here[76] is a (hopefully, useful) fast glance under the hood, though it is certainly insufficient for professional numerical research work.

The simplest of the numerical approaches to the solution of partial differential equations, such as the Poisson or the Laplace equations (1.41)-(1.42), is the *finite-difference* method,[77] in which the sought continuous scalar function $f(\mathbf{r})$, such as the potential $\phi(\mathbf{r})$, is represented by its values in discrete points of a rectangular grid (frequently called *mesh*) of the corresponding dimensionality – see Fig. 33.

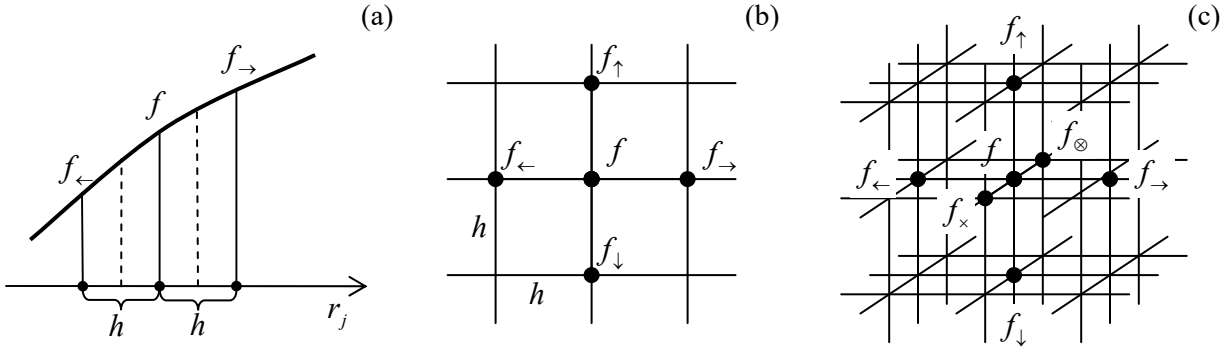


Fig. 2.33. The general idea of the finite-difference method in (a) one, (b) two, and (c) three dimensions.

Each partial second derivative of the function is approximated by the formula that readily follows from linear approximations of the function f and then its partial derivatives – see Fig. 33a:

$$\frac{\partial^2 f}{\partial r_j^2} \equiv \frac{\partial}{\partial r_j}\left(\frac{\partial f}{\partial r_j}\right) \approx \frac{1}{h}\left(\frac{\partial f}{\partial r_j}\bigg|_{r_j+h/2} - \frac{\partial f}{\partial r_j}\bigg|_{r_j-h/2}\right) \approx \frac{1}{h}\left[\frac{f_\rightarrow - f}{h} - \frac{f - f_\leftarrow}{h}\right] = \frac{f_\rightarrow + f_\leftarrow - 2f}{h^2}, \quad (2.220)$$

where $f_\rightarrow \equiv f(r_j + h)$ and $f_\leftarrow \equiv f(r_j - h)$. (The relative error of this approximation is of the order of $h^4\partial^4 f/\partial r_j^4$.) As a result, the action of a 2D Laplace operator on the function f may be approximated as

$$\frac{\partial^2 f}{\partial x^2} + \frac{\partial^2 f}{\partial y^2} \approx \frac{f_\rightarrow + f_\leftarrow - 2f}{h^2} + \frac{f_\uparrow + f_\downarrow - 2f}{h^2} = \frac{f_\rightarrow + f_\leftarrow + f_\uparrow + f_\downarrow - 4f}{h^2}, \quad (2.221)$$

and of the 3D operator, as

$$\frac{\partial^2 f}{\partial x^2} + \frac{\partial^2 f}{\partial y^2} + \frac{\partial^2 f}{\partial z^2} \approx \frac{f_\rightarrow + f_\leftarrow + f_\uparrow + f_\downarrow + f_\otimes + f_\times - 6f}{h^2}. \quad (2.222)$$

(The notation used in Eqs. (221)-(222) should be clear from Figs. 33b and 33c, respectively.)

[75] See, for example, MA Secs. 16 (iii) and (iv).

[76] It is almost similar to that given in CM Sec. 8.5 and is reproduced here for the reader's convenience, illustrated with examples from this (EM) course.

[77] For more details see, e.g., R. Leveque, *Finite Difference Methods for Ordinary and Partial Differential Equations*, SIAM, 2007.

As a simple example, let us use this scheme to find the electrostatic potential distribution inside a cylindrical box with conducting walls and square cross-section $a \times a$, using an extremely coarse mesh with step $h = a/2$ (Fig. 34). In this case, our function, the electrostatic potential $\phi(x, y)$, equals zero at the side and bottom walls, and V_0 at the top lid, so, according to Eq. (221), the 2D Laplace equation may be approximated as

$$\frac{0 + 0 + V_0 + 0 - 4\phi}{(a/2)^2} = 0. \tag{2.223}$$

The resulting value for the potential in the center of the box is $\phi = V_0/4$.

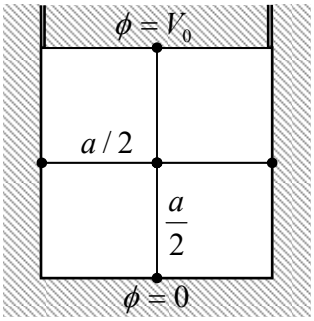


Fig. 2.34. Numerically solving an internal 2D boundary problem for a conducting, cylindrical box with a square cross-section, using a very coarse mesh (with $h = a/2$).

Surprisingly, this is the *exact* value! This may be proved either by solving this problem by the variable separation method, just as this has been done for a similar 3D problem in Sec. 5, or just from the following Green's-function argument. If all four walls of our 2D volume were biased to the voltage V_0, there would be no electric field in it at all, so the middle-point potential would be equal to V_0 as well. However, from the point of view of Eq. (210) with no bulk charge, $\rho(\mathbf{r}) = 0$, this result may be legitimately viewed as the linear superposition of the four contributions of the potentials $\phi_k = V_0$ of each wall. Since for this symmetric geometry, the corresponding geometrical factors are equal, the contribution of one wall, with $\phi_k = 0$ on all other walls (as in our current problem), has to equal $V_0/4$.

For a similar 3D problem (a cubic box), with a similar 3D mesh, Eq. (222) yields

$$\frac{0 + 0 + V_0 + 0 + 0 + 0 - 6\phi}{(a/2)^2} = 0, \tag{2.226}$$

so $\phi = V_0/6$. Using the same Green's-function argument, now for six walls of the cube, we see that this result is also exact! (This fact also follows from our variable-separation result expressed by Eqs. (95) and (99) with $a = b = c$.)

Though such exact results should be considered as a happy coincidence rather than the general law, they still show that numerical methods, even with relatively crude meshes, may be more computationally efficient than some "analytical" approaches, like the variable separation method with its infinite-sum results that, in most cases, require computers anyway – at least for the result's comprehension and analysis.

A more powerful (but also much more complex) approach is the *finite-element* method in which the discrete point mesh, typically with triangular cells, is (automatically) generated in accordance with the system geometry.[78] Such mesh generators provide higher point concentration near sharp convex

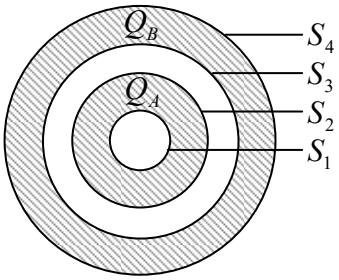
[78] See, e.g., CM Fig. 8.14.

parts of conductor surfaces, where the field concentrates and hence the potential changes faster, thus ensuring a better accuracy-to-speed tradeoff than the finite-difference methods on a uniform grid. The price to pay for this improvement is the algorithm's complexity which makes its adjustments much harder. Unfortunately, in this series, I do not have time for going into the details of that method and have to refer the reader to the special literature on this subject.[79]

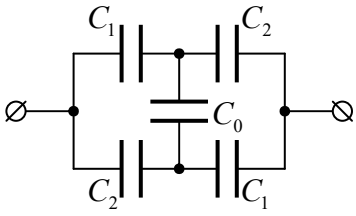
2.12. Exercise problems

2.1. Calculate the force (per unit area) exerted on a conducting surface by an external electric field normal to it. Compare the result with the electric field's definition given by Eq. (1.6), and comment.

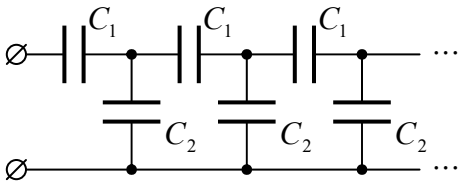
2.2. Electric charges Q_A and Q_B have been placed on two conducting concentric spherical shells – see the figure on the right. What is the full charge of each of the surfaces S_1-S_4?



2.3. Calculate the mutual capacitance between the terminals of the lumped-capacitor circuit shown in the figure on the right. Analyze and interpret the result.

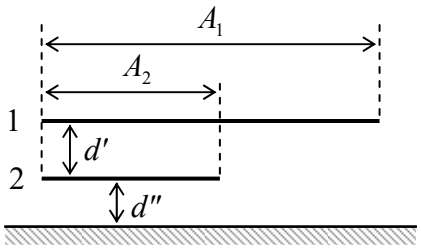


2.4. Calculate the mutual capacitance between the terminals of the semi-infinite lumped-capacitor circuit shown in the figure on the right, and find the law of the applied voltage's decay along the system. Analyze and interpret the result.



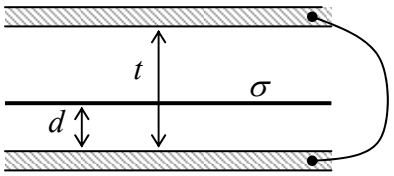
2.5. A system of two thin conducting plates is located over a ground plane as shown in the figure on the right, where A_1 and A_2 are the areas of the indicated plate parts, while *d'* and *d"* are the distances between them. Neglecting the fringe effects, calculate:

(i) the effective capacitance of each plate, and
(ii) their mutual capacitance.

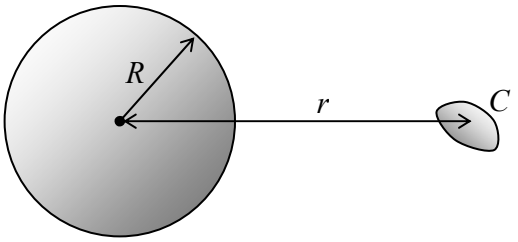


[79] See, e.g., either C. Johnson, *Numerical Solution of Partial Differential Equations by the Finite Element Method*, Dover, 2009, or T. Hughes, *The Finite Element Method*, Dover, 2000.

2.6. A wide and thin film carrying a uniformly distributed electric charge of areal density σ is placed inside a plane capacitor whose plates are connected with a wire – see the figure on the right. Neglecting the fringe effects, calculate the surface charges of the plates and the net force exerted on the film (per unit area).



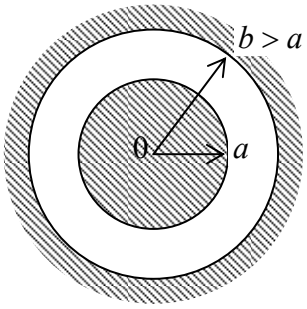
2.7. A relatively small conductor (possibly, of an irregular shape) with self-capacitance C is located at distance r from the center of a conducting sphere of radius R – see the figure on the right. In the first approximation in C, find the reciprocal capacitance matrix of the system. Use the matrix to calculate its potential energy and the force of the conductor interaction for two cases:



(i) the conductor *charges* Q are equal, and

(ii) the conductor *potentials* ϕ are kept equal.

2.8. Use the Gauss law to calculate the mutual capacitance of the following two-electrode systems, both with the cross-section shown in Fig. 7 (reproduced on the right):

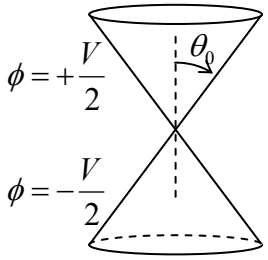


(i) a conducting sphere in the center of a spherical cavity inside another conductor, and

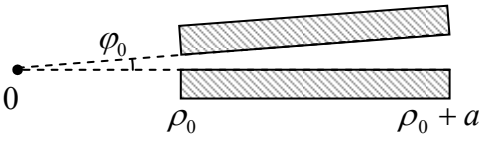
(ii) a long conducting round cylinder on the axis of a cylindrical cavity inside another conductor, i.e. a coaxial cable. (In this case, we speak about the capacitance per unit length).

Compare the results with those obtained in Sec. 2.2 using the Laplace equation.

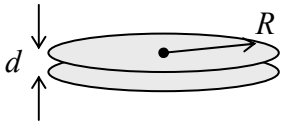
2.9. Calculate the electrostatic potential distribution around two barely separated conductors in the form of coaxial round cones (see the figure on the right), with voltage V between them. Compare the result with that of a similar 2D problem, with the cones replaced with plane-face wedges. Can you calculate the mutual capacitances between the conductors in these systems? If not, can you estimate them?



2.10. Calculate the mutual capacitance between two rectangular planar electrodes of area $A = a \times l$, with a very small angle φ_0 between them – see the figure on the right.



2.11. Using the results for a single thin round disk, obtained in Sec. 4, consider a system of two such disks at a small distance $d << R$ from each other – see the figure on the right. In particular, calculate:

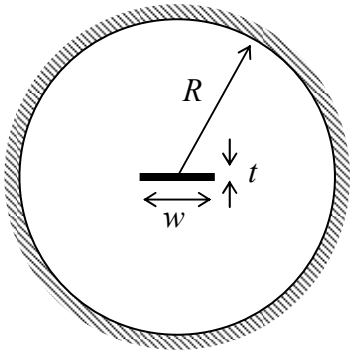


(i) the reciprocal capacitance matrix of the system,

(ii) the mutual capacitance between the disks,
(iii) the partial capacitance of one disk, and
(iv) the effective capacitance of one disk,

– all in the first nonvanishing approximation in $d/R << 1$. Compare the results (ii)-(iv) and interpret their similarities and differences.

2.12.* Calculate the mutual capacitance (per unit length) between two cylindrical conductors forming a system with the cross-section shown in the figure on the right, in the limit $t << w << R$.



Hint: You may like to use the elliptic coordinates mentioned in Sec. 4. They are defined by the following equality:

$$x + iy = c\cosh(\mu + i\nu), \qquad (*)$$

where c is a constant.

2.13. Calculate the mutual capacitance (per unit length) between two similar, long, parallel wires, each with a round cross-section of radius R, whose axes are separated by distance $d > 2R$. Explore and interpret the result in the limits $R \to 0$ and $R \to 2d$.

Hint: You may like to use the 2D orthogonal *bipolar coordinates* $\{\tau, \sigma\}$ defined by the following relations with the Cartesian coordinates $\{x, y\}$:

$$x = a\frac{\sinh\tau}{\cosh\tau - \cos\sigma}, \qquad y = a\frac{\sin\sigma}{\cosh\tau - \cos\sigma}, \qquad \text{with } -\infty < \tau < +\infty, \quad -\pi \le \sigma \le +\pi.$$

In these coordinates, the Laplace operator is

$$\nabla^2 = \frac{1}{a^2}(\cosh\tau - \cos\sigma)\left(\frac{\partial^2}{\partial\tau^2} + \frac{\partial^2}{\partial\sigma^2}\right).$$

2.14. Formulate 2D electrostatic problems that may be simply solved using each of the following analytic functions of the complex variable $\boldsymbol{z} \equiv x + iy$:

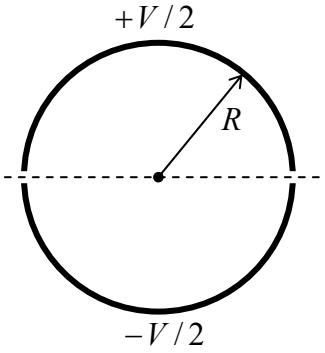
(i) $\boldsymbol{w} = \ln \boldsymbol{z}$,
(ii) $\boldsymbol{w} = \boldsymbol{z}^{1/2}$,
(iii) $\boldsymbol{w} = \boldsymbol{z} + 1/\boldsymbol{z}$,

and solve these problems.

2.15. On each side of a cylindrical volume with a rectangular cross-section $a \times b$, with no electric charges inside it, the electric field's component normal to the side's plane is constant, and also equal and opposite to that on the opposite side. Calculate the distribution of the electric potential inside the volume, provided that the magnitude of the normal components on the sides of length b equals E. Suggest a practicable method to implement such potential distribution.

2.16. Complete the solution of the problem shown in Fig. 12, by calculating the distribution of the surface charge on the semi-planes. Can you calculate the mutual capacitance between the semi-planes (per unit length of the system)? If not, can you estimate it?

2.17. A straight, long, thin, round-cylindrical conducting pipe has been cut, along its axis, into two equal parts – see the figure on the right.



(i) Use the conformal mapping method to calculate the distributions of the electrostatic potential created by voltage V applied between the two parts, both outside and inside the pipe, and of the surface charge.

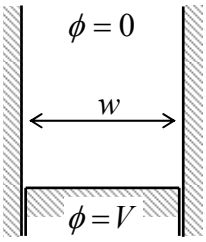
(ii)* Calculate the mutual capacitance between the pipe's halves (per unit length), taking into account a small width $2t << R$ of the cut.

Hints: In Task (i), you may like to use the complex function

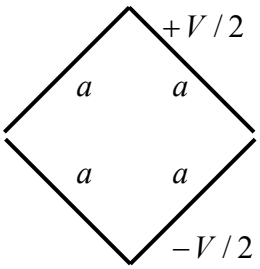
$$w = \ln \frac{R+z}{R-z},$$

while in Task (ii), you may use the solution of the previous problem.

2.18. A gap of constant width w between two grounded conducting semi-spaces is closed, from one side, with a conducting plunger biased with voltage V, so that the cross-section of the system looks like the figure on the right shows. Use the variable separation method to calculate the distribution of the electrostatic potential within the gap.

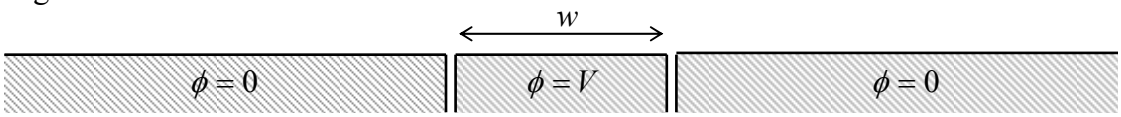


2.19. Use the variable separation method to calculate the electrostatic potential's distribution inside a very long thin-wall metallic box with a quadratic cross-section, cut and voltage-biased as shown in the figure on the right. (Assume that the cut's width is negligibly small.)



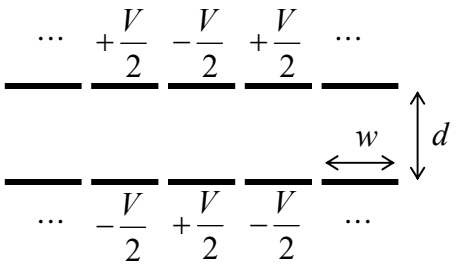
2.20. Solve Problem 17(i) by using the variable separation method, and compare the results.

2.21. Use the variable separation method to calculate the potential distribution above the plane surface of a conductor, with a strip of width w separated by very thin cuts, and biased with voltage V – see the figure below.



2.22. The previous problem is now modified: the cut-out and voltage-biased part of the conducting plane is now not a strip, but a square with side w. Calculate the potential distribution above the conductor's surface.

2.23. Each electrode of a large plane capacitor is cut into parallel long strips of equal width w, with very narrow gaps between them. These strips are kept at alternating potentials, as shown in the figure on the right. Use the variable separation method to calculate the electrostatic potential distribution in space, and explore the limit $w << d$.



2.24. Complete the cylinder problem started in Sec. 7 (see Fig. 17), for the cases when the top lid's voltage is fixed as follows:

(i) $V = V_0 J_1(\xi_{11}\rho/R)\sin\varphi$, where $\xi_{11} \approx 3.832$ is the first root of the Bessel function $J_1(\xi)$;

(ii) $V = V_0 = \text{const}$.

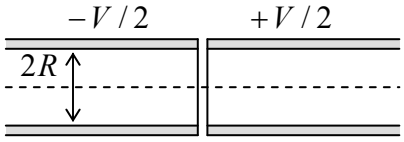
For both cases, calculate the electric field at the centers of the lower and upper lids. (For Task (ii), an answer including series and/or integrals is acceptable.)

2.25. For the infinitely long periodic system sketched in Fig. 21, assuming that $t << h, R$:

(i) calculate and sketch the electrostatic potential's distribution inside the system for various values of the ratio R/h, and

(ii) simplify the results for the limit $R/h \to 0$.

2.26. A long round cylindrical conducting pipe is split, with a very narrow cut normal to its axis, into two parts that are voltage-biased as the figure on the right shows. Use two different approaches to calculate the force exerted by the resulting field upon a charged particle flying along the pipe close to its axis. Can the system work as an electrostatic lens?



2.27. Use the variable separation method to find the potential distribution inside and outside of a thin spherical shell of radius R, with a fixed potential distribution on it: $\phi(R,\theta,\varphi) = V_0 \sin\theta\cos\varphi$.

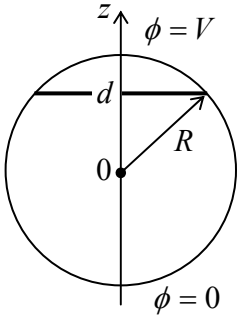
2.28. A thin spherical shell carries an electric charge with areal density $\sigma = \sigma_0\cos\theta$. Calculate the spatial distribution of the electrostatic potential and the electric field, both inside and outside the shell.

2.29. Use the variable separation method to solve the problem already considered in Sec. 10: calculate the potential distribution both inside and outside of a thin spherical shell of radius R, separated with a very thin cut along the central plane $z = 0$ into two halves, with voltage V applied between them – see Fig. 32. Analyze the solution; in particular, compare the field at the z-axis, for $z > R$, with Eq. (218).

Hint: You may like to use the following integral of a Legendre polynomial with an odd index $l = 1, 3, 5\ldots = 2n - 1$:[80]

[80] As a reminder, the *double factorial* (also called "semifactorial") operator (!!) is similar to the usual factorial operator (!), but with the product limited to numbers of the same parity as its argument – in our particular case, of the odd numbers in the numerator and even numbers in the denominator.

$$I_n \equiv \int_0^1 P_{2n-1}(\xi)\,d\xi = \frac{1}{n!}\cdot\left(\frac{1}{2}\right)\cdot\left(-\frac{3}{2}\right)\cdot\left(-\frac{5}{2}\right)\cdot\ldots\cdot\left(\frac{3}{2}-n\right) \equiv (-1)^{n-1}\,\frac{(2n-3)!!}{2n(2n-2)!!}.$$



2.30. Calculate, up to terms $O(1/r^2)$, the long-range electric field induced by a split and voltage-biased conducting sphere – similar to that discussed in Sec. 10 (see Fig. 32) and in the previous problem, but with the cut's plane at an arbitrary distance $d < R$ from the center – see the figure on the right.

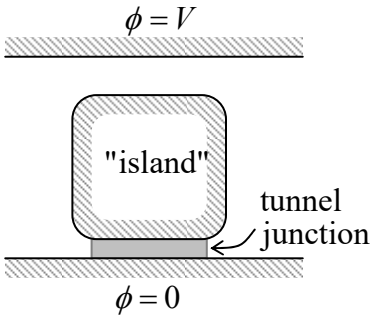
2.31. Calculate the field distribution in the simple electrostatic lens that was the subject of Problem 1.9, provided that the separation of the two field regions is provided by a thin conducting membrane, with a round hole of radius R.

Hint: You may like to use the fact that the general axially symmetric solution of the Laplace equation in the oblate ellipsoidal coordinates (59) may be represented in the following variable-separation form:

$$\phi = \sum_{n=0}^{\infty}\left[p_n \mathcal{P}_n(i\sinh\alpha) + q_n \mathcal{Q}_n(i\sinh\alpha)\right]\mathcal{P}_n(\cos\beta),$$

where p_n and q_n are constants, $\mathcal{P}_n$ are the Legendre polynomials (2.169), which are sometimes called the *Legendre functions of the first kind*, while $\mathcal{Q}_n$ are the *Legendre functions of the second kind* (briefly mentioned, in a different context, in Sec. 2.8) that may be defined by the following recurrence relations:

$$\mathcal{Q}_0(\xi) = \frac{1}{2}\ln\frac{1+\xi}{1-\xi}, \qquad \mathcal{Q}_1(\xi) = \mathcal{P}_1(\xi)\mathcal{Q}_0(\xi) - 1, \qquad \mathcal{Q}_{n>2}(\xi) = \frac{2n-1}{n}\xi\mathcal{Q}_{n-1}(\xi) - \frac{n-1}{n}\mathcal{Q}_{n-2}(\xi).$$

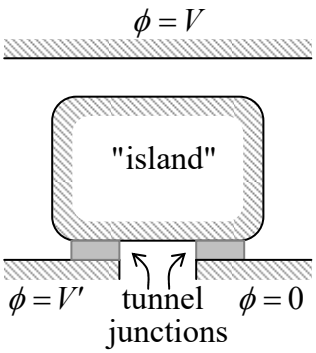


2.32. A small conductor (in this context, usually called a *single-electron island*) is placed between two conducting electrodes, with voltage V applied between them. The gap between the island and one of the electrodes is so narrow that electrons may tunnel quantum-mechanically through this "junction" – see the figure on the right. Neglecting thermal excitation effects, calculate the equilibrium charge of the island as a function of V.

Hint: To solve this problem, you do not need to know much about the quantum-mechanical tunneling between conductors, besides that such tunneling of an electron, together with energy relaxation of the resulting excitations, may be considered a single inelastic (energy-dissipating) event.[81] At negligible thermal excitations, such an event takes place only if it decreases the total potential energy of the system.

[81] Strictly speaking, this statement, implying negligible quantum-mechanical coherence of the tunneling events, is correct only if the junction transparency is so low that its effective electric resistance is much higher than the fundamental quantum unit of resistance, $R_Q \equiv \pi\hbar/2e^2 \approx 6.5$ kΩ (see, e.g., QM Sec. 3.2). However, this condition is satisfied in most experimental tunnel junctions.

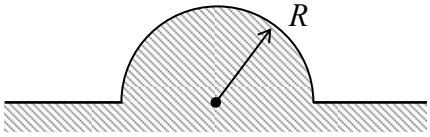
2.33. The system discussed in the previous problem is now generalized as the figure on the right shows. If the voltage V' applied between the two bottom electrodes is sufficiently large, electrons can successively tunnel through two junctions of this system (called the *single-electron transistor*), carrying dc current between these electrodes. Neglecting thermal excitations, calculate the region of voltages V and V' where such a current is fully suppressed (*Coulomb-blocked*).



2.34. Use the charge image method to calculate the full surface charges induced in the plates of a very broad, voltage-unbiased plane capacitor of thickness D by a point charge q separated from one of the electrodes by distance d. Suggest at least one alternative method to obtain the same result.

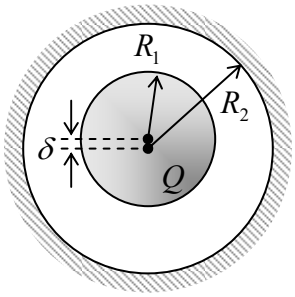
2.35. Use the charge image method to calculate the potential energy of the electrostatic interaction between a point charge placed in the center of a spherical cavity that had been carved inside a grounded conductor, and the cavity's walls. Looking at the result, could it be obtained in a simpler way (or ways)?

2.36. Use the method of charge images to find the Green's function of the system shown in the figure on the right, where the bulge on the (otherwise, plane) surface of a conductor has the shape of a semi-sphere of radius R.



2.37.* Use the spherical inversion expressed by Eq. (198), to develop an iterative method for a more and more precise calculation of the mutual capacitance between two similar conducting spheres of radius R, with their centers separated by distance $d > 2R$.

2.38.* A conducting sphere of radius R_1, carrying an electric charge Q, is placed inside a spherical cavity of radius $R_2 > R_1$, carved inside another bulk conductor. Calculate the electric force exerted on the sphere if its center is displaced by a small distance $\delta << R_1, R_2 - R_1$ from that of the cavity – see the figure on the right.



2.39. Within the simple models of the electric field screening in conductors, discussed in Sec. 2.1, analyze the partial screening of the electric field of a point charge q by a planar conducting film of constant thickness $t << \lambda$, where λ is (depending on charge carrier statistics) either the Debye or the Thomas-Fermi screening length – see, respectively, Eqs. (8) or (10). Assume that the distance d between the charge and the film is much larger than t.

2.40. Prove the following expansion of the simplest Green's function (204) into a series over the Legendre polynomials:

$$\frac{1}{|\mathbf{r}-\mathbf{r}'|}=\frac{1}{r_>}\sum_{l=0}^{\infty}\left(\frac{r_<}{r_>}\right)^l \mathcal{P}_l(\cos\theta),$$

where $r_>$ is the largest of the two scalars $r \equiv |\mathbf{r}| \geq 0$ and $r' \equiv |\mathbf{r}'| \geq 0$, while $r_<$ is the smallest of them.

2.41. Use the expansion that was the subject of the previous problem to confirm the analysis, in Sec. 2.9 of the lecture notes, of the system shown in Fig. 29: a grounded conducting sphere of radius R, and a point charge q located at distance $d > R$ from its center.

2.42. Suggest a convenient definition of the Green's function for 2D electrostatic problems, and calculate it for:

(i) the unlimited free space, and
(ii) the free space above a conducting plane.

Use the latter result to re-solve Problem 21.

2.43. A conducting plane is separated into two parts with a very narrow straight cut, and voltage V is applied between the resulting half-planes – see the figure below. Use the Green's function method to find the distribution of the electrostatic potential and the electric field everywhere in the space. Compare the result with Eq. (83). In hindsight, could the problem be solved in an even simpler way (or ways)?

2.44. Use the last result of Problem 42 and one of the conformal mappings discussed in Sec. 4 to find one more solution of Problem 18.

2.45. Calculate the 2D Green's functions for the free spaces:

(i) outside a round conducting cylinder, and
(ii) inside a round cylindrical hole in a conductor.

2.46. Solve Problem 17(i) using the Green's function method.

2.47. Solve the 2D boundary problem that was discussed in Sec. 11 (Fig. 34) by using:

(i) the finite difference method with the finer square mesh $h = a/3$, and
(ii) the variable separation method.

Compare the results at the mesh points, and comment.

This page is intentionally left blank

Chapter 3. Dipoles and Dielectrics

In contrast to conductors, the motion of charges in dielectrics is restricted to the atom/molecule interiors, so the electric polarization of these materials by an external field takes a different form. This issue is the main subject of this chapter, but in preparation for its analysis, we have to start with a general discussion of the electric field induced by spatially restricted systems of charges.

3.1. Electric dipole

Let us consider a localized system of charges, of a linear scale a, and derive a simple approximate expression for the electrostatic field induced by the system at a distant point $\mathbf{r}$. For that, let us select a reference frame with the origin either somewhere inside the system, or at a distance of the order of a from it (Fig. 1).

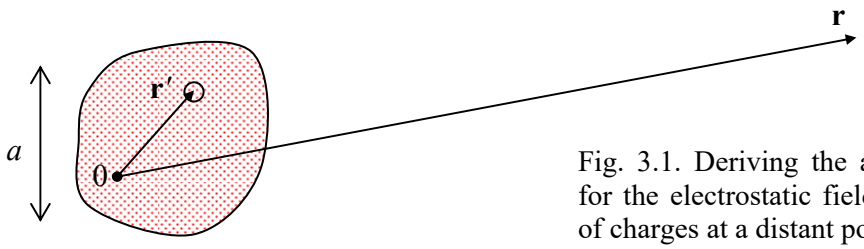


Fig. 3.1. Deriving the approximate expression for the electrostatic field of a localized system of charges at a distant point ($r >> r' \sim a$).

Then the positions of all charges of the system satisfy the following condition:

$$r' << r\,. \tag{3.1}$$

Using this condition, we can expand the general expression (1.38) for the electrostatic potential $\phi(\mathbf{r})$ of the system into the Taylor series in small parameter r'. For any function of type $f(\mathbf{r} - \mathbf{r}')$, the expansion may be represented as[1]

$$f(\mathbf{r}-\mathbf{r}') = f(\mathbf{r}) - \sum_{j=1}^{3} r_j' \frac{\partial f}{\partial r_j}(\mathbf{r})\Big|_{\mathbf{r}'=0} + \frac{1}{2!}\sum_{j,j'=1}^{3} r_j' r_{j'}' \frac{\partial^2 f}{\partial r_j \partial r_{j'}}(\mathbf{r})\Big|_{\mathbf{r}'=0} - \ldots\,. \tag{3.2}$$

Applying this formula to the fraction $1/|\mathbf{r} - \mathbf{r}'|$ in Eq. (1.38) (i.e. essentially to the free-space Green's function), we get the so-called *multipole expansion* of the electrostatic potential:

$$\phi(\mathbf{r}) = \frac{1}{4\pi\varepsilon_0}\left(\frac{1}{r}Q + \frac{1}{r^3}\sum_{j=1}^{3} r_j p_j + \frac{1}{2r^5}\sum_{j,j'=1}^{3} r_j r_{j'} \mathcal{Q}_{jj'} + \ldots\right), \tag{3.3}$$

whose $\mathbf{r}$-independent parameters are defined as follows:

$$Q \equiv \int \rho(\mathbf{r}')d^3r', \qquad p_j \equiv \int \rho(\mathbf{r}')r_j' d^3r', \qquad \mathcal{Q}_{jj'} \equiv \int \rho(\mathbf{r}')\left(3r_j' r_{j'}' - r'^2\delta_{jj'}\right)d^3r'\,. \tag{3.4}$$

[1] See, e.g., MA Eq. (2.11b).

(Indeed, the two leading terms of the expansion (2) may be rewritten in the vector form $f(\mathbf{r}) - \mathbf{r}'\cdot\nabla f(\mathbf{r})$, and the gradient of such a spherically-symmetric function $f(r) = 1/r$ is just $\mathbf{n}_r df/dr$, so

$$\frac{1}{|\mathbf{r}-\mathbf{r}'|} \approx \frac{1}{r} - \mathbf{r}'\cdot\mathbf{n}_r \frac{d}{dr}\left(\frac{1}{r}\right) = \frac{1}{r} + \mathbf{r}'\cdot\frac{\mathbf{r}}{r^3}, \tag{3.5}$$

immediately giving the two first terms of Eq. (3). The proof of the third, *quadrupole* term in Eq. (3) is similar but a bit longer, and is left for the reader's exercise.)

Evidently, the scalar parameter Q in Eqs. (3)-(4) is just the total charge of the system. The constants p_j may be considered as Cartesian components of the following vector:

$$\mathbf{p} \equiv \int \rho(\mathbf{r}')\mathbf{r}' d^3 r', \tag{3.6}$$

Electric dipole moment

called the system's *electric dipole moment*, and $\mathcal{Q}_{jj'}$ are Cartesian elements of a tensor – system's *electric quadrupole moment*. If $Q \neq 0$, all higher terms on the right-hand side of Eq. (3), at large distances (1), are just small corrections to the first one, and in many cases may be ignored. However, the net charge of many systems is exactly zero, the most important examples being neutral atoms and molecules. For such neural systems, the second (*dipole*) term in Eq. (3) is, most frequently, the leading one. Such systems are called *electric dipoles*. Due to their importance, let us rewrite the expression for the dipole term in three other, mathematically equivalent forms:

$$\phi_{\rm d} \equiv \frac{1}{4\pi\varepsilon_0}\frac{\mathbf{r}\cdot\mathbf{p}}{r^3} \equiv \frac{1}{4\pi\varepsilon_0}\frac{p\cos\theta}{r^2} \equiv \frac{1}{4\pi\varepsilon_0}\frac{pz}{\left(x^2+y^2+z^2\right)^{3/2}}, \tag{3.7}$$

Electric dipole's potential

that are more convenient for some applications. Here θ is the angle between the vectors $\mathbf{p}$ and $\mathbf{r}$, and in the last (Cartesian) representation, the z-axis is directed along the vector $\mathbf{p}$. Fig. 2a shows equipotential surfaces of the dipole field – or rather their cross-sections by any plane in which the vector $\mathbf{p}$ resides.

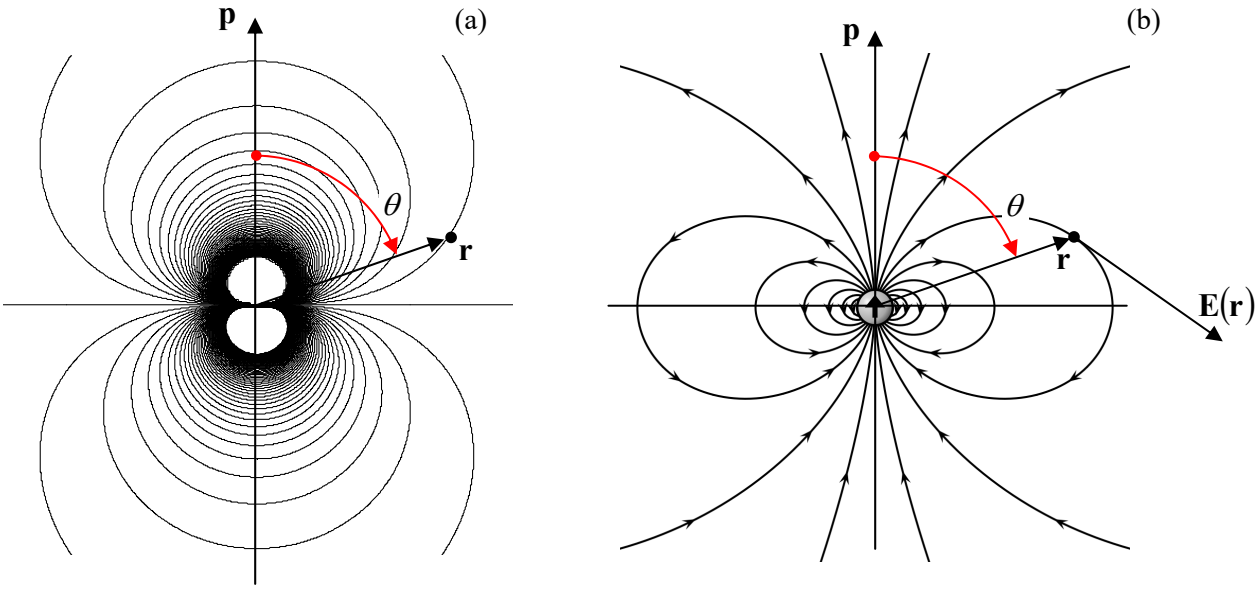


Fig. 3.2. (a) The equipotential surfaces and (b) the electric field lines of a dipole. (Panel (b) adapted from http://en.wikipedia.org/wiki/Dipole under the GNU Free Documentation License.)

The simplest example of a system whose field, at large distances, approaches the dipole field (7), is two equal but opposite point charges ("poles"), $+q$ and $-q$, with the radius vectors, respectively, $\mathbf{r}_+$ and $\mathbf{r}_-$:

$$\rho(\mathbf{r}) = (+q)\delta(\mathbf{r}-\mathbf{r}_+) + (-q)\delta(\mathbf{r}-\mathbf{r}_-)\,. \tag{3.8}$$

For this system (sometimes called the *physical dipole*), Eq. (4) yields

$$\mathbf{p} = (+q)\mathbf{r}_+ + (-q)\mathbf{r}_- = q(\mathbf{r}_+ - \mathbf{r}_-) = q\mathbf{a}\,, \tag{3.9}$$

where $\mathbf{a}$ is the vector connecting the points $\mathbf{r}_-$ and $\mathbf{r}_+$. Note that in this case (and indeed for all systems with $Q = 0$), the dipole moment does not depend on the choice of the reference frame's origin.

A less trivial example of a dipole is a conducting sphere of radius R in a uniform external electric field $\mathbf{E}_0$. As a reminder, its field was calculated in Sec. 2.8, and its result is expressed by Eq. (2.176). The first term in the parentheses of that relation describes just the external field (2.173), so the field of the sphere itself (i.e. that of the surface charge induced by $\mathbf{E}_0$) is given by the second term:

$$\phi_s = \frac{E_0 R^3}{r^2}\cos\theta\,. \tag{3.10}$$

Comparing this expression with the second form of Eq. (7), we see that the sphere has an *induced* dipole moment

$$\mathbf{p} = 4\pi\varepsilon_0 \mathbf{E}_0 R^3. \tag{3.11}$$

This is an interesting example of a virtually pure dipole field: at all points outside the sphere ($r > R$), the field has neither a quadrupole moment nor any higher moments.

Other examples of dipole fields are given by two more systems discussed in Chapter 2 – see Eqs. (2.215) and (2.219). Those systems, however, do have higher-order multipole moments, so for them, Eq. (7) gives only the long-distance approximation.

Now returning to the general properties of the dipole field (7), let us calculate its major characteristics. First of all, we may use Eq. (7) to calculate the electric field of a dipole:

$$\mathbf{E}_d = -\nabla\phi_d = -\frac{1}{4\pi\varepsilon_0}\nabla\left(\frac{\mathbf{r}\cdot\mathbf{p}}{r^3}\right) = -\frac{1}{4\pi\varepsilon_0}\nabla\left(\frac{p\cos\theta}{r^2}\right). \tag{3.12}$$

This differentiation is easiest in the spherical coordinates, using the well-known expression for the gradient of a scalar function in these coordinates[2] and taking the z-axis parallel to the dipole moment $\mathbf{p}$. From the last form of Eq. (12), we immediately get

Electric dipole's field

$$\boxed{\mathbf{E}_d = \frac{p}{4\pi\varepsilon_0 r^3}\left(2\mathbf{n}_r\cos\theta + \mathbf{n}_\theta\sin\theta\right) \equiv \frac{1}{4\pi\varepsilon_0}\frac{3\mathbf{r}(\mathbf{r}\cdot\mathbf{p}) - \mathbf{p}r^2}{r^5}\,.} \tag{3.13}$$

Fig. 2b above shows the electric field lines given by Eqs. (13). The most important features of this result are a faster drop of the field's magnitude ($E_d \propto 1/r^3$, rather than $E \propto 1/r^2$ for a point charge), and the change of the signs of its radial component as a function of the polar angle $\theta \in [0, \pi]$.

[2] See, e.g., MA Eq. (10.8) with $\partial/\partial\varphi = 0$.

Next, let us use Eq. (1.55) to calculate the potential energy of interaction between a dipole and an external electric field. Assuming that this field does not change much at distances of the order of a (Fig. 1), we may expand its potential $\phi_{ext}(\mathbf{r})$ into the Taylor series, and keep only two leading terms:

$$U = \int \rho(\mathbf{r})\phi_{ext}(\mathbf{r})d^3r \approx \int \rho(\mathbf{r})\left[\phi_{ext}(0) + \mathbf{r}\cdot\nabla\phi_{ext}(0)\right]d^3r \equiv Q\phi_{ext}(0) - \mathbf{p}\cdot\mathbf{E}_{ext}. \qquad (3.14)$$

The first term is the potential energy the system would have if it were just a point charge. If the net charge Q is zero, that term disappears, and the leading contribution is due to the dipole moment:

$$U = -\mathbf{p}\cdot\mathbf{E}_{ext}, \qquad \text{for } \mathbf{p} = \text{const}. \qquad (3.15a)$$

Dipole's energy in external field

Note that this result is only valid for a *fixed* dipole, with $\mathbf{p}$ independent of $\mathbf{E}_{ext}$. In the opposite limit, when the dipole is *induced* by the field, i.e. $\mathbf{p} \propto \mathbf{E}_{ext}$ (you may have one more look at Eq. (11) to see an example of such a proportionality), we need to start with Eq. (1.60) rather than Eq. (1.55), getting

$$U = -\frac{1}{2}\mathbf{p}\cdot\mathbf{E}_{ext}, \qquad \text{for } \mathbf{p} \propto \mathbf{E}_{ext}. \qquad (3.15b)$$

In particular, combining Eqs. (13) and Eq. (15a), we may get the following important formula for the interaction of two independent dipoles:

$$U_{int} = \frac{1}{4\pi\varepsilon_0}\frac{\mathbf{p}_1\cdot\mathbf{p}_2 r^2 - 3(\mathbf{r}\cdot\mathbf{p}_1)(\mathbf{r}\cdot\mathbf{p}_2)}{r^5} = \frac{1}{4\pi\varepsilon_0}\frac{p_{1x}p_{2x} + p_{1y}p_{2y} - 2p_{1z}p_{2z}}{r^3}, \qquad (3.16)$$

where $\mathbf{r}$ is the vector connecting the dipoles, and the z-axis is directed along this vector. It is easy to prove (this exercise is left for the reader) that if the magnitude p of each dipole moment is fixed (the approximation valid, in particular, for weak interaction of so-called *polar molecules*), this potential energy reaches its minimum at the parallel orientation of the dipoles along the line connecting them. Note also that in this case, U_{int} is proportional to $1/r^3$. On the other hand, if each moment $\mathbf{p}$ has a random value plus a component due to its polarization by the electric field of its counterpart: $\Delta\mathbf{p}_{1,2} \propto \mathbf{E}_{2,1} \propto 1/r^3$, their average interaction energy (which may be calculated from Eq. (16) with the additional factor ½) is always negative and is proportional to $1/r^6$. Such negative potential describes, in particular, the long-range, attractive part (the so-called *London dispersion force*) of the interaction between electrically neutral atoms and molecules.[3]

According to Eqs. (15), the electric field should "try" to reach the minimum of U by aligning the dipole vector's direction with its own. The direct quantitative description of this effect is the torque $\boldsymbol{\tau}$ exerted by the field. The simplest way to calculate it is to sum up all the elementary torques $d\boldsymbol{\tau} = \mathbf{r}\times d\mathbf{F}_{ext} = \mathbf{r}\times\mathbf{E}_{ext}(\mathbf{r})\rho(\mathbf{r})d^3r$ exerted on all elementary charges of the system:

$$\boldsymbol{\tau} = \int \mathbf{r}\times\mathbf{E}_{ext}(\mathbf{r})\rho(\mathbf{r})d^3r \approx \mathbf{p}\times\mathbf{E}_{ext}(0), \qquad (3.17)$$

where at the last step, the spatial dependence of the external field $\mathbf{E}_{ext}(\mathbf{r})$ was again neglected. This dependence cannot, however, be ignored at the calculation of the *total force* exerted by the field on the dipole (with $Q = 0$). Indeed, Eqs. (15) shows that if the field is constant, the dipole's energy is

[3] Several calculations of this force, using various models, are described in the QM and SM parts of this series.

independent of its spatial location and hence the net force is zero. However, if the field has a non-zero gradient, a total force does appear; for a field-independent dipole,

$$\mathbf{F} = -\nabla U = \nabla(\mathbf{p}\cdot\mathbf{E}_{\text{ext}}), \tag{3.18}$$

where the derivative has to be taken at the dipole's position (in our notation, at $\mathbf{r} = 0$). If the dipole that is being moved in a field retains its magnitude and orientation, then the last formula is equivalent to[4]

$$\mathbf{F} = (\mathbf{p}\cdot\nabla)\mathbf{E}_{\text{ext}}. \tag{3.19}$$

Alternatively, the last expression may be obtained similarly to Eq. (14):

$$\mathbf{F} = \int \rho(\mathbf{r})\mathbf{E}_{\text{ext}}(\mathbf{r})d^3r \approx \int \rho(\mathbf{r})\left[\mathbf{E}_{\text{ext}}(0) + (\mathbf{r}\cdot\nabla)\mathbf{E}_{\text{ext}}\right]d^3r = Q\mathbf{E}_{\text{ext}}(0) + (\mathbf{p}\cdot\nabla)\mathbf{E}_{\text{ext}}. \tag{3.20}$$

Finally, let me add a note on the so-called *coarse-grain model* of the dipole. The dipole approximation explored above is asymptotically correct only *at large distances*, $r >> a$. However, for some applications (including the forthcoming discussion of the molecular field effects in Sec. 3) it is beneficial to have an expression that might be formally used *everywhere* in space, though maybe without exact details at $r \sim a$, giving the correct result for the space average of the electric field,

$$\overline{\mathbf{E}} \equiv \frac{1}{V}\int_V \mathbf{E}d^3r, \tag{3.21}$$

where V is a regularly-shaped volume much larger than a^3, for example, a sphere of radius $R >> a$, with the dipole at its center. For the field $\mathbf{E}_{\text{d}}$ given by Eq. (13), such an average is zero. Indeed, let us consider the Cartesian components of that vector in a reference frame with the z-axis directed along the vector $\mathbf{p}$. Due to the axial symmetry of the field, the averages of the components E_x and E_y vanish. Let us use Eq. (13) to spell out the "vertical" component of the field (parallel to the dipole moment vector):

$$E_z \equiv \mathbf{E}_{\text{d}}\cdot\frac{\mathbf{p}}{p} = \frac{1}{4\pi\varepsilon_0 r^3}\left(2\mathbf{n}_r\cdot\mathbf{p}\cos\theta - \mathbf{n}_\theta\cdot\mathbf{p}\sin\theta\right) = \frac{p}{4\pi\varepsilon_0 r^3}\left(2\cos^2\theta - \sin^2\theta\right). \tag{3.22}$$

Integrating this expression over the whole solid angle $\Omega = 4\pi$, at fixed r, using a convenient variable substitution $\cos\theta \equiv \xi$, we get

$$\oint_{4\pi} E_z d\Omega = 2\pi\int_0^\pi E_z \sin\theta d\theta = \frac{p}{2\varepsilon_0 r^3}\int_0^\pi \left(2\cos^2\theta - \sin^2\theta\right)\sin\theta d\theta = \frac{p}{2\varepsilon_0 r^3}\int_{-1}^{+1}\left(3\xi^2 - 1\right)d\xi = 0. \tag{3.23}$$

On the other hand, the *exact* electric field of an *arbitrary* charge distribution, with the total dipole moment $\mathbf{p}$, obeys the following equality:

$$\int_V \mathbf{E}(\mathbf{r})d^3r = -\frac{\mathbf{p}}{3\varepsilon_0} \equiv -\frac{1}{4\pi\varepsilon_0}\frac{4\pi}{3}\mathbf{p}, \tag{3.24}$$

where the integration is over *any* sphere containing all the charges. (A proof of this formula by using Eqs. (1.9) and (1.22) is left for the reader's exercise.) The origin of the difference is illustrated in Fig. 3 on the example of a physical dipole, i.e. a system of two equal but opposite charges – see Eqs. (8)-(9).

[4] The equivalence may be proved, for example, by using MA Eq. (11.6) with $\mathbf{f} = \mathbf{p} = \text{const}$ and $\mathbf{g} = \mathbf{E}_{\text{ext}}$, taking into account that according to the general Eq. (1.28), $\nabla\times\mathbf{E}_{\text{ext}} = 0$.

The zero average (23) of the dipole field (13) does not take into account the contribution from the region between the charges where Eq. (13) is not valid, and the field is directed mostly against the dipole vector (9).

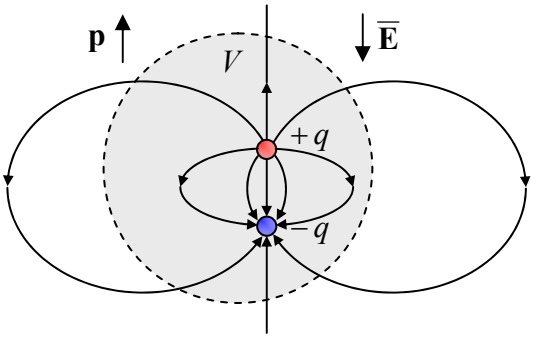


Fig. 3.3. A sketch illustrating the origin of Eq. (24) for a physical dipole.

So, in order to be used as a reasonable coarse-grain model, Eq. (13) may be modified as follows:

$$\mathbf{E}_{\rm cg} = \frac{1}{4\pi\varepsilon_0}\left[\frac{3\mathbf{r}(\mathbf{r}\cdot\mathbf{p}) - \mathbf{p}r^2}{r^5} - \frac{4\pi}{3}\mathbf{p}\,\delta(\mathbf{r})\right], \tag{3.25}$$

with the average (21) satisfying Eq. (24). Evidently, such a modification does not change the field at large distances $r >> a$, i.e. in the region where the expansion (3), and hence Eq. (13), are valid.

3.2. Dipole media

Now let us generalize Eq. (7) to the case of several (possibly, many) dipoles $\mathbf{p}_j$ located at arbitrary points $\mathbf{r}_j$. Using the linear superposition principle, we get

$$\phi_{\rm d}(\mathbf{r}) = \frac{1}{4\pi\varepsilon_0}\sum_j \mathbf{p}_j \cdot \frac{\mathbf{r}-\mathbf{r}_j}{\left|\mathbf{r}-\mathbf{r}_j\right|^3}. \tag{3.26}$$

If our system (medium) contains many similar dipoles, distributed in space with density $n(\mathbf{r})$, we may approximate the last sum with a *macroscopic potential*, which is the average of the genuine ("microscopic") potential (26) over a local volume much larger than the distance between the dipoles, and as a result, is given by the integral

$$\phi_{\rm d}(\mathbf{r}) = \frac{1}{4\pi\varepsilon_0}\int \mathbf{P}(\mathbf{r}')\cdot\frac{\mathbf{r}-\mathbf{r}'}{\left|\mathbf{r}-\mathbf{r}'\right|^3}d^3r', \qquad \text{with } \mathbf{P}(\mathbf{r}) \equiv n(\mathbf{r})\mathbf{p}, \tag{3.27}$$

where the vector $\mathbf{P}(\mathbf{r})$, called the *electric polarization*, has the physical meaning of the net dipole moment per unit volume. (Note that by its definition, $\mathbf{P}(\mathbf{r})$ is also a "macroscopic" field.)

Now comes a very impressive trick, which is the basis of all the theory of "macroscopic" electrostatics (and eventually, "macroscopic" electrodynamics). Just as was done at the derivation of Eq. (5), Eq. (27) may be rewritten in the equivalent form

$$\phi_{\rm d}(\mathbf{r}) = \frac{1}{4\pi\varepsilon_0}\int \mathbf{P}(\mathbf{r}')\cdot\nabla'\frac{1}{\left|\mathbf{r}-\mathbf{r}'\right|}d^3r' \ , \tag{3.28}$$

where ∇' means the del operator (in this particular case, the gradient) acting in the "source space" of vectors $\mathbf{r}'$. The right-hand side of Eq. (28), applied to any volume V limited by a closed surface S, may be readily integrated by parts to give[5]

$$\phi_{\rm d}(\mathbf{r}) = \frac{1}{4\pi\varepsilon_0}\oint_S \frac{P_n(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|}d^2r' - \frac{1}{4\pi\varepsilon_0}\int_V \frac{\nabla'\cdot\mathbf{P}(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|}d^3r'. \tag{3.29}$$

If the surface does not carry an infinitely dense (δ-functional) sheet of additional dipoles,[6] or it is just very distant, the first term on the right-hand side is negligible. Now comparing the second term with the basic equation (1.38) for the electric potential, we see that this term may be interpreted as the field of certain *effective* electric charges with density

Effective charge density

$$\rho_{\rm ef} = -\nabla\cdot\mathbf{P}. \tag{3.30}$$

Figure 4 illustrates the physics of this key relation for a cartoon model of a simple multi-dipole system: a layer of uniformly distributed two-point-charge units oriented normally to the layer's surface. (In this case, $\nabla\cdot\mathbf{P} = dP/dx$.) One can see that the $\rho_{\rm ef}$ defined by Eq. (30) may be interpreted as the density of the uncompensated surface charges of polarized elementary dipoles.

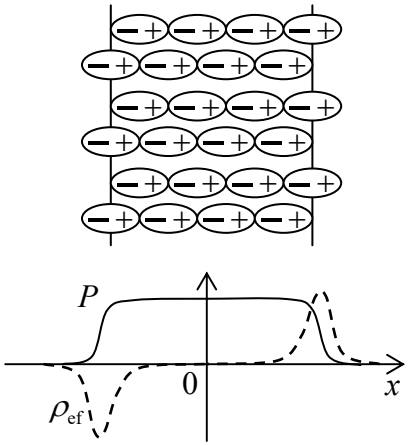


Fig. 3.4. The spatial distributions of the polarization and effective charges in a layer of similar elementary dipoles (schematically).

Next, from Sec. 1.2, we already know that Eq. (1.38) is equivalent to the inhomogeneous Maxwell equation (1.27) for the electric field, so the *macroscopic* electric field of the dipoles (defined as $\mathbf{E}_{\rm d} = -\nabla\phi_{\rm d}$, where $\phi_{\rm d}$ is given by Eq. (27)) obeys a similar equation, with the effective charge density (30).

Now let us consider a more general case when a system, besides the compensated charges of the dipoles, also has certain stand-alone charges – not parts of the dipoles already taken into account in the polarization $\mathbf{P}$. As was discussed in Sec. 1.1, if we average this charge over the inter-point-charge distances, i.e. approximate it with a continuous "macroscopic" density $\rho(\mathbf{r})$, then its macroscopic

[5] To prove this (almost evident) formula strictly, it is sufficient to apply the divergence theorem given by MA Eq. (12.2), to the vector function $\mathbf{f} = \mathbf{P}(\mathbf{r}')/|\mathbf{r}-\mathbf{r}'|$, in the "source space" of radius-vectors $\mathbf{r}'$.

[6] Just like in the case of Eq. (1.9), we may always describe such a dipole sheet using the second term in Eq. (29), by including a delta-functional part into the polarization distribution $\mathbf{P}(\mathbf{r}')$.

electric field also obeys Eq. (1.27), but with the stand-alone charge density. Due to the linear superposition principle, for the *total macroscopic field* **E** of these charges and dipoles, we may write

$$\nabla \cdot \mathbf{E} = \frac{1}{\varepsilon_0}\left(\rho + \rho_{\text{ef}}\right) = \frac{1}{\varepsilon_0}\left(\rho - \nabla \cdot \mathbf{P}\right). \tag{3.31}$$

This is already the main result of the "macroscopic" electrostatics. However, it is evidently tempting (and very convenient for applications) to rewrite Eq. (31) in a different form by carrying the dipole-related term of this equality over to its left-hand side. The resulting formula is called the *macroscopic Maxwell equation for* **D**:

$$\boxed{\nabla \cdot \mathbf{D} = \rho\,,} \tag{3.32}$$

Maxwell equation for **D**

where **D**(**r**) is a new "macroscopic" field, called the *electric displacement* (in some older texts, "electric induction"), defined as[7]

$$\boxed{\mathbf{D} \equiv \varepsilon_0 \mathbf{E} + \mathbf{P}\,.} \tag{3.33}$$

Electric displacement

The comparison of Eqs. (32) and (1.27) shows that **D** (or more strictly, the fraction $\mathbf{D}/\varepsilon_0$) may be interpreted as the "would-be electric field" that *would be* created by stand-alone charges in the absence of the dipole medium polarization. If should be distinguished from the **E** participating in Eqs. (31) and (33), i.e. from the genuine electric field, if averaged over a spatial scale of the order of the distance between elementary charges and dipoles.

In order to get an even better gut feeling of the fields **E** and **D**, let us first rewrite the macroscopic Maxwell equation (32) in the integral form. Applying the divergence theorem to an arbitrary volume V limited by surface S, we get the following *macroscopic Gauss law*:

$$\boxed{\oint_S D_n d^2 r = \int_V \rho\, d^3 r \equiv Q\,,} \tag{3.34}$$

Macroscopic Gauss law

where Q is the *stand-alone* charge inside volume V.

This general result may be used to find the boundary conditions for **D** at a sharp interface between two different dielectrics. (The analysis is applicable to a dielectric/free-space boundary as well.) For that, let us apply Eq. (34) to a flat pillbox formed at the interface (see the solid rectangle in Fig. 5), which is sufficiently small on the spatial scales of the dielectric's nonuniformity and the interface's curvature, but still contains many elementary dipoles. Assuming that the interface does not have stand-alone surface charges, we immediately get

$$\boxed{\left(D_n\right)_1 = \left(D_n\right)_2\,,} \tag{3.35}$$

Boundary condition for **D**

i.e. the normal component of the electric displacement has to be continuous. Note that a similar statement for the macroscopic electric field **E** is generally not valid, because the polarization vector **P** may have, and typically does have a leap at a sharp interface (say, due to the different polarizability of

[7] Note that according to its definition (33), the dimensionality of **D** in the SI units is different from that of **E**. In contrast, in the Gaussian units, the electric displacement is defined as $\mathbf{D} = \mathbf{E} + 4\pi\mathbf{P}$, so $\nabla\cdot\mathbf{D} = 4\pi\rho$ (the relation $\rho_{\text{ef}} = -\nabla\cdot\mathbf{P}$ remains the same as in the SI units), and the dimensionalities of **D** and **E** coincide. This coincidence is a certain perceptional handicap because it is frequently convenient to consider the scalar components of **E** as generalized forces, and those of **D** as generalized coordinates (see Sec. 5 below), and it is somewhat comforting to have their dimensionalities different, as they are in the SI units.

the two different dielectrics), providing a surface layer of the effective charges (30) – see again the example shown in Fig. 4.

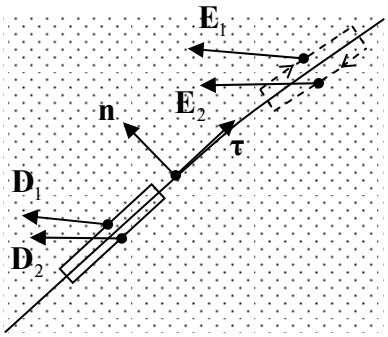


Fig. 3.5. Deriving the boundary conditions at an interface between two dielectrics, using a Gauss pillbox (shown as a solid-line rectangle) and a circulation contour (dashed-line rectangle). Here **n** and **τ** are the unit vectors that are, respectively, normal and tangential to the interface. Note that due to the leap of polarization, the field lines are generally "refracted" at the interface – see Fig. 11b for an example.

However, we still can make an important statement about the behavior of **E** at the interface. Indeed, the macroscopic electric fields defined by Eqs. (29) and (31), are evidently still potential ones, and hence obey the macroscopic Maxwell equation similar to Eq. (1.28):

Macroscopic Maxwell equation for **E**

$$\nabla \times \mathbf{E} = 0 . \tag{3.36}$$

Integrating this equality along a narrow contour stretched along the interface (see the dashed rectangle in Fig. 5), we get

Boundary condition for **E**

$$\left(E_\tau\right)_1 = \left(E_\tau\right)_2 . \tag{3.37}$$

Note that this condition is compatible with (and may be derived from) the continuity of the macroscopic electrostatic potential ϕ related to the macroscopic field **E** by the relation similar to Eq. (1.33), $\mathbf{E} = -\nabla\phi$, at each point of the interface: $\phi_1 = \phi_2$.

In order to see how these boundary conditions work, let us consider the simple problem shown in Fig. 6. A very broad plane capacitor, with zero voltage between its conducting plates (as may be enforced, for example, by their connection with an external wire), is partly filled with a material with a uniform polarization $\mathbf{P}_0$,[8] oriented normal to the plates. Let us calculate the spatial distribution of the fields **E** and **D**, and also the surface charge density of each conducting plate.

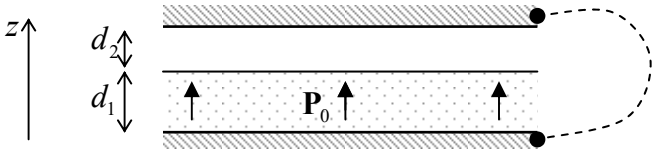


Fig. 3.6. A simple system whose analysis requires Eq. (35).

Due to the symmetry of the system, the vectors **E** and **D** are both normal to the plates and do not depend on the position in the capacitor's plane, so we can limit the fields' analysis to the calculation of their z-components $E(z)$ and $D(z)$. In this case, the Maxwell equation (32) is reduced to $dD/dz = 0$ inside each layer (but not at their border!), so within each of them, D is constant – say, some D_1 in the layer with $\mathbf{P} = \mathbf{P}_0$, and certain D_2 in the free-space layer, where $\mathbf{P} = 0$. As a result, according to Eq. (33), the (macroscopic) electric field inside each layer is also constant:

[8] As will be discussed in the next section, this is a good approximation for the so-called *electrets*, and also for *hard ferroelectrics* in not very high electric fields.

$$D_1 = \varepsilon_0 E_1 + P_0, \qquad D_2 = \varepsilon_0 E_2 . \tag{3.38}$$

Since the voltage between the plates is zero, we may also require the integral of E, taken along a path connecting the plates, to vanish. This gives us one more relation:

$$E_1 d_1 + E_2 d_2 = 0. \tag{3.39}$$

Still, the three equations (38)-(39) are insufficient to calculate the four fields in the system ($E_{1,2}$ and $D_{1,2}$). The decisive help comes from the boundary condition (35):

$$D_1 = D_2 . \tag{3.40}$$

(Note that it is valid because the layer interface does not carry *stand-alone* electric charges, even though it has a *polarization surface charge*, whose areal density may be calculated by integrating Eq. (30) across the interface: $\sigma_{ef} = P_0$. Note also that in our simple system, Eq. (37) is identically satisfied due to the system's symmetry, and hence does not give any additional information.)

Now solving the resulting system of four equations (38)-(40), we readily get

$$E_1 = -\frac{P_0}{\varepsilon_0}\frac{d_2}{d_1 + d_2}, \qquad E_2 = \frac{P_0}{\varepsilon_0}\frac{d_1}{d_1 + d_2}, \qquad D_1 = D_2 = D = P_0 \frac{d_1}{d_1 + d_2}. \tag{3.41}$$

The areal densities of the electrode surface charges may now be readily calculated by the integration of Eq. (32) across each surface:

$$\sigma_1 = -\sigma_2 = D = P_0 \frac{d_1}{d_1 + d_2}. \tag{3.42}$$

Note that due to the spontaneous polarization of the lower layer's material, the capacitor plates are charged even in the absence of voltage between them and that this charge is a function of the second electrode's position (d_2).[9] Also notice a substantial similarity between this system (Fig. 6), and the one whose analysis was the subject of Problem 2.6.

3.3. Polarization of dielectrics

The general relations derived in the previous section may be used to describe the electrostatics of any dielectrics – materials with bound electric charges (and hence with negligible dc electric conduction). However, to form a full system of equations necessary to solve electrostatics problems, they have to be complemented by certain constitutive relations between the vectors **P** and **E**.[10]

In most materials, in the absence of an external electric field, the elementary dipoles **p** either equal zero or have a random orientation in space, so the net dipole moment of each macroscopic volume

[9] This effect is used in most modern microphones. In such a device, the sensed sound wave's pressure bends a thin conducting membrane playing the role of one of the capacitor's plates, and thus modulates the thickness (in Fig. 6, d_2) of the air gap adjacent to the electret layer. This modulation produces proportional variations of the charges (42), and hence the corresponding electric current flowing between the plates, which is picked up by readout electronics. According to J. West (who, together with G. Sessler, invented the electret microphone in 1962), currently more than 2 billion of these devices are fabricated each year.

[10] In the problem solved at the end of the previous section, the role of such relation was played by the equality $\mathbf{P}_0 = \text{const}$.

(still containing many such dipoles) equals zero: $\mathbf{P} = 0$ at $\mathbf{E} = 0$. Moreover, if the field changes are sufficiently slow, most materials may be characterized by a unique dependence of $\mathbf{P}$ on $\mathbf{E}$. Then using the Taylor expansion of function $\mathbf{P}(\mathbf{E})$, we may argue that in relatively low electric fields the function should be well approximated by a linear dependence between these two vectors. Such dielectrics are called *linear* (or "simple"). In an isotropic media, the coefficient of proportionality should be just a scalar. In the SI units, this scalar is defined by the following relation:

Electric susceptibility

$$\mathbf{P} = \chi_e \varepsilon_0 \mathbf{E}\,, \tag{3.43}$$

with the dimensionless constant χ_e called the *electric susceptibility*. However, it is much more common to use, instead of χ_e, another dimensionless parameter,[11]

Dielectric constant

$$\kappa \equiv 1 + \chi_e\,, \tag{3.44}$$

which is sometimes called the "relative electric permittivity", but much more often, the *dielectric constant*. This parameter is very convenient, because combining Eqs. (43) and (44),

$$\mathbf{P} = (\kappa - 1)\varepsilon_0 \mathbf{E}. \tag{3.45}$$

and then plugging the resulting relation into the general Eq. (33), we get simply

$$\mathbf{D} = \kappa\varepsilon_0 \mathbf{E}, \qquad \text{or} \quad \mathbf{D} = \varepsilon \mathbf{E}, \tag{3.46}$$

where another popular parameter,[12]

Electric permittivity

$$\varepsilon \equiv \kappa\varepsilon_0 \equiv (1 + \chi_e)\varepsilon_0\,. \tag{3.47}$$

ε is called the *electric permittivity* of the material.[13] Table 1 gives the approximate values of the dielectric constant for several representative materials.

In order to understand the range of these values, let me discuss (briefly and rather superficially[14]) the two simplest mechanisms of electric polarization. The first of them is typical for liquids and gases of *polar* atoms/molecules, which have their own, spontaneous dipole moments $\mathbf{p}$. (A typical example is the water molecule H_2O, with the negative oxygen ion offset from the line connecting two positive hydrogen ions, thus producing a spontaneous dipole moment $p = ea$, with $a \approx 0.38\times10^{-10}$m $\sim r_B$.) In the absence of an external electric field, the orientation of such dipoles may be random, with the average polarization $\mathbf{P} = n\langle\mathbf{p}\rangle$ equal to zero – see the top panel of Fig. 7a.

---

[11] In older physics literature, the dielectric constant is often denoted by the letter ε_r (with the index "r" meaning "relative"), while in electrical engineering publications, its notation is frequently K.

[12] The reader may be perplexed by the use of three different but uniquely related parameters (χ_e, $\kappa \equiv 1 + \chi_e$, and $\varepsilon \equiv \kappa\varepsilon_0$) for the description of just one scalar property. Unfortunately, such redundancy is typical for physics, whose different sub-field communities have different, well-entrenched traditions.

[13] In the Gaussian units, χ_e is defined by the following relation: $\mathbf{P} = \chi_e \mathbf{E}$, while ε is defined just as in the SI units, $\mathbf{D} = \varepsilon\mathbf{E}$. Because of that, in the Gaussian units, the constant ε is dimensionless and equals $(1 + 4\pi\chi_e)$. As a result, $\varepsilon_{Gaussian} = (\varepsilon/\varepsilon_0)_{SI} \equiv \kappa$, so $(\chi_e)_{Gaussian} = (\chi_e)_{SI}/4\pi$, sometimes creating confusion between the numerical values of the latter parameter – dimensionless in both systems.

[14] While I believe this discussion is very useful, it is quantitatively valid only for relatively sparse media, with low concentration ($n << 1/a^3$) of elementary atomic/molecular dipoles of size scale a. Indeed, in some condensed materials, with $na^3 \sim 1$, even the notion of the dipole moment $\mathbf{p}$ with a single atomic cell is ambiguous.

Table 3.1. Dielectric constants of a few representative (and/or practically important) dielectrics

Material	κ
Air (at ambient conditions)	1.00054
Teflon (polytetrafluoroethylene, $[C_2F_4]_n$)	2.1
Silicon dioxide (amorphous)	3.9
Glasses (of various compositions)	3.7–10
Castor oil	4.5
Silicon[(a)]	11.7
Water (at 100°C)	55.3
Water (at 20°C)	80.1
Barium titanate ($BaTiO_3$, at 20°C)	~1,600

[(a)] Anisotropic materials, such as silicon crystals, require a *susceptibility tensor* to give an exact description of the linear relation of the vectors **P** and **E**. However, most important crystals (including Si) are only weakly anisotropic, so they may be reasonably well characterized with a scalar (angle-average) susceptibility.

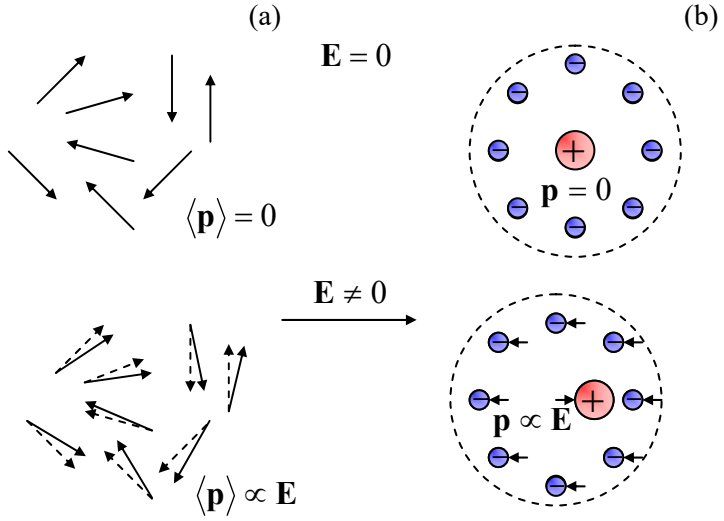


Fig. 3.7. Crude cartoons of two mechanisms of the induced electrical polarization: (a) a partial ordering of spontaneous elementary dipoles, and (b) an elementary dipole induction. The upper two panels correspond to $\mathbf{E} = 0$, and the lower two panels, to $\mathbf{E} \neq 0$.

A relatively weak external field does not change the magnitude of the dipole moments significantly, but according to Eqs. (15a) and (17), tries to orient them along the field, creating a non-zero vector average $\langle \mathbf{p} \rangle$ directed along the vector $\mathbf{E}_m$, where $\mathbf{E}_m$ is the microscopic field at the point of the dipole's location – cf. two panels of Fig. 7a. If the field is not two high ($p\langle E_m \rangle << k_B T$), the induced average polarization $\langle \mathbf{p} \rangle$ is proportional to $\mathbf{E}_m$. If we write this proportionality relation in the following traditional form,

$$\langle \mathbf{p} \rangle = \alpha \mathbf{E}_m , \tag{3.48}$$

Atomic polarizability

where α is called the *atomic* (or, sometimes, "molecular") *polarizability*, this means that α is positive. If the concentration n of such elementary dipoles is low, the contribution of their own fields into the

microscopic field acting on each dipole is negligible, and we may identify $\mathbf{E}_m$ with the macroscopic field $\mathbf{E}$. As a result, the second of Eqs. (27) yields

$$\mathbf{P} \equiv n\langle \mathbf{p} \rangle = \alpha n \mathbf{E} . \tag{3.49}$$

Comparing this relation with Eq. (45), we get

$$\kappa = 1 + \frac{\alpha n}{\varepsilon_0}, \tag{3.50}$$

so $\kappa > 1$ (i.e. $\chi_e = \alpha n/\varepsilon_0 > 0$). Note that at this particular polarization mechanism (illustrated on the lower panel of Fig. 7a), the thermal motion "tries" to randomize the dipole orientation, i.e. reduce its ordering by the field, so we may expect α, and hence $\chi_e \equiv \kappa - 1$ to increase as temperature T is decreased – the so-called *paraelectricity*. Indeed, the basic statistical mechanics[15] shows that in this case, the electric susceptibility follows the so-called *Curie law* $\chi_e \propto 1/T$.

The materials of the second, much more common class consist of *non-polar* atoms without intrinsic spontaneous polarization. A crude classical image of such an atom is an isotropic cloud of negatively charged electrons surrounding a positively charged nucleus – see the top panel of Fig. 7b. The external electric field shifts the positive charge in the direction of the vector $\mathbf{E}$, and the negative charges in the opposite direction, thus creating a similarly directed average dipole moment $\langle \mathbf{p} \rangle$.[16] At relatively low fields, this average moment is proportional to $\mathbf{E}$, so we again arrive at Eq. (48), with $\alpha > 0$, and if the dipole concentration n is sufficiently low, also at Eq. (50), with $\kappa - 1 > 0$. So, the dielectric constant is larger than 1 for both polarization mechanisms – please have one more look at Table 1.

In order to make a crude but physically transparent estimate of the difference $\kappa - 1$, let us consider the following toy model of a non-polar dielectric: a set of similar conducting spheres of radius R, distributed in space with a low density $n << 1/R^3$. At such density, the electrostatic interaction of the spheres is negligible, and we can use Eq. (11) for the induced dipole moment of a single sphere. Then the polarizability definition (48) yields $\alpha = 4\pi\varepsilon_0 R^3$, so Eq. (50) gives

$$\kappa = 1 + 4\pi R^3 n . \tag{3.51}$$

Let us use this result for a crude estimate of the dielectric constant of air at the so-called *ambient conditions*, meaning the normal atmospheric pressure $\mathcal{P} = 1.013\times10^5$ Pa and temperature $T = 300$ K. At these conditions the molecular density n may be, with a few-percent accuracy, found from the well-known equation of state of an ideal gas:[17] $n \approx \mathcal{P}/k_B T \approx (1.013\times10^5)/(1.38\times10^{-23}\times300) \approx 2.45\times10^{25}$ m^{-3}. The molecule of the air's main component, N_2, has a van-der-Waals radius[18] of 1.55×10^{-10} m. Taking this radius for the R of our crude model, we get $\chi_e \equiv \kappa - 1 \approx 1.15\times10^{-3}$. Comparing this number with the

[15] See, e.g., SM Chapter 2.

[16] Realistically, these effects are governed by quantum mechanics, so the average here should be understood not only in the statistical-mechanical but also (and mostly) in the quantum-mechanical sense. Because of that, for non-polar atoms, α is typically a very weak function of temperature, at least on the usual scale $T \sim 300$K.

[17] If needed, see, e.g., SM Secs. 1.4 and 3.1.

[18] Such radius is defined by the requirement that the volume of the corresponding sphere, if used in the van-der-Waals equation (see, e. g., SM Sec. 4.1), gives the best fit to the experimental equation of state $n = n(P, T)$.

first line of Table 1, we see that the model gives a surprisingly reasonable result: to get the experimental value, it is sufficient to decrease the effective R of the sphere by just ~30%, to ~1.2×10^{-10} m.[19]

This result may encourage us to try using Eq. (51) for a larger density n. For example, as a crude model for a non-polar crystal, let us assume that the conducting spheres form a simple cubic lattice with the period $a = 2R$ (i.e., the neighboring spheres virtually touch). With this, $n = 1/a^3 = 1/8R^3$ and Eq. (44) yields $\kappa = 1 + 4\pi/8 \approx 2.5$. This estimate provides a reasonable semi-qualitative explanation for the values of κ listed in a few middle rows of Table 1. However, at such small distances, the electrostatic dipole-dipole interaction should be already essential, so this simple model cannot even approximately describe the values of κ much larger than 1, listed in the last rows of the table.

Such high values may be explained by the so-called *molecular field effect*: each elementary dipole is polarized not only by the external field, as Eq. (49) assumes, but by the field of neighboring dipoles as well. Ottavino-Fabrizio Mossotti in 1850 and (almost 30 years later) Rudolf Clausius suggested what is now known, rather unfairly, as the *Clausius-Mossotti formula*,[20] which describes this effect reasonably well in many non-polar materials. In our notation, it reads[21]

$$\frac{\kappa-1}{\kappa+2}=\frac{\alpha n}{3\varepsilon_0},\qquad \text{so } \kappa=1+\frac{\alpha n/\varepsilon_0}{1-\alpha n/3\varepsilon_0}. \tag{3.52}$$

Clausius-Mossotti formula

If the dipole density is low in the sense $n << \varepsilon_0/\alpha$, this relation is reduced to Eq. (50) corresponding to independent dipoles. However, at higher dipole density, κ and hence $\chi_e \equiv \kappa - 1$ increase faster and tend to infinity as the density-polarizability product approaches some critical value n_c, equal to $3\varepsilon_0/\alpha$ in the Clausius-Mossotti approximation.[22] This means that the zero-polarization state becomes unstable even in the absence of an external electric field.

This instability is a linear-theory (i.e. low-field) manifestation of a substantially nonlinear effect – the formation, in some materials, of spontaneous polarization even in the absence of an external electric field. Such materials are called *ferroelectrics*, and may be experimentally recognized by the hysteretic behavior of their polarization as a function of the applied (external) electric field – see Fig. 8. As the plots show, the polarization of a ferroelectric depends on the applied field's history. For example, the direction of its spontaneous *remnant polarization* P_R may be switched by first applying, and then removing a sufficiently high field (larger than the so-called *coercive field* E_C – see Fig. 8) of the opposite orientation. The physics of this switching is rather involved; the polarization vector **P** of a ferroelectric material is typically constant only within each of the spontaneously formed spatial regions (called *domains*), with a typical size of a few tenths of a micron, and different (frequently, opposite) directions of the vector **P** in adjacent domains. The change of the applied electric field results not in the

[19] As will be discussed in QM Chapter 6, for a hydrogen atom in its ground state, the low-field polarizability may be calculated analytically: $\alpha = (9/2)\times4\pi\varepsilon_0 r_B^3$, corresponding to our metallic-ball model with a close value of the effective radius: $R = (9/2)^{1/3} r_B \approx 1.65\, r_B \approx 0.87\times10^{-10}$ m.

[20] Applied to the high-frequency electric field, with κ replaced by the square of the refraction coefficient at the field's frequency (see Chapter 7), this formula is known as the *Lorenz-Lorentz relation*.

[21] The proof of Eq. (52), by using Eq. (24) for the molecular field's evaluation, is left for the reader's exercise.

[22] The Clausius-Mossotti formula does not give quantitatively correct results for many condensed materials, notably including ferroelectrics. For a review of modern approaches to the theory of their polarization, see, e.g., the paper by R. Resta and D. Vanderbilt in the review collection by K. Rabe, C. Ahn, and J.-M. Triscone (eds.), *Physics of Ferroelectrics: A Modern Perspective*, Springer, 2010.

switching of the direction of **P** inside each domain, but rather in a shift of the domain walls, resulting in the change of the average polarization of the sample.

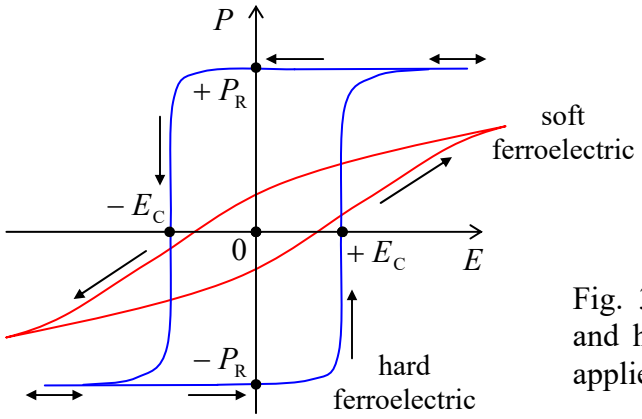


Fig. 3.8. The average polarization of soft and hard ferroelectrics as functions of the applied electric field (schematically).

Depending on the ferroelectric's material, temperature, and the sample's geometry (a solid crystal, a ceramic material, or a thin film), the hysteretic loops may be rather different, ranging from a rather smooth form in the so-called *soft ferroelectrics* (which include most ferroelectric thin films) to an almost rectangular form in *hard ferroelectrics* – see Fig. 8. In low fields, soft ferroelectrics behave essentially as linear paraelectrics, but with a very high average dielectric constant – see the bottom line of Table 1 for such a classical material as $BaTiO_3$ (which is a soft ferroelectric at temperatures below $T_c \approx 120°C$, and a paraelectric above this critical temperature). On the other hand, the polarization of a hard ferroelectric in the fields below its coercive field remains virtually constant, and the analysis of their electrostatics may be based on the condition $\mathbf{P} = \mathbf{P}_R = \text{const}$ – already used in the problem discussed in the end of the previous section.[23] This condition is even more applicable to the so-called *electrets* – synthetic polymers with a spontaneous polarization that remains constant even in very high electric fields.

Some materials exhibit even more complex polarization effects, for example, *antiferroelectricity*, *helielectricity*, and (practically very valuable) *piezoelectricity*. Unfortunately, I do not have time for a discussion of these exotic phenomena in this course;[24] the main reason I am mentioning them is to emphasize again that the constitutive relation $\mathbf{P} = \mathbf{P}(\mathbf{E})$ is material-specific rather than fundamental. However, most insulators, in practicable fields, behave as linear dielectrics, so the next section will be committed to the discussion of their electrostatics.

---

[23] Due to this property, hard ferroelectrics, such as the lead zirconate titanate (PZT) and strontium bismuth tantalite (SBT), with high remnant polarization P_R (up to ~1 C/m^2), may be used in nonvolatile random-access memories (dubbed either FRAM or FeRAM) – see, e.g., J. Scott, *Ferroelectric Memories*, Springer, 2000. In a cell of such a memory, binary information is stored in the form of one of two possible directions of spontaneous polarization at $\mathbf{E} = 0$ (see Fig. 8). Unfortunately, the time of spontaneous depolarization of ferroelectric thin films is typically well below 10 years – the industrial standard for data retention in nonvolatile memories, and this time may be decreased even more by "fatigue" from the repeated polarization recycling at information recording. Due to these reasons, the industrial production of FRAM is currently just a tiny fraction of the nonvolatile memory market, which is dominated by floating-gate memories – see, e.g., Sec. 4.2 below.

[24] For detailed coverage of ferroelectrics, I can recommend the encyclopedic monograph by M. Lines and A. Glass, *Principles and Applications of Ferroelectrics and Related Materials*, Oxford U. Press, 2001, and the recent review collection edited by K. Rabe *et al*., that was cited above.

3.4. Electrostatics of linear dielectrics

First, let us consider the simplest but very important problem: how is the electrostatic field of a set of stand-alone charges of density $\rho(\mathbf{r})$ modified if it is placed into a uniform linear dielectric medium that obeys Eq. (46) with a dielectric constant κ constant in the whole region we are interested in. In this case, we may combine Eqs. (32) and (46) to write

$$\nabla \cdot \mathbf{E} = \frac{\rho}{\varepsilon}. \tag{3.53}$$

As a reminder, in the free space, we had a similar equation (1.27), but with a different constant, $\varepsilon_0 = \varepsilon/\kappa$. Hence all the results discussed in Chapter 1 are valid inside a uniform linear dielectric, for the macroscopic field the $\mathbf{E}$ (and the corresponding macroscopic electrostatic potential ϕ), if they are reduced by the factor of $\kappa > 1$. Thus, the most straightforward result of the induced polarization of a dielectric medium is the *electric field reduction*. This is a very important effect, especially taking into account the very high values of κ in such common dielectrics as water – see Table 1. Indeed, it is the reduction of the attraction between positive and negative ions (called, respectively, *cations* and *anions*) in water that enables their substantial dissociation and hence almost all biochemical reactions, which are the basis of the biological cell functions – and hence of the life itself.

Let us apply this general result to the important particular case of the plane capacitor (Fig. 2.3) filled with a linear, uniform dielectric. Applying the macroscopic Gauss law (34) to a pillbox-shaped volume on the conductor surface, we get the following relation,

$$\sigma = D_n = \varepsilon E_n = -\varepsilon \frac{\partial \phi}{\partial n}, \tag{3.54}$$

which differs from Eq. (2.3) only by the replacement $\varepsilon_0 \to \varepsilon \equiv \kappa\varepsilon_0$. Hence, for a fixed field E_n, the charge density calculated for the free-space case should be increased by the factor of κ – that's it. In particular, this means that the capacitance (2.28) has to be increased by this factor:

$$C = \frac{\kappa\varepsilon_0 A}{d} \equiv \frac{\varepsilon A}{d}. \tag{3.55}$$

C of a planar capacitor

(As a reminder, this increase of C by κ has been already incorporated, without proof, into some estimates made in Secs. 2.1 and 2.2, to make them realistic.)

If a linear dielectric is nonuniform, the situation is more complex. For example, let us consider the case of a sharp interface between two otherwise uniform dielectrics, free of stand-alone charges. In this case, we still may use Eq. (37) for the tangential component of the macroscopic electric field, and also Eq. (36), with $D_n = \varepsilon E_n$, for its normal component, getting

$$\left(\varepsilon E_n\right)_1 = \left(\varepsilon E_n\right)_2, \qquad \text{i.e. } \varepsilon_1 \frac{\partial \phi_1}{\partial n} = \varepsilon_2 \frac{\partial \phi_2}{\partial n}. \tag{3.56}$$

Boundary condition for E_n

Let us apply these boundary conditions, first of all, to consider how carving a slit of some width d and a much smaller thickness $t << d$ from inside a dielectric, changes an initially uniform electric field $\mathbf{E}_0$, depending on its orientation – see Fig. 9.

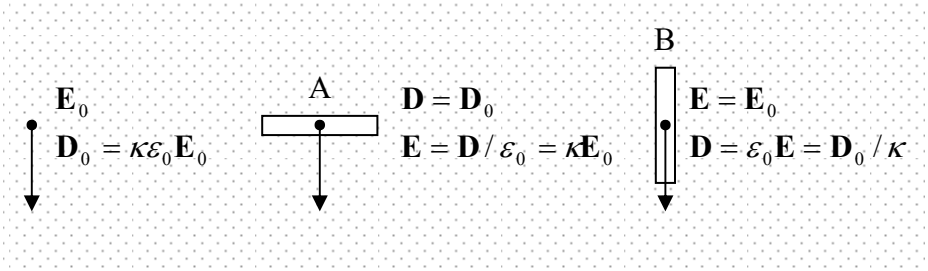


Fig. 3.9. Fields inside two narrow slits cut in a linear dielectric.

First of all, intuition tells us that regardless of its orientation, a slit cannot change the field far from it; moreover, at $t \to 0$, it cannot modify substantially even the field right outside its "major" (broader) surfaces. This conclusion may be supported either by direct calculations (see, e.g., the problem illustrated by Fig. 11 below), or by energy arguments: at $t << d$, any potential energy decrease due to the field change inside the slit's volume (proportional to td) cannot compensate its increase in the outer volume proportional to d^2. However, it may induce some local field changes – inside the slit, and even outside it, close to its "minor" surfaces.

To calculate the inner field for case A, with the slit's plane normal to the applied field, we may apply Eq. (56) to its major surfaces (shown horizontal), to prove that the vector **D** should be continuous. But according to Eq. (46), this means that in the free space inside the slit, the electric field should equal $\mathbf{D}/\varepsilon_0$, and hence be κ times higher than the field $\mathbf{E}_0 = \mathbf{D}/\kappa\varepsilon_0$ far from the slit. This field, and hence **D**, may be measured by a sensor placed inside the gap, so the electric displacement is not an entirely mathematical construct.[25] On the contrary, for case B, with the slit's plane parallel to the initial field, we may apply Eq. (37) to the major (now, vertical) interfaces of the slit, to see that now the electric field **E** is continuous, while the electric displacement $\mathbf{D} = \varepsilon_0\mathbf{E}$ inside the gap is a factor of κ lower than its value in the dielectric. (Similarly to case A, any perturbations of the field uniformity, caused by the compliance with Eq. (56) at the minor surfaces, settle down at distances ~t from them.)

For other problems with piecewise-constant ε, with more complex geometries, we may need to apply the methods studied in Chapter 2. In particular, in the simplest cases, we can select such a set of orthogonal coordinates that the electrostatic potential depends on just one of them. Consider, for example, two types of filling a plane capacitor with two different dielectrics – see Fig. 10.

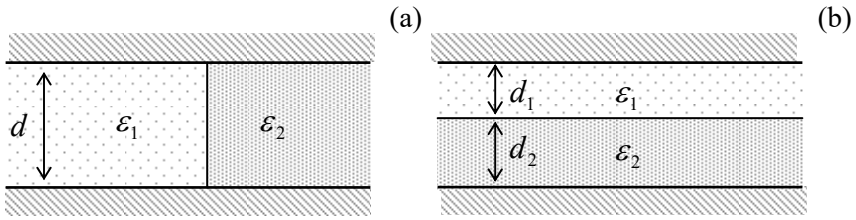


Fig. 3.10. Plane capacitors filled with two different dielectrics.

In case (a), the voltage V between the electrodes is the same for each part of the capacitor, telling us that at least far from the dielectric interface, the electric field is vertical, uniform, and constant ($E = V/d$). Hence the boundary condition (37) is satisfied even if such a distribution is valid near the surface

[25] Superficially, this result violates the boundary condition (37) at the vertical ("minor") surfaces of the gap. This apparent contradiction is resolved by the fact the thin slit can deform the field both inside and outside it, at distances of the order of t around these interfaces, but not far beyond them, so the above relations for **E** and **D** are valid at most of the slit area.

as well, i.e. at any point of the system. The only effect of different values of ε in the two parts is that the electric displacement $D = \varepsilon E$ and hence electrodes' surface charge density $\sigma = D$ are different in them. Thus we can calculate the electrode charges $Q_{1,2}$ of the two parts independently, and then add up the results to get the total mutual capacitance

$$C = \frac{Q_1 + Q_2}{V} = \frac{1}{d}\left(\varepsilon_1 A_1 + \varepsilon_2 A_2\right). \tag{3.57}$$

Note that this formula may be interpreted as the total capacitance of two separate lumped capacitors connected (by wires) *in parallel*. This is natural, because we may cut the system along the dielectric interface, without any effect on the fields in either part, and then connect the corresponding electrodes by external wires, again without any effect on the system – besides very close vicinities of the capacitor's edges, where the fringe

Case (b) may be analyzed just as in the problem illustrated by Fig. 6, by applying Eq. (34) to a Gaussian pillbox with one lid inside the (for example) bottom electrode, and the other lid inside any of the layers. As a result, we see that D anywhere inside the system should be equal to the surface charge density σ of the electrode, i.e. constant. Hence, according to Eq. (46), the electric field E inside each dielectric layer is also constant: in the top layer, it is $E_1 = D_1/\varepsilon_1 = \sigma/\varepsilon_1$, while in bottom layer, $E_2 = D_2/\varepsilon_2 = \sigma/\varepsilon_2$. Integrating the field E across the whole capacitor, we get

$$V = \int_0^{d_1+d_2} E(z)dz = E_1 d_1 + E_2 d_2 = \left(\frac{d_1}{\varepsilon_1} + \frac{d_2}{\varepsilon_2}\right)\sigma, \tag{3.58}$$

so the mutual capacitance per unit area

$$\frac{C}{A} \equiv \frac{\sigma}{V} = \left[\frac{d_1}{\varepsilon_1} + \frac{d_2}{\varepsilon_2}\right]^{-1}. \tag{3.59}$$

Note that this result is similar to the total capacitance of an *in-series* connection of two plane capacitors based on each of the layers. This is also natural because we could insert an uncharged, thin conducting sheet (rather than a cut as in the previous case) at the layer interface, which is an equipotential surface, without changing the field distribution in any part of the system. Then we could thicken the conducting sheet as much as we liked (and possibly shape its internal part into a thin wire), also without changing the fields in the dielectric parts of the system, and hence the capacitance.

Proceeding to problems with more complex geometry, let us consider the system shown in Fig. 11a: a dielectric sphere placed into an initially uniform external electric field $\mathbf{E}_0$. According to Eq. (53) for the macroscopic electric field, and the definition of the macroscopic electrostatic potential, $\mathbf{E} = -\nabla\phi$, the potential satisfies the Laplace equation both inside and outside the sphere, though not at its border. Due to the spherical symmetry of the dielectric sample, this problem invites the variable separation method in spherical coordinates, which was discussed in Sec. 2.8. From that discussion, we already know, in particular, the general solution (2.172) of the Laplace equation outside of the sphere. To satisfy the uniform-field condition at $r \to \infty$, we have to reduce this solution to

$$\phi_{r\geq R} = -E_0 r\cos\theta + \sum_{l=1}^{\infty} \frac{b_l}{r^{l+1}} \mathcal{P}_l(\cos\theta). \tag{3.60}$$

Inside the sphere, we can also use Eq. (2.172), but keeping only the radial functions finite at $r \to 0$:

$$\phi_{r\le R} = \sum_{l=1}^{\infty} a_l r^l \mathcal{P}_l(\cos\theta)\,. \tag{3.61}$$

Now, spelling out the boundary conditions (37) and (56) at $r = R$, we see that for all coefficients a_l and b_l with $l \ge 2$, we get homogeneous linear equations (just like for the conducting sphere discussed in Sec. 2.8) that have only trivial solutions. Hence, all these terms may be dropped, while for the only surviving terms with $l = 1$, proportional to the Legendre polynomial $\mathcal{P}_1(\cos\theta) \equiv \cos\theta$, we get two equations:

$$-E_0 - \frac{2b_1}{R^3} = \kappa a_1, \qquad -E_0 R + \frac{b_1}{R^2} = a_1 R. \tag{3.62}$$

Solving this simple system of linear equations for a_1 and b_1, and plugging the result into Eqs. (60) and (61), we get the final solution of the problem:

$$\phi_{r\ge R} = E_0\left(-r + \frac{\kappa-1}{\kappa+2}\frac{R^3}{r^2}\right)\cos\theta, \qquad \phi_{r\le R} = -E_0\frac{3}{\kappa+2} r\cos\theta\,. \tag{3.63}$$

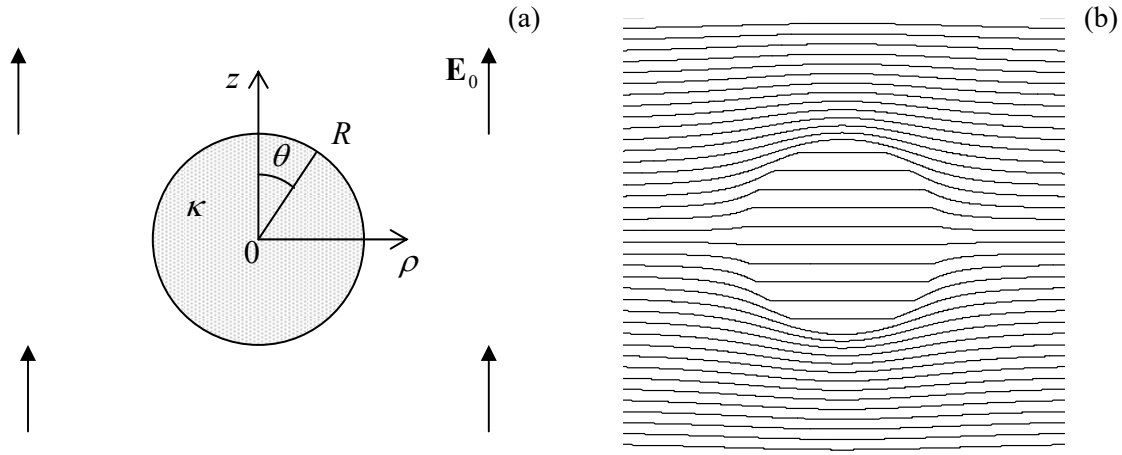


Fig. 3.11. A dielectric sphere in an initially uniform electric field: (a) the problem, and (b) the equipotential surfaces, as given by Eq. (63), for $\kappa = 3$.

Figure 11b shows the equipotential surfaces given by this solution, for a particular value of the dielectric constant κ. Note that according to Eq. (62), at $r \ge R$ the dielectric sphere, just as the conducting sphere in a similar problem, produces (on top of the uniform external field) a pure dipole field, with the dipole moment

$$\mathbf{p} = 4\pi R^3 \frac{\kappa-1}{\kappa+2}\varepsilon_0\mathbf{E}_0 \equiv 3V\frac{\kappa-1}{\kappa+2}\varepsilon_0\mathbf{E}_0, \qquad \text{where } V = \frac{4\pi}{3}R^3. \tag{3.64}$$

This is an evident generalization of Eq. (11), to which Eq. (64) tends at $\kappa \to \infty$. By the way, this property is common: for their electrostatic properties, conductors may be adequately described as dielectrics with $\kappa \to \infty$.

Another remarkable feature of Eqs. (63) is that the electric field and polarization inside the sphere are uniform, with R-independent values

$$\mathbf{E} = \frac{3}{\kappa+2}\mathbf{E}_0, \qquad \mathbf{D} \equiv \kappa\varepsilon_0\mathbf{E} = \varepsilon_0\frac{3\kappa}{\kappa+2}\mathbf{E}_0, \qquad \mathbf{P} \equiv \mathbf{D} - \varepsilon_0\mathbf{E} = 3\varepsilon_0\frac{\kappa-1}{\kappa+2}\mathbf{E}_0. \tag{3.65}$$

In the limit $\kappa \to 1$ (for example, the "sphere made of free space", i.e. no sphere at all), the electric field inside it naturally tends to the external one, and its polarization vanishes. In the opposite limit $\kappa \to \infty$, the electric field inside the sphere vanishes. Curiously enough, in this limit the electric displacement inside the sphere remains finite: $\mathbf{D} \to 3\varepsilon_0\mathbf{E}_0$.

More complex problems with piecewise-uniform dielectrics also may be addressed by the methods discussed in Chapter 2, and hopefully, the reader will be able to use them to solve a few such problems offered in Sec. 6, on their own. Let me discuss just one of such problems because it exhibits a new feature of the charge image method that was discussed in Secs 2.9 (and is the basis of Green's function approach – see Sec. 2.10). Consider the system shown in Fig. 12: a point charge near a dielectric half-space; it obviously parallels the system discussed in Sec. 2.9 – see Fig. 2.26.

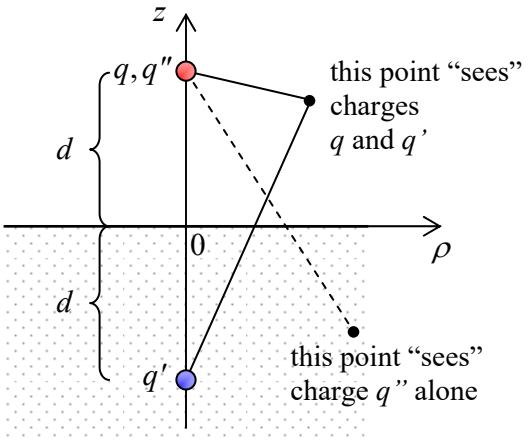


Fig. 3.12. Charge images for a dielectric half-space.

As for the case of a conducting half-space, the Laplace equation for the electrostatic potential in the upper half-space $z > 0$ (besides the charge point $\rho = 0$, $z = d$) may be satisfied using a single image charge q' at the point with $\rho = 0$ and $z = -d$, but now q' may differ from $(-q)$. In addition, in contrast to the case analyzed in Sec. 2.9, we should also calculate the field inside the dielectric (at $z \le 0$). This field cannot be contributed by the image charge q', because that would give a potential divergence at its location. Thus, in the dielectric-filled half-space we should try to use the real point source only, but with a re-normalized charge q'' rather than the genuine charge q – see Fig. 12. As a result, we may look for the potential distribution in the form

$$\phi(\rho,z) = \frac{1}{4\pi\varepsilon_0}\times\begin{cases}\left[\dfrac{q}{\left(\rho^2+(z-d)^2\right)^{1/2}} + \dfrac{q'}{\left(\rho^2+(z+d)^2\right)^{1/2}}\right], & \text{for } z \ge 0,\\ \dfrac{q''}{\left(\rho^2+(z-d)^2\right)^{1/2}}, & \text{for } z \le 0,\end{cases} \tag{3.66}$$

at this stage of solution, with unknown q' and q''. Plugging this equality into the boundary conditions (37) and (56) at $z = 0$ (with $\partial/\partial n = \partial/\partial z$), we see that they are indeed satisfied (so Eq. (66) does express the solution of the boundary problem), provided that the effective charges q' and q'' obey the following relations:

$$q - q' = \kappa q'', \qquad q + q' = q'' . \tag{3.67}$$

Solving this simple system of linear equations, we get

$$q' = -\frac{\kappa - 1}{\kappa + 1} q, \qquad q'' = \frac{2}{\kappa + 1} q . \tag{3.68}$$

If $\kappa \to 1$, then $q' \to 0$, and $q'' \to q$ – both facts very natural because in this limit (no polarization at all) we have to recover the unperturbed field of the initial point charge in both semi-spaces. In the opposite limit $\kappa \to \infty$ (which, as was discussed above, may describe a conducting half-space), $q' \to -q$ (repeating the result we have discussed in detail in Sec. 2.9), and $q'' \to 0$. The last result means that in this limit, the electric field **E** in the dielectric tends to zero – as it should.

In conclusion of this section, please note that if the permittivity ε of a linear dielectric is a continuous rather than piecewise function of coordinates, the distribution of the electrostatic potential ϕ may be found from Eq. (32) with the electric displacement given by Eq. (46): $\mathbf{D} = \varepsilon(\mathbf{r})\mathbf{E} = -\varepsilon(\mathbf{r})\nabla\phi$. However, analytical solutions of the resulting partial differential equation of the second order may be found only for rare particular cases; one of them is offered in Sec. 6 for the reader's exercise.

3.5. Electric field energy in a dielectric

In Chapter 1, we have obtained two key results for the electrostatic energy: Eq. (1.55) for a charge interaction with an independent ("external") field, and a similarly structured formula (1.60), but with an additional factor ½, for the field induced by the charges under consideration. These relations are universal, i.e. valid for dielectrics as well, provided that the charge density includes *all* charges – including those bound into the elementary dipoles. However, for most applications, it is convenient to recast them into a form where these bound charges participate not explicitly, but only via the macroscopic polarization effects they create.

If a field is created only by the stand-alone charges under consideration and is proportional to $\rho(\mathbf{r})$ (requiring that we deal with linear dielectrics), we can repeat all the argumentation of the beginning of Sec. 1.3, and again arrive at Eq. (1.60), provided that ϕ is now the macroscopic field's potential. Now we can recast this result in the terms of fields – essentially as this was done in Eqs. (1.62)-(1.64), but now making a clear difference between the macroscopic electric field $\mathbf{E} = -\nabla\phi$ and the electric displacement field **D**, which obeys the macroscopic Maxwell equation (32). Plugging $\rho(\mathbf{r})$ expressed from that equation, into Eq. (1.60), we get

$$U = \frac{1}{2}\int (\nabla \cdot \mathbf{D})\,\phi\, d^3 r . \tag{3.69}$$

Using the fact[26] that for differentiable functions ϕ and **D**,

$$(\nabla \cdot \mathbf{D})\phi = \nabla \cdot (\phi\, \mathbf{D}) - (\nabla\phi) \cdot \mathbf{D} , \tag{3.70}$$

we may rewrite Eq. (69) as

$$U = \frac{1}{2}\int \nabla \cdot (\phi\, \mathbf{D})\, d^3 r - \frac{1}{2}\int (\nabla\phi) \cdot \mathbf{D}\, d^3 r . \tag{3.71}$$

[26] See, e.g., MA Eq. (11.4a).

The divergence theorem, applied to the first term on the right-hand side, reduces it to a surface integral of ϕD_n. (As a reminder, in Eq. (1.63) the integral was of $\phi(\nabla\phi)_n \propto \phi E_n$.) If the surface of the volume we are considering is sufficiently far, this surface integral vanishes. On the other hand, the gradient in the second term of Eq. (71) is just (minus) field **E**, so it gives

$$U = \frac{1}{2}\int \mathbf{E}\cdot\mathbf{D}\, d^3r = \frac{1}{2}\int E(\mathbf{r})\,\varepsilon(\mathbf{r})E(\mathbf{r})\, d^3r \equiv \frac{\varepsilon_0}{2}\int \kappa(\mathbf{r})E^2(\mathbf{r})\, d^3r\,. \tag{3.72}$$

This expression is a natural generalization of Eq. (1.65), and shows that we can, as we did in free space, represent the electrostatic energy in a local form:[27]

$$U = \int u(\mathbf{r})d^3r, \qquad \text{with } u = \frac{1}{2}\mathbf{E}\cdot\mathbf{D} = \frac{\varepsilon}{2}E^2 = \frac{D^2}{2\varepsilon}\,. \tag{3.73}$$

Field energy in a linear dielectric

As a sanity check, in the trivial case $\varepsilon = \varepsilon_0$ (i.e. $\kappa = 1$), this result is reduced to Eq. (1.65).

Of course, Eq. (73) is valid only for linear dielectrics, because our starting point, Eq. (1.60), is only valid if ϕ is proportional to ρ. To make our calculation more general, we should intercept the calculations of Sec. 1.3 at an earlier stage, at which this proportionality had not yet been used. For example, the first of Eqs. (1.56) may be rewritten, in the continuous form, as

$$\delta U = \int \phi(\mathbf{r})\delta\rho(\mathbf{r})d^3r\,, \tag{3.74}$$

where the symbol δ means a small variation of the function – e.g., its change in time, sufficiently slow to ignore the relativistic and magnetic-field effects. Applying such variation to Eq. (32), and plugging the resulting relation $\delta\rho = \nabla\cdot\delta\mathbf{D}$ into Eq. (74), we get

$$\delta U = \int (\nabla\cdot\delta\mathbf{D})\phi\, d^3r\,. \tag{3.75}$$

(Note that in contrast to Eq. (69), this expression does not have the front factor ½.) Now repeating the same calculations as in the linear case, for the energy density's *variation* we get a remarkably simple (and general!) formula,

$$\delta u = \mathbf{E}\cdot\delta\mathbf{D} \equiv \sum_{j=1}^{3} E_j \delta D_j\,, \tag{3.76}$$

Energy density's variation

where the last expression uses the Cartesian components of the vectors **E** and **D**. This is as far as we can go for the general dependence **D**(**E**). If the dependence is linear and isotropic, as in Eq. (46), then $\delta\mathbf{D} = \varepsilon\delta\mathbf{E}$ and

$$\delta u = \varepsilon\mathbf{E}\cdot\delta\mathbf{E} \equiv \varepsilon\delta\left(\frac{E^2}{2}\right). \tag{3.77}$$

The integration of this expression over the whole variation, from the field equal to zero to a certain final distribution **E**(**r**), brings us back to Eq. (73).

An important role of Eq. (76), in its last form, is to indicate that from the point of view of analytical mechanics, the Cartesian coordinates of **E** may be interpreted as generalized forces, and those

[27] In the Gaussian units, each of the last three expressions should be divided by 4π.

of **D** as generalized coordinates of the field's effect on a unit volume of the dielectric. This allows one, in particular, to form the proper *Gibbs potential energy*[28] of a system with an electric field **E**(**r**) fixed, at every point, by some external source:

Gibbs potential energy

$$U_G = \int_V u_G(\mathbf{r}) d^3 r, \quad u_G(\mathbf{r}) = u(\mathbf{r}) - \mathbf{E}(\mathbf{r}) \cdot \mathbf{D}(\mathbf{r}). \tag{3.78}$$

The essence of this notion is that if the generalized external force (in our case, **E**) is fixed, the stable equilibrium of the system corresponds to the minimum of U_G, rather than of the potential energy U as such – in our case, that of the field in our system.

As the simplest illustration of this important concept, let us consider a very long cylinder (with an arbitrary cross-section shape), made of a uniform linear dielectric, placed into a uniform external electric field parallel to the cylinder's axis – see Fig. 13.

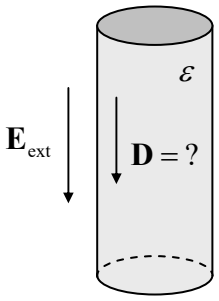


Fig. 3.13. A cylindrical dielectric sample in a longitudinal external electric field.

For this simple problem, the equilibrium value of **D** inside the cylinder may be, of course, readily found without any appeal to energies. Indeed, the solution of the Laplace equation inside the cylinder, with the boundary condition (37) is evident: $\mathbf{E}(\mathbf{r}) = \mathbf{E}_{ext}$, and so Eq. (46) immediately yields $\mathbf{D}(\mathbf{r}) = \varepsilon\mathbf{E}_{ext}$. One may wonder why the minimum of the potential energy U, given by Eq. (73) in its last form,

$$\frac{U}{V} = \frac{D^2}{2\varepsilon}, \tag{3.79}$$

corresponds to a different (zero) value of **D**, but let us recall that Eq. (73) was derived for the case when the electric field is created by the stand-alone charges in the system under consideration. If it is created by external sources, we have to use the Gibbs potential energy (78) instead. For our current uniform case, this energy per unit volume of the cylinder is

$$\frac{U_G}{V} = \frac{U}{V} - \mathbf{E} \cdot \mathbf{D} = \frac{D^2}{2\varepsilon} - \mathbf{E} \cdot \mathbf{D} \equiv \sum_{j=1}^{3} \left(\frac{D_j^2}{2\varepsilon} - E_j D_j \right), \tag{3.80}$$

and its minimum as a function of every Cartesian component of **D** corresponds to the correct value of the displacement: $D_j = \varepsilon E_j$, i.e. to $\mathbf{D} = \varepsilon\mathbf{E} = \varepsilon\mathbf{E}_{ext}$. So, the systems' equilibrium indeed corresponds to the minimum of the Gibbs potential energy (78) rather than of the energy (73).

[28] See, e.g., CM Sec. 1.4, in particular Eq. (1.41). Note that as Eq. (78) clearly illustrates, once again, that the difference between the potential energies U_G and U, usually discussed in courses of thermodynamics and statistical physics as the difference between the Gibbs and Helmholtz free energies (see, e.g., SM 1.4), is much more general than the effects of random thermal motion addressed by these disciplines.

Now note that Eq. (80), at this equilibrium point (only!), may be rewritten as

$$\frac{U_G}{V} = \frac{U}{V} - \mathbf{E}\cdot\mathbf{D} = \frac{D^2}{2\varepsilon} - \frac{\mathbf{D}}{\varepsilon}\cdot\mathbf{D} \equiv -\frac{D^2}{2\varepsilon}, \tag{3.81}$$

i.e. formally coincides with Eq. (79), besides the (perhaps, somewhat counter-intuitive) *opposite sign*. A similar but more general relation (not limited to linear dielectrics and uniform fields) may be obtained by taking the variation of the u_G expressed by Eq. (78), and then using Eq. (76):

$$\delta u_G = \delta u - \delta(\mathbf{E}\cdot\mathbf{D}) = \mathbf{E}\cdot\delta\mathbf{D} - (\delta\mathbf{E}\cdot\mathbf{D} + \mathbf{E}\cdot\delta\mathbf{D}) \equiv -\mathbf{D}\cdot\delta\mathbf{E}. \tag{3.82}$$

In order to see how this expression works, let us plug **D** from Eq. (33):

$$\delta u_G = -(\varepsilon_0\mathbf{E} + \mathbf{P})\cdot\delta\mathbf{E} \equiv -\delta\left(\frac{\varepsilon_0 E^2}{2}\right) - \mathbf{P}\cdot\delta\mathbf{E}. \tag{3.83}$$

So far, this relation is general. In the particular case when the polarization **P** is field-independent, we may integrate Eq. (83) over the full electric field's variation, say from 0 to some finite value **E**, getting

$$u_G = -\frac{\varepsilon_0 E^2}{2} - \mathbf{P}\cdot\mathbf{E}. \tag{3.84}$$

Again, the Gibbs energy is relevant only if **E** is dominated by an external field $\mathbf{E}_{ext}$ independent of the orientation of **P**. If, in addition, $\mathbf{P}(\mathbf{r}) \neq 0$ only in some finite volume V, we may integrate Eq. (84) over that volume, getting

$$U_G = -\mathbf{p}\cdot\mathbf{E}_{ext} + \text{const}, \qquad \text{with } \mathbf{p} \equiv \int_V \mathbf{P}(\mathbf{r})d^3r, \tag{3.85}$$

where the "const" means the terms independent of **p**. In this expression, we may readily recognize Eq. (15a) for an electric dipole **p** of a fixed magnitude, which was obtained in Sec. 1 in a different way. This comparison illustrates again that U_G is nothing mysterious; it is just the relevant part of the potential energy of the system in a fixed external field, including the energy of its interaction with the field.

Finally, in the other important case of a linear dielectric, when according to Eqs. (45) and (47), $\mathbf{P} = (\varepsilon - \varepsilon_0)\mathbf{E}$, the similar integration of the general Eq. (83) over the field yields the additional factor ½:

$$U_G = -\frac{1}{2}\int_V \mathbf{P}\cdot\mathbf{E}_{ext}d^3r + \text{const}. \tag{3.86}$$

This expression may be very convenient for analyses of the forces exerted by electric fields on linear dielectric media – see, for, example, a few exercises on this topic, offered at the end of this chapter.

3.6. Exercise problems

3.1. Prove Eqs. (3)-(4), starting from Eqs. (1.38) and (3.2).

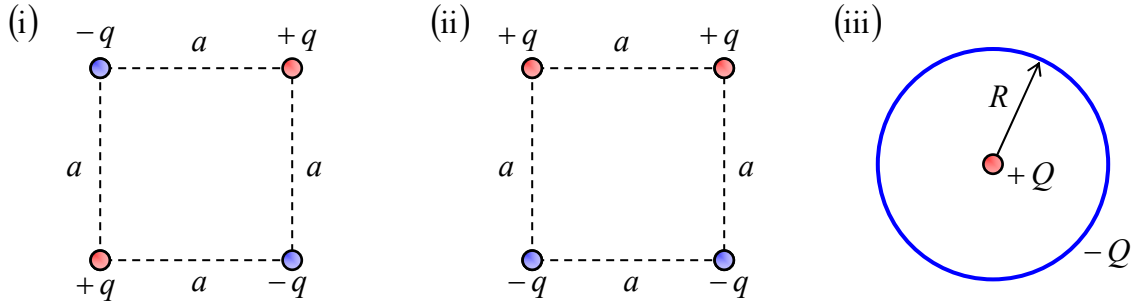
3.2. A thin ring of radius R is charged with a constant linear density λ. Calculate the exact electrostatic potential distribution along the symmetry axis of the ring, and prove that at large distances, $r >> R$, the three leading terms of its multipole expansion are indeed correctly described by Eqs. (3)-(4).

3.3. In suitable reference frames, calculate the dipole and quadrupole moments of the following systems (see the figures below):

(i) four point charges of the same magnitude but alternating signs, placed in the corners of a square;

(ii) a similar system but with a pair charge sign alternation; and

(iii) a point charge in the center of a thin ring carrying a similar but opposite charge uniformly distributed along its circumference.



3.4. Calculate the dipole and quadrupole moments of a thin spherical shell of radius R, carrying an electric charge with the areal density $\sigma = \sigma_0 \cos\theta$. Discuss the relation between the results and the solution of Problem 2.28.

3.5. For a regular cubic lattice of similarly oriented identical dipoles, calculate the electric field it creates at the location of each dipole.

3.6. Without carrying out an exact calculation, can you predict the spatial dependence of the interaction between various electric multipoles, including point charges (in this context, frequently called electric *monopoles*), dipoles, and quadrupoles? Based on these predictions, what is the functional dependence of the interaction between *homonuclear* diatomic molecules such as H_2, N_2, O_2, etc., on the distance between them when the distance is much larger than the molecular size?

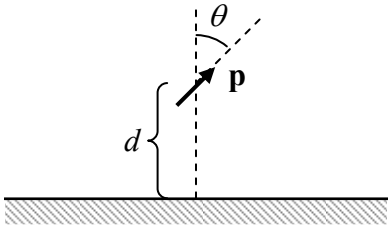
3.7. Two similar electric dipoles, of a fixed magnitude p, located at a fixed distance r from each other, are free to change their directions. What stable equilibrium position(s) they may take as a result of their electrostatic interaction?

3.8. An electric dipole is located above a grounded infinite conducting plane (see the figure on the right). Calculate:

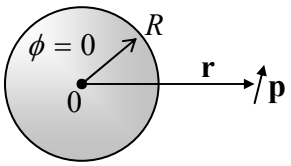
(i) the distribution of the induced charge in the conductor,

(ii) the dipole-to-plane interaction energy, and

(ii) the force and the torque exerted on the dipole.



3.9. Calculate the net charge Q induced in a grounded conducting sphere of radius R by a dipole $\mathbf{p}$ located at point $\mathbf{r}$ outside the sphere – see the figure on the right.



3.10. Use two different approaches to calculate the energy of interaction between a grounded conductor and an electric dipole **p** placed in the center of a spherical cavity of radius R, carved in the conductor.

3.11. A plane separating two halves of otherwise free space is densely and uniformly (with a constant areal density n) filled with electric dipoles, with similar moments **p** oriented normally to the plane.

(i) Use two different approaches to calculate the electrostatic potential at distances $d >> 1/n^{1/2}$ on both sides of the plane.

(ii) Give a physical interpretation of your result.

(iii) Use the result to calculate the potential distribution created in space by a spherical surface of radius R, densely and uniformly filled with radially oriented dipoles.

3.12. Prove Eq. (24).

Hint: You may like to use the basic Eq. (1.9) to spell out the left-hand side of Eq. (1.24), change the order of integration over **r** and **r**', and then contemplate the physical sense of the inner integral.

3.13. A sphere of radius R is made of a material with a uniform spontaneous polarization $\mathbf{P}_0$. Calculate the electric field everywhere in space – both inside and outside the sphere, and compare the result for the internal field with Eq. (24).

3.14. Calculate the electric field at the center of a cube made of a material with the uniform spontaneous polarization $\mathbf{P}_0$ of arbitrary orientation.

3.15. Derive the Clausius-Mossotti formula (52) by combining Eq. (24) with the result of the solution of Problem 5.

3.16. Stand-alone charge Q is distributed, in some way, within the volume of a body made of a uniform linear dielectric with a dielectric constant κ. Calculate the total polarization charge Q_{ef} residing on the surface of the body, provided that it is surrounded by free space.

3.17. In two separate experiments, a thin plane sheet of a linear dielectric with $\kappa = \text{const}$ is placed into a uniform external electric field $\mathbf{E}_0$, in two different ways:

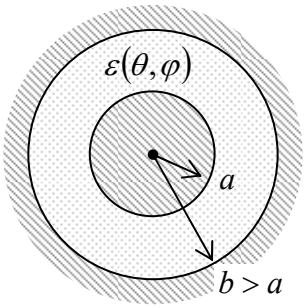
(i) with the sheet's surfaces parallel to the electric field, and

(ii) with its surfaces normal to the field.

For each case, find the electric field **E**, the electric displacement **D**, and the polarization **P** inside the dielectric, sufficiently far from the sheet's edges.

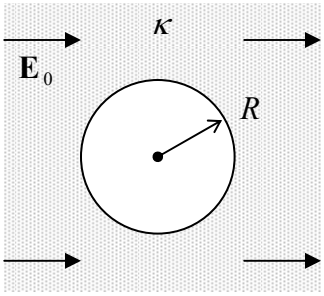
3.18. A fixed dipole **p** is placed in the center of a spherical cavity of radius R, carved inside a uniform linear dielectric. Calculate the electric field distribution everywhere in the system.

Hint: You may start with the assumption that the field at $r > R$ has a distribution typical for a dipole. However, be ready for surprises.



3.19. A spherical capacitor (see the figure on the right) is filled with a linear dielectric whose permittivity ε depends on the spherical angles θ and φ, but not on the distance r from the system's center. Derive an explicit expression for its capacitance C.

3.20. A spherical capacitor similar to that considered in the previous problem is now filled with a linear dielectric whose permittivity depends only on the distance from the center. Obtain an explicit expression for its capacitance, and spell it out for the particular case $\varepsilon(r) = \varepsilon(a)(r/a)^n$.



3.21. A uniform electric field $\mathbf{E}_0$ has been created (by distant external sources) inside a uniform linear dielectric. Find the electric field's change created by carving out a cavity in the shape of a round cylinder of radius R, with its axis normal to the external field – see the figure on the right.

3.22. Similar small spherical particles, made of a linear dielectric, are dispersed in free space with a low concentration $n << 1/R^3$, where R is the particle's radius. Calculate the average dielectric constant of such a medium. Compare the result with the apparent but wrong answer

$$\bar{\kappa} - 1 = (\kappa - 1) n V, \qquad \text{(WRONG!)}$$

(where κ is the dielectric constant of the particle's material and $V = (4\pi/3)R^3$ is its volume), and explain the origin of the difference.

3.23. A straight thin filament, uniformly charged with linear density λ, is positioned parallel to the plane separating two uniform linear dielectrics, at a distance d from it. Calculate the electric potential's distribution everywhere in the system.

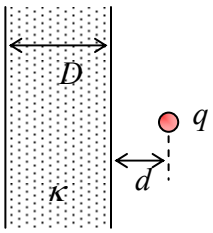
3.24. A point charge q is located at a distance $d > R$ from the center of a sphere of radius R, made of a uniform linear dielectric with permittivity ε.

(i) Calculate the electrostatic potential's distribution in all the space, for an arbitrary ratio d/R.

(ii) For large d/R, use two different approaches to calculate the interaction force and the energy of interaction between the sphere and the charge, in the first nonzero approximation in $R/d << 1$.

Hint: Task (i) cannot be carried out using the method of charge images, so you may like to use the expansion of the function $1/|\mathbf{r} - \mathbf{r}'|$ in the series over the Legendre polynomials, whose proof was the subject of Problem 2.40.

3.25. Calculate the spatial distribution of the electrostatic potential induced by a point charge q located at distance d from a very wide parallel plate, of thickness D, made of a uniform linear dielectric – see the figure on the right.

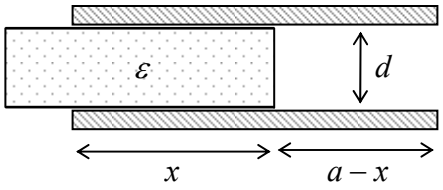


3.26. Discuss the physical nature of Eq. (76). Apply your conclusions to a material with a fixed (field-independent) polarization $\mathbf{P}_0(\mathbf{r})$, and calculate the electric field's energy of a uniformly polarized sphere (see Problem 13 above).

3.27. Use Eqs. (73) and (82) to calculate the force of attraction of a plane capacitor's plates (per unit area), for two cases:

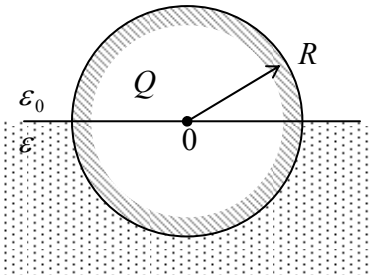
(i) the capacitor is charged to voltage V, and then disconnected from the battery,[29] and
(ii) the capacitor remains connected to the battery.

3.28. A slab made of a linear dielectric is partly inserted into a plane capacitor – see the figure on the right. Assuming the simplest (cylindrical) geometry of the system, calculate the force exerted by the field on the slab, for the same two cases as in the previous problem



3.29. For each of the two capacitors shown in Fig. 10, calculate the electric force exerted on the interface between two different dielectrics, in terms of the fields in the system.

3.30. One half of a conducting sphere of radius R, carrying electric charge Q, is submerged into a half-space filled with a linear dielectric with permittivity ε – see the figure on the right. Calculate the electric force exerted on the sphere by the dielectric.



[29] "Battery" is a common if misleading term for what is usually a single *galvanic element*. (The last term stems from the name of Luigi Galvani, a pioneer of electric current studies. Another term derived from his name is the *galvanic connection*, meaning a direct connection of two conductors, enabling a dc current flow – see the next chapter.) The term "battery" had to be, in all fairness, reserved for the connection of several galvanic elements in series – as was pioneered in 1800 by L. Galvani's friend Alexander Volta.

Chapter 4. DC Currents

The goal of this chapter is to discuss the distribution of stationary ("dc") currents in conducting samples and their "global" characteristics such as resistance. In the most important case of linear ("Ohmic") conductivity, the current distribution is governed by the same Laplace and Poisson equations whose solution methods were discussed in detail in the previous chapters. Because of that, we can piggyback on most approaches discussed earlier, enabling me to keep this chapter rather brief.

4.1. Continuity equation and the Kirchhoff laws

Until this point, our discussion of conductors has been limited to the cases when they are separated with *insulators* (meaning either the free space or some dielectric media), preventing any continuous motion of charges from one conductor to another, even if there is a non-zero voltage (and hence electric field) between them – see Fig. 1a.

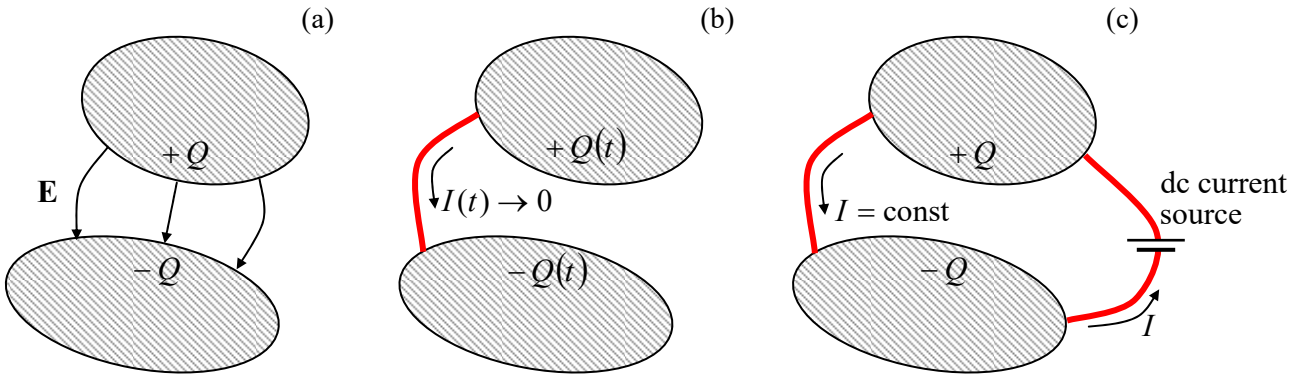


Fig. 4.1. Two oppositely charged conductors: (a) in the electrostatic situation, (b) at the charge relaxation through an additional narrow conductor ("wire"), and (c) in a system sustaining a dc current I.

Now let us connect the two conductors with a *wire* – a thin, elongated conductor (Fig. 1b). Then the electric field causes the motion of charge carriers in the wire, from the conductor with a higher electrostatic potential toward that with lower potential, until the potentials equilibrate. Such a process is called *charge relaxation*. The main equation governing this process may be obtained from the fundamental experimental fact (already mentioned in Sec. 1.1) that electric charges cannot appear or disappear – though opposite charges may recombine with the conservation of the net charge. As a result, the charge Q in a conductor may change only due to the *electric current* I through the wire:

$$\frac{dQ}{dt} = -I(t); \tag{4.1}$$

this relation may be understood as the definition of the current.[1]

[1] Just as a (hopefully, unnecessary :-) reminder, in the SI units the current is measured in amperes (A). In legal metrology, the ampere (rather than the coulomb, which is defined as 1C = 1A × 1s) is a primary unit. (Its formal definition will be discussed in the next chapter.) In the Gaussian units, Eq. (1) remains the same, so the current's unit is the statcoulomb per second – the so-called *statampere*.

Let us express Eq. (1) in a differential form, introducing the notion of the *current density* **j**(**r**). This vector may be defined via the following relation for the elementary current dI crossing an elementary area dA (Fig. 2):

$$dI = jdA\cos\theta \equiv (j\cos\theta)dA \equiv j_n dA\,, \tag{4.2}$$

where θ is the angle between the direction normal to the surface and the charge carrier motion direction, which is taken for the direction of the vector **j**.

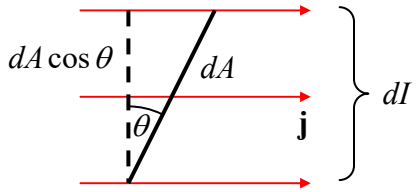


Fig. 4.2. The current density vector **j**.

With that definition, Eq. (1) may be rewritten as

$$\frac{d}{dt}\int_V \rho\, d^3r = -\oint_S j_n d^2r\,, \tag{4.3}$$

where V is an arbitrary but stationary volume limited by the closed surface S. Applying to this volume the same divergence theorem as was repeatedly used in previous chapters, we get

$$\int_V \left[\frac{\partial\rho}{\partial t} + \nabla\cdot\mathbf{j}\right] d^3r = 0\,. \tag{4.4}$$

Since the volume V is arbitrary, this equation may be true only if

$$\boxed{\frac{\partial\rho}{\partial t} + \nabla\cdot\mathbf{j} = 0\,.} \tag{4.5}$$

Continuity equation

This is the fundamental *continuity equation* – which is true even for time-dependent phenomena.[2]

The charge relaxation, illustrated by Fig. 1b, is of course a dynamic, time-dependent process. However, electric currents may also exist in stationary situations, when a certain *current source*, for example a battery, drives the current against the electric field, and thus replenishes the conductor charges and sustains currents at a certain time-independent level – see Fig. 1c. (This process requires a persistent replenishment of the electrostatic energy of the system from either a source or a large storage of energy of a different kind – say, the chemical energy of the battery.) Let us discuss the laws governing the distribution of such *dc currents*. In this case ($\partial/\partial t = 0$), Eq. (5) reduces to a very simple equation

$$\nabla\cdot\mathbf{j} = 0\,. \tag{4.6}$$

This relation acquires an even simpler form in the particular but important case of *dc electric circuits* (Fig. 3) – the systems that may be fairly represented as direct ("galvanic") connections of components of two types:

[2] Similar differential relations are valid for the density of any conserved quantity, for example for mass in classical dynamics (see, e.g., CM Sec. 8.3), and for the probability, as it is defined in statistical physics (SM Sec. 5.6) and in quantum mechanics (QM Sec. 1.4).

(i) relatively-small-size (*lumped*) *circuit elements*, meaning either a passive resistor, or a current source, etc. – generally, any "black box" with two or more *terminals*, and

(ii) *perfectly conducting wires*, with a negligible drop of the electrostatic potential along them, that are galvanically connected at certain points called *nodes* (or "junctions").

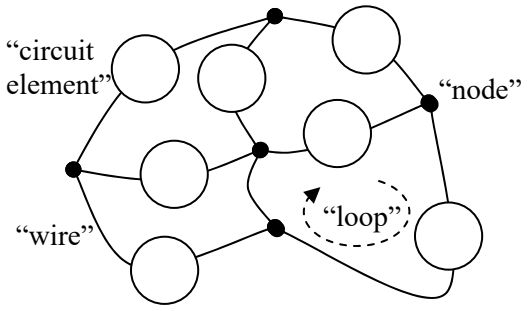


Fig. 4.3. A typical system obeying Kirchhoff laws.

In the standard circuit theory, the electric charges of the nodes are considered negligible,[3] and we may integrate Eq. (6) over the closed surface drawn around any node to get a simple equality

$$\sum_j I_j = 0\,, \tag{4.7a}$$

where the summation is over all the wires (numbered with index j) connected in the node. On the other hand, according to its definition (2.25), the voltage V_k across each circuit element may be represented as the difference of the electrostatic potentials of the adjacent nodes, $V_k = \phi_k - \phi_{k-1}$. Summing such differences around any closed loop of the circuit (Fig. 3), we get all terms canceled, so

$$\sum_k V_k = 0\,. \tag{4.7b}$$

These relations are called, respectively, the *1st* and *2nd Kirchhoff laws*[4] – or sometimes the *node rule* (7a) and the *loop rule* (7b). They may seem elementary, and their genuine power is in the mathematical fact that any set of Eqs. (7) covering every node and every circuit element of the system at least once, gives a system of equations sufficient for the calculation of all currents and voltages in it – provided that the relation between the current and voltage is known for each circuit element.

It is almost evident that in the absence of current sources, the system of equations (7) has only the trivial solution: $I_j = 0$, $V_k = 0$ – with the exotic exception of superconductivity, to be discussed in Sec. 6.3. The current sources that allow non-zero current flows may be described by their *electromotive forces* (*e.m.f.*) $\mathcal{V}_k$, having the dimensionality of voltage, which have to be taken into account in the corresponding terms V_k of the sum (7b). Let me hope that the reader has some experience of using Eqs. (7) for analyses of simple circuits – say, consisting of several resistors and batteries, so I can save our time by skipping their discussion. Still, due to their practical importance, I would recommend the reader to carry out a self-test by solving a couple of problems offered at the beginning of Sec. 6.

[3] In many cases, the charge accumulation/relaxation may be described without an explicit violation of Eq. (7a), just by adding other circuit elements, *lumped capacitors* (see Fig. 2.5 and its discussion), to the circuit under analysis. The resulting circuit may be used to describe not only the transient processes but also periodic ac currents. However, it is convenient for me to postpone the discussion of such *ac circuits* until Chapter 6, where one more circuit element type, *lumped inductances*, will be introduced.

[4] Named after Gustav Kirchhoff (1824-1887) – who also suggested the differential form (8) of the Ohm law.

4.2. The Ohm law

As was mentioned above, the relations spelled out in Sec. 1 are sufficient for forming a closed system of equations for finding electric current and field in a system only if they are complemented with some constitutive relations between the scalars I and V in each lumped circuit element, or alternatively between the macroscopic (atomic–scale–averaged) vectors $\mathbf{j}$ and $\mathbf{E}$ at each point of the material of such an element. The simplest of such relations is the famous *Ohm law* whose differential (or "local") form is

$$\mathbf{j} = \sigma \mathbf{E}, \tag{4.8}$$

Ohm law

where σ is a constant called the *Ohmic conductivity* (or just the "conductivity" for short).[5] Though the Ohm law (discovered, in its simpler form, by Georg Simon Ohm in 1827) is one of constitutive rather than fundamental relations, and is approximate for *any* conducting medium, we can argue that if:

(i) the medium carries no current at $\mathbf{E} = 0$ (mind superconductors!),

(ii) the medium is isotropic or virtually isotropic (a notable exception: some organic conductors),

(iii) the *mean free path* l of the current carriers (the notion to be discussed in detail in SM Ch. 6) in this medium is much smaller than the characteristic scale a of the spatial variations of $\mathbf{j}$ and $\mathbf{E}$,

then the law may be viewed as the leading, linear term of the Taylor expansion of the local relation $\mathbf{j}(\mathbf{E})$, and thus is general for relatively low fields.

Table 1 gives approximate experimental values of σ for some representative (and/or practically important) materials. Note that the range of these values is very broad, even without going to such extremes as very pure metallic crystals at very low temperatures, where σ may reach $\sim 10^{12}$ S/m.

Table 4.1. Ohmic dc conductivities for some materials at 20°C.

Material	σ (S/m)
Teflon (PTFE, $[C_2F_4]_n$)	10^{-22}-10^{-24}
Silicon dioxide	10^{-16}-10^{-19}
Various glasses	10^{-10}-10^{-14}
Deionized water	$\sim 10^{-6}$
Seawater	5
Silicon *n*-doped to 10^{16}cm^{-3}	2.5×10^{2}
Silicon *n*-doped to 10^{19}cm^{-3}	1.6×10^{4}
Silicon *p*-doped to 10^{19}cm^{-3}	1.1×10^{4}
Nichrome (alloy 80% Ni + 20% Cr)	0.9×10^{6}
Aluminum	3.8×10^{7}
Copper	6.0×10^{7}
Zinc crystal along *a*-axis	1.65×10^{7}
Zinc crystal along *c*-axis	1.72×10^{7}

[5] In SI units, the conductivity is measured in S/m, where one siemens (S) is the reciprocal of the ohm: $1\text{S} \equiv (1\Omega)^{-1} \equiv 1\text{A}/1\text{V}$. The constant reciprocal to conductivity, $1/\sigma$, is called *resistivity* and is commonly denoted by the letter ρ. I will, however, try to avoid using this notion, because in these notes this letter is already overused.

In order to get a better feeling of what these values mean, let us consider a very simple system (Fig. 4): a plane capacitor of area $A >> d^2$, filled with a material that has not only a dielectric constant κ, but also some Ohmic conductivity σ, with much more conductive electrodes.

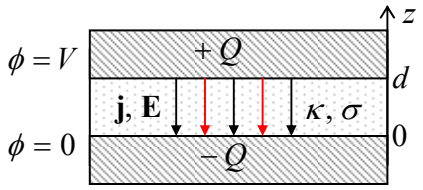


Fig. 4.4. A "leaky" plane capacitor.

Assuming that these properties are compatible with each other,[6] we may assume that the distribution of the electric potential (not too close to the capacitor's edges) still obeys Eq. (2.39), so the electric field is normal to the electrode surfaces and uniform, with $E = V/d$. Then, according to Eq. (6), the current density is also uniform, $j = \sigma E = \sigma V/d$. From here, the total current between the plates is

$$I = jA = \sigma EA = \sigma \frac{V}{d} A. \tag{4.9}$$

On the other hand, from Eqs. (2.26) and (3.45), the instantaneous value of the total charge of the top electrode is $Q = CV = (\kappa\varepsilon_0 A/d)V$. Plugging these relations into Eq. (1), we see that the speed of charge (and voltage) relaxation is independent of the geometric parameters A and d of the capacitor:

$$\frac{dV}{dt} = -\frac{V}{\tau_r}, \qquad \text{with } \tau_r \equiv \frac{\varepsilon_0 \kappa}{\sigma} \equiv \frac{\varepsilon}{\sigma}, \tag{4.10}$$

so the *relaxation time constant* τ_r may be used to characterize the gap-filling material as such.

As we already know (see Table 3.1), for most practical materials the dielectric constant κ is within one order of magnitude from 10, so the numerator in the second of Eqs. (10) is of the order of 10^{-10} (SI units). As a result, according to Table 1, the charge relaxation time ranges from $\sim 10^{14}$s (more than a million years!) for the best insulators like Teflon (polytetrafluoroethylene, PTFE),[7] to $\sim 10^{-18}$s for the least resistive metals. What is the physics behind such a huge range of σ, and why, for some materials, Table 1 gives them with such a large uncertainty? As in Chapters 2 and 3, in this course, I have time only for a brief, admittedly superficial discussion of these issues.[8]

If the charge carriers move almost as classical particles (e.g., in plasmas or non-degenerate semiconductors), a very reasonable description of the conductivity is given by the famous *Drude formula*.[9] In his picture, due to a weak electric field, the charge carriers are accelerated in its direction (on top of their random motion in all directions, with the average velocity vector equal to zero):

$$\frac{d\mathbf{v}}{dt} = \frac{q}{m}\mathbf{E}, \tag{4.11}$$

and as a result, their velocity acquires the average value

[6] As will be discussed in Chapter 6, this is true only if σ is not too high.
[7] This polymer is broadly used in engineering and physical experiment, due to its many remarkable properties.
[8] A more detailed discussion of this issue may be found in SM Chapter 6.
[9] It was suggested by Paul Drude in 1900.

$$\mathbf{v} = \frac{d\mathbf{v}}{dt}\tau = \frac{q}{m}\mathbf{E}\tau, \tag{4.12}$$

where the phenomenological parameter $\tau = l/v$ (not to be confused with τ_r!) may be understood as the average time since the last scattering event. From here, the current density:[10]

$$\mathbf{j} = qn\mathbf{v} = \frac{q^2 n\tau}{m}\mathbf{E}, \qquad \text{i.e. } \sigma = \frac{q^2 n\tau}{m}. \tag{4.13a}$$

Drude formula: two versions

(Notice the independence of σ of the charge sign.) Another form of the same result, more popular in the physics of semiconductors, is

$$\sigma = q^2 n\mu, \qquad \text{with } \mu = \frac{\tau}{m}, \tag{4.13b}$$

where the parameter μ, defined by the relation $\mathbf{v} \equiv \mu\mathbf{E}$, is called the *charge carrier mobility.*

Most good conductors (e.g., metals) are essentially degenerate Fermi gases (or liquids), in which the average thermal energy of a particle, $k_B T$ is much lower than the Fermi energy ε_F. In this case, a quantum theory is needed for the calculation of σ. Such a theory was developed by A. Sommerfeld in 1927 (and is sometimes called the *Drude-Sommerfeld model*). I have no time to discuss it in this course,[11] and here will only notice that for a nearly ideal, isotropic Fermi gas the result is reduced to Eq. (13), with a certain effective value of τ, so it may be used for estimates of σ, with due respect to the quantum theory of scattering. In a typical metal, n is very high (~10^{23} cm^{-3}) and is fixed by the atomic structure, so the sample quality may only affect σ via the scattering time τ.

At room temperature, the scattering of electrons by thermally-excited lattice vibrations (*phonons*) dominates, so τ and σ are high but finite, and do not change much from one sample to another. (Hence the relatively accurate values given for metals in Table 1.) On the other hand, at $T \to 0$, quantum mechanics says a perfect crystal should not exhibit scattering at all, and its conductivity should be infinite. In practice, this is never true (for one, due to electron scattering from imperfect boundaries of finite-size samples), and the effective conductivity σ is infinite (or practically infinite, at least above the largest measurable values ~10^{20} S/m) only in superconductors.[12]

On the other hand, the conductivity of quasi-insulators (including deionized water) and semiconductors depends mostly on the carrier density n, which is much lower than in metals. From the point of view of quantum mechanics, this happens because the ground-state wavefunctions of charge carriers are localized within an atom (or molecule), and their energies are separated from those of excited states, with space-extended wavefunctions, by a large energy gap – often called the *bandgap*. For example, in SiO_2 the bandgap approaches 9 eV, equivalent to ~4,000 K. This is why even at room temperatures the density of thermally-excited free charge carriers in good insulators is negligible. In these materials, n is determined by impurities and vacancies, and may depend on a particular chemical synthesis or other fabrication technology, rather than on the fundamental properties of the material. (On the contrary, the carrier mobility μ in these materials is almost technology-independent.)

[10] Note that **j** in Eq. (8) is defined as an already macroscopic variable, averaged over inter-particle distances, so no additional average sign is necessary in the first of Eqs. (13a).

[11] For such a discussion see, e.g., SM Sec. 6.3.

[12] The electrodynamic properties of superconductors are so interesting (and fundamentally important) that I will discuss them in more detail in Chapter 6.

The practical importance of the fabrication technology may be illustrated by the following example. In the cells of the so-called *floating-gate memories*, in particular, the *flash memories*, which currently dominate the nonvolatile digital memory technology, data bits are stored as small electric charges ($Q \sim 10^{-16}$ C $\sim 10^3$ e) of highly doped silicon islands (so-called *floating gates*) separated from the rest of the integrated circuit with ~10-nm-thick layers of silicon dioxide, SiO_2. Such layers are fabricated by high-temperature oxidation of virtually perfect silicon crystals. The conductivity of the resulting high-quality (though amorphous) material is so low, $\sigma \sim 10^{-19}$ S/m, that the relaxation time τ_r, defined by Eq. (10), is well above 10 years – the industrial standard for data retention in nonvolatile memories. To appreciate how good this technology is, the cited value should be compared with the typical conductivity $\sigma \sim 10^{-16}$ S/m of the usual, bulk SiO_2 ceramics.[13]

To conclude this chapter, let me note that the Ohm law, for all its importance, is not a universal law of nature. As a reminder of this fact, in Sec. 5 below I describe two very simple systems (leaving their analysis for the reader's exercise) whose *I-V* relation is nonlinear even for very small currents.

4.3. Boundary problems

For an Ohmic conducting medium, we may combine Eqs. (6) and (8) to get the following differential equation

$$\nabla \cdot (\sigma \nabla \phi) = 0 . \tag{4.14}$$

For a uniform conductor (σ = const), Eq. (14) is reduced to the Laplace equation for the (macroscopic) electrostatic potential ϕ. As we already know from Chapters 2 and 3, its solution depends on the boundary conditions. These conditions, in turn, depend on the interface type.

(i) Conductor-conductor interface. Applying the continuity equation (6) to a Gauss-type pillbox at the interface of two different conductors (Fig. 5), we get

$$(j_n)_1 = (j_n)_2, \tag{4.15}$$

so if the Ohm law (8) is valid inside each medium, then

$$\sigma_1 \frac{\partial \phi_1}{\partial n} = \sigma_2 \frac{\partial \phi_2}{\partial n} . \tag{4.16}$$

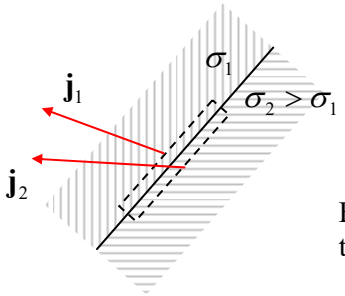


Fig. 4.5. DC current's "refraction" at the interface between two different conductors.

[13] This course is not an appropriate platform to discuss details of the floating-gate memory technology. However, I think that every educated physicist should know its basics, because such memories are presently the driver of all semiconductor integrated circuit technology development, and hence of the whole information technology progress. Perhaps the best available general book on this topic is still the relatively old review collection by J. Brewer and M. Gill (eds.), *Nonvolatile Memory Technologies with Emphasis on Flash*, IEEE Press, 2008.

Also, since the electric field should be finite, its potential ϕ has to be continuous across the interface – the condition that may also be written as

$$\frac{\partial \phi_1}{\partial \tau} = \frac{\partial \phi_2}{\partial \tau}. \tag{4.17}$$

Both these conditions (and hence the solutions of the boundary problems using them) are similar to those for the interface between two dielectrics – cf. Eqs. (3.46)-(3.47). Note that using the Ohm law, Eq. (17) may be rewritten as

$$\frac{1}{\sigma_1}(j_\tau)_1 = \frac{1}{\sigma_2}(j_\tau)_2. \tag{4.18}$$

Comparing it with Eq. (15) we see that, generally, the current density's magnitude changes at the interface: $j_1 \neq j_2$. It is also curious that if $\sigma_1 \neq \sigma_2$, the current line slope changes at the interface (Fig. 5), qualitatively similar to the refraction of light rays in optics – see Chapter 7.

(ii) Conductor-electrode interface. An *electrode* is defined as a body made of a "perfect conductor", i.e. of a medium with $\sigma \to \infty$. Then, at a fixed current density at the interface, the electric field in the electrode tends to zero, and hence it may be described by the equality

$$\phi = \phi_j = \text{const}, \tag{4.19}$$

where constants ϕ_j may be different for different electrodes (numbered with index j). Note that with such boundary conditions, the Laplace boundary problem becomes exactly the same as in electrostatics – see Eq. (2.35) – and hence we can use the methods (and some solutions :-) discussed in Chapter 2 for finding the dc current distribution.

(iii) Conductor-insulator interface. For the description of a good insulator, we can use the equality $\sigma = 0$, so Eq. (16) yields the following boundary condition,

$$\frac{\partial \phi}{\partial n} = 0, \tag{4.20}$$

for the potential derivative *inside the conductor*. From the Ohm law (8) in the form $\mathbf{j} = -\sigma \nabla \phi$, we see that this is just the very natural requirement for the dc current not to flow into an insulator. Now note that this condition makes the Laplace problem inside the conductor completely well-defined, and independent of the potential distribution in the adjacent insulator. On the contrary, due to the continuity of the electrostatic potential at the border, its distribution inside the surrounding insulator has to follow that inside the conductor.

Let us discuss this conceptual issue on the following (apparently, trivial) example: dc current in a uniform wire of length l and a cross-section of area A. The reader certainly knows the answer:

$$I = \frac{V}{R}, \qquad \text{where } R \equiv \frac{V}{I} = \frac{l}{\sigma A}, \tag{4.21}$$

Uniform wire's resistance

where the constant R is called the wire's *resistance*.[14]

[14] The first of Eqs. (21) is essentially the (historically, initial) integral form of the Ohm law, and is valid not only for a uniform wire but also for Ohmic conductors of any geometry in that I and V may be clearly defined.

However, let us derive this result formally from our theoretical framework. For the simple geometry shown in Fig. 6a, this is easy to do. Here the potential evidently has a linear 1D distribution

$$\phi = \text{const} - \frac{x}{l}V, \tag{4.22}$$

both in the conductor and the surrounding free space, with both boundary conditions (16) and (17) satisfied at the conductor-insulator interfaces, and the condition (20) satisfied at the conductor-electrode interfaces. As a result, the electric field is constant and has only one Cartesian component: $E_x = V/l$, so inside the conductor

$$j_x = \sigma E_x, \quad I = j_x A, \tag{4.23}$$

giving us the well-known Eq. (21).

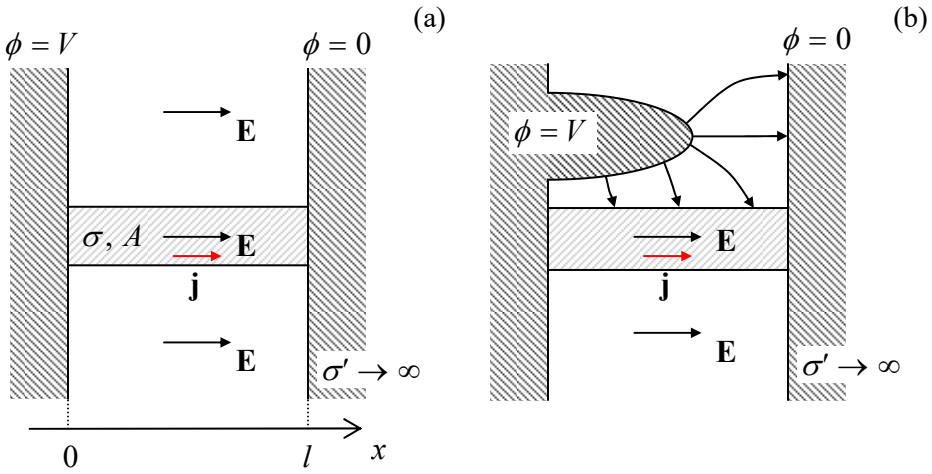


Fig. 4.6. (a) An elementary problem and (b) a (slightly) less obvious problem of the field distribution at dc current flow (schematically).

However, what about the geometry shown in Fig. 6b? In this case, the field distribution in the free space around the conductor is dramatically different, but according to the boundary problem defined by Eqs. (14) and (20), inside the conductor, the solution is exactly the same as it was in the former case. Now, the Laplace equation in the surrounding insulator has to be solved with the boundary values of the electrostatic potential, "dictated" by the distribution of the current (and hence potential) in the conductor. Note that as a result, the electric field lines are generally *not* normal to the conductor's surface, because the surface is not equipotential – see Eq. (22) again.

Let us solve a problem in that this *conduction hierarchy* may be followed analytically to the very end. Consider an empty spherical cavity carved in a conductor with an initially uniform current flow with a constant density $\mathbf{j}_0 = \mathbf{n}_z j_0$ (Fig. 7a). Following the hierarchy, we have to solve the boundary problem in the conducting part of the system, i.e. outside the sphere (at $r \geq R$), first. Since the problem is evidently axially symmetric, we already know the general solution of the Laplace equation – see Eq. (2.172). Moreover, we know that in order to match the uniform field distribution at $r \to \infty$, all coefficients a_l but one ($a_1 = -E_0 = -j_0/\sigma$) have to be zero, and that the boundary conditions at $r = R$ will give zero solutions for all coefficients b_l but one (b_1), so

$$\phi = -\frac{j_0}{\sigma} r\cos\theta + \frac{b_1}{r^2}\cos\theta, \qquad \text{for } r \geq R. \tag{4.24}$$

In order to find the remaining coefficient b_1, we have to use the boundary condition (20) at $r = R$:

$$\left.\frac{\partial\phi}{\partial r}\right|_{r=R} = \left(-\frac{j_0}{\sigma} - \frac{2b_1}{R^3}\right)\cos\theta = 0\,. \tag{4.25}$$

This gives $b_1 = -j_0R^3/2\sigma$, so, finally,

$$\phi(r,\theta) = -\frac{j_0}{\sigma}\left(r + \frac{R^3}{2r^2}\right)\cos\theta, \qquad \text{for } r \geq R\,. \tag{4.26}$$

(Note that this potential distribution corresponds to the dipole moment $\mathbf{p} = -\mathbf{E}_0R^3/2$. It is straightforward to check that if the spherical cavity was cut in a dielectric, the potential distribution outside it would be similar, with $\mathbf{p} = -\mathbf{E}_0R^3(\kappa - 1)/(\kappa + 2)$. In the limit $\kappa \to \infty$, these two results coincide, despite the rather different type of the problem: in the dielectric case, there is no current at all.)

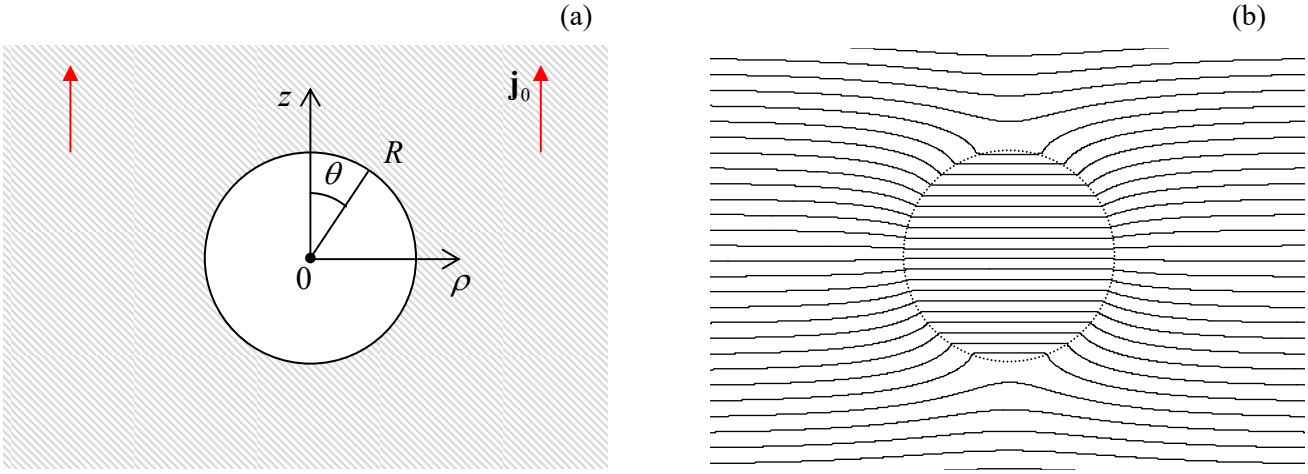


Fig. 4.7. A spherical cavity carved in a uniform conductor: (a) the problem's geometry, and (b) the equipotential surfaces as given by Eqs. (26) and (28).

Now, as the second step in the conductivity hierarchy, we may find the electrostatic potential distribution $\phi(r,\theta)$ in the insulator, in this particular case inside the empty cavity (at $r \leq R$). It should also satisfy the Laplace equation with the boundary values at $r = R$, "dictated" by the distribution (26):

$$\phi(R,\theta) = -\frac{3}{2}\frac{j_0}{\sigma}R\cos\theta\,. \tag{4.27}$$

We could again solve this problem by the formal variable separation (keeping in the general solution (2.172) only the term proportional to a_l, which does not diverge at $r \to 0$), but if we notice that the boundary condition (27) depends on just one Cartesian coordinate, $z = R\cos\theta$, the solution may be just guessed:

$$\phi(r,\theta) = -\frac{3}{2}\frac{j_0}{\sigma}z = -\frac{3}{2}\frac{j_0}{\sigma}r\cos\theta, \qquad \text{at } r \leq R\,. \tag{4.28}$$

Indeed, it evidently satisfies the Laplace equation and the boundary condition (27), and corresponds to a constant electric field parallel to the vector $\mathbf{j}_0$ and equal to $3j_0/2\sigma$ – see Fig. 7b. Again, the cavity surface is *not* equipotential, and the electric field lines at $r \leq R$ are *not* normal to it at almost all points.

More generally, the conductivity hierarchy says that static electrical fields and charges outside conductors (e.g., electric wires) do not affect currents flowing in the wires, and it is physically very clear

why. For example, if a charge in the free space is slowly moved close to a wire, it (in accordance with the linear superposition principle) only induces an additional surface charge (see Sec. 2.1) that screens the external charge's field, without participating in the current flow inside the conductor.

Besides this conceptual issue, the two examples given above may be considered as applications of the first two methods discussed in Chapter 2 – the orthogonal coordinates (Fig. 6) and the variable separation (Fig. 7) – to dc current distribution problems. As the reader may recall, in that chapter we also discussed the method of charge images. It turns out that its analog may be also used for the solution of some dc conductivity problems. Indeed, let us consider a spherically-symmetric potential distribution of the electrostatic potential, similar to that given by the basic Eq. (1.35):

$$\phi = \frac{c}{r}. \tag{4.29}$$

As we know from Chapter 1, this is a particular solution of the 3D Laplace equation at all points but $r = 0$. In free space, this distribution would correspond to a point charge $q = 4\pi\varepsilon_0 c$; but what about a uniform Ohmic conductor? Calculating the corresponding electric field and current density,

$$\mathbf{E} = -\nabla\phi = \frac{c}{r^3}\mathbf{r}, \qquad \mathbf{j} = \sigma\mathbf{E} = \sigma\frac{c}{r^3}\mathbf{r}, \tag{4.30}$$

we see that the total current flowing from the origin through a sphere of an arbitrary radius r does not depend on the radius:

$$I = Aj = 4\pi r^2 j = 4\pi\sigma c. \tag{4.31}$$

Plugging the resulting coefficient c into Eq. (29), we get

$$\phi = \frac{I}{4\pi\sigma r}. \tag{4.32}$$

Hence the Coulomb-type distribution of the electric potential in a conductor is possible (at least at some distance from the singular point $r = 0$), and describes the dc current I flowing out of a small-size electrode – or *into* such an electrode if the coefficient c is negative. Such *current injection* may be readily implemented experimentally; think for example about an insulated wire with a small bare end, inserted into a poorly conducting soil – an important method in geophysical research. Such point injection is even simpler in 2D situations – think about a wire attached, within a small spot, to a thin resistive layer, such as the thin films used for wiring in microelectronics.[15]

Now let the current injection point $\mathbf{r}'$ be close to a plane interface between the conductor and an insulator (Fig. 8). In this case, besides the Laplace equation, we should satisfy the boundary condition,

$$j_n = \sigma E_n = -\sigma\frac{\partial\phi}{\partial n} = 0, \tag{4.33}$$

at the interface. It is clear that this can be done by replacing the insulator with an imaginary similar conductor with an additional current injection point, at the mirror image point $\mathbf{r}''$. Note, however, that in contrast to charge images, the sign of the imaginary current has to be *similar*, not opposite, to the initial one, so the total electrostatic potential inside the conducting semi-space is

[15] Note that in such layers, the current distribution near the injection point is different, $j \propto 1/r$ rather than $1/r^2$.

$$\phi(\mathbf{r}) = \frac{I}{4\pi\sigma}\left(\frac{1}{|\mathbf{r}-\mathbf{r}'|} + \frac{1}{|\mathbf{r}-\mathbf{r}''|}\right). \tag{4.34}$$

(The image current's sign would be opposite at the interface between a conductor with moderate conductivity and a perfect conductor ("electrode"), whose potential should be virtually constant.)

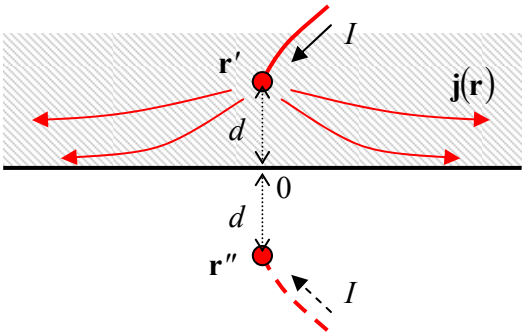


Fig. 4.8. Applying the method of images for the current injection analysis.

This result may be readily used, for example, to calculate the current density at a plane surface of a uniform conductor, as a function of distance ρ from point 0 – the surface's point closest to the current injection site – see Fig. 8. At such surface, Eq. (34) yields

$$\phi = \frac{I}{2\pi\sigma}\frac{1}{\left(\rho^2 + d^2\right)^{1/2}}, \tag{4.35}$$

so the current density is:

$$j_\rho = \sigma E_\rho = -\sigma\frac{\partial\phi}{\partial\rho} = \frac{I}{2\pi}\frac{\rho}{\left(\rho^2 + d^2\right)^{3/2}}. \tag{4.36}$$

Deviations from Eqs. (35) and (36) may be used to find and characterize conductance inhomogeneities, say, those due to mineral deposits in the Earth's crust.[16]

So, the methods used in electrostatics to calculate the potential distribution in linear dielectrics may be also used to find such distributions in Ohmic conductors. Moreover, some of these methods are more valuable in this field. For example, in electrostatics, the effective methods of solution of the 2D Laplace equation, discussed in Secs. 2.3-2.6, could be only applied to cylindrical geometries. At Ohmic conduction, this equation is also valid in some 3D cases. A practically important example is the current flow in thin resistive layers where, due to the conductivity hierarchy principle, the 3D-distributed field outside a layer, induced by the 2D-distributed current in it, does not affect the flow and in many cases is not important. A few problems of this kind, formulated in Sec. 5, are left for the reader's exercise.

4.4. Energy dissipation

Let me conclude this brief chapter with an ultra-short discussion of energy dissipation in conductors. In contrast to the electrostatic situations in insulators (vacuum or dielectrics), at dc

[16] The current injection may be also produced, due to electrochemical reactions, by an ore mass itself, so one need only measure (and correctly interpret :-) the resulting potential distribution – the so-called *self-potential method* – see, e.g., Sec. 6.1 in W. Telford *et al.*, *Applied Geophysics*, 2nd ed., Cambridge U. Press, 1990.

conduction, the electrostatic energy U is "dissipated" (i.e. transferred to heat) at a certain rate $\mathscr{P} \equiv -dU/dt$, with the dimensionality of power.[17] The rate of this *energy dissipation* may be evaluated by calculating the power of the electric field's work on a single moving charge:

$$\mathscr{P}_1 = \mathbf{F} \cdot \mathbf{v} = q\mathbf{E} \cdot \mathbf{v} . \tag{4.37}$$

After the summation over all charges, Eq. (37) gives us the average power of energy dissipation. If the charge density n is uniform, multiplying by it both parts of this relation, and taking into account that $qn\mathbf{v} = \mathbf{j}$, for the energy dissipation rate in a unit volume we get the following *Joule law*[18]

General Joule law

$$\mathscr{p} \equiv \frac{\mathscr{P}}{V} = \frac{\mathscr{P}_1 N}{V} = \mathscr{P}_1 n = q\mathbf{E} \cdot \mathbf{v}n = \mathbf{E} \cdot \mathbf{j} . \tag{4.38}$$

In the case of the Ohmic conductivity (8), this expression may be also rewritten in two other forms:

Joule law for Ohmic conductivity

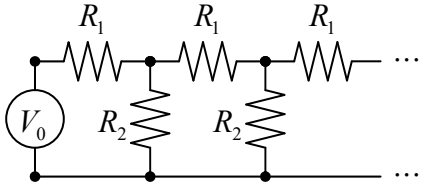
$$\mathscr{p} = \sigma E^2 = \frac{j^2}{\sigma} . \tag{4.39}$$

With our electrostatics background, it is also straightforward (and hence left for the reader's exercise) to prove that the dc current distribution in a uniform Ohmic conductor, at a fixed voltage distribution along its surface, corresponds to the minimum of the total dissipation in the sample,

$$\mathscr{P} \equiv \int_V \mathscr{p} d^3 r = \sigma \int_V E^2 d^3 r . \tag{4.40}$$

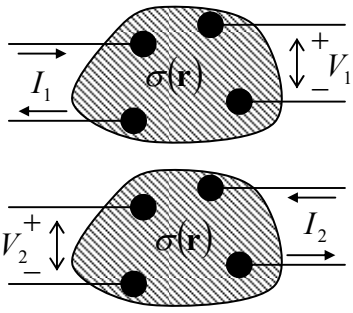
4.5. Exercise problems

4.1. DC voltage V_0 is applied to the end of a semi-infinite chain of lumped Ohmic resistors, shown in the figure on the right. Calculate the voltage across the j^{th} link of the chain.



4.2. It is well known that properties of many dc current sources (e.g., batteries) may be reasonably well represented as a connection in series of a *perfect voltage source* and an Ohmic *internal resistance*. Discuss the option, and possible advantages, of using a different equivalent circuit that would include a *perfect current source*.

4.3. Prove the following *Rayleigh-Lorentz-Carson reciprocity relation*: the results of the two separate experiments shown schematically in the figure on the right, with an arbitrary Ohmic conductor with four electrodes/terminals, are related as $I_1 V_2 = I_2 V_1$.



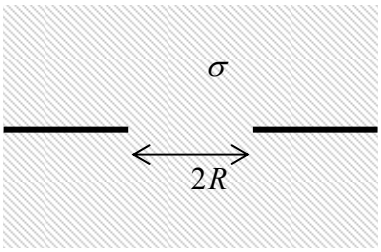
Hint: Try to apply the same approach as was used to prove Green's reciprocity relation of electrostatics in Problem 1.18, but with proper modifications.

[17] If this electric field and hence the electrostatic energy are time-independent, the energy is replenished at the same rate from the current source(s).

[18] Named after James Prescott Joule, who quantified this effect in 1841.

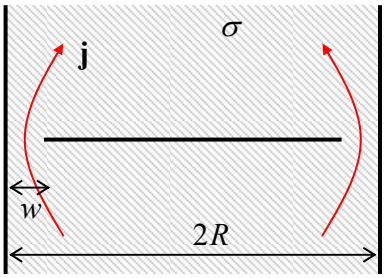
4.4. Calculate the resistance between two large uniform Ohmic conductors separated by a very thin, plane, insulating partition with a circular hole of radius R in it – see the figure on the right.

Hint: You may like to use the oblate spheroidal coordinates that were discussed in Sec. 2.4.



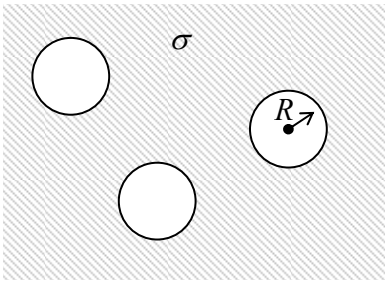
4.5. A very narrow plane crack inside a round conducting wire of radius R does not reach its surface by a small distance w – see the figure on the right. Assuming that the Ohmic conductivity σ of the wire's material is otherwise constant, calculate the electric resistance of the obstacle in the first approximation in small $w/R << 1$.

Hint: You may like to use the same elliptic coordinates as were employed at the solution of Problem 2.12.



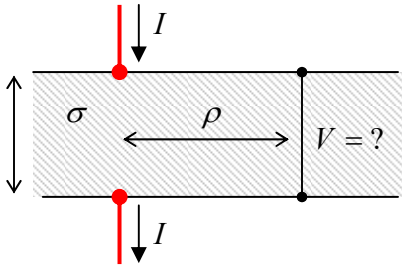
4.6. Calculate the effective (average) conductivity σ_{ef} of a medium with many empty spherical cavities of radius R, carved at random positions in a uniform Ohmic conductor (see the figure on the right), in the limit of a low density $n << R^{-3}$ of the cavities.

Hint: You may like to use the analogy with an electric-dipole medium – see, e.g., Sec. 3.2.

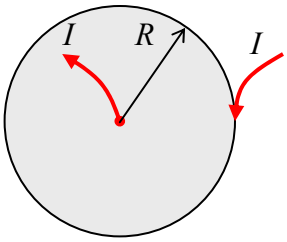


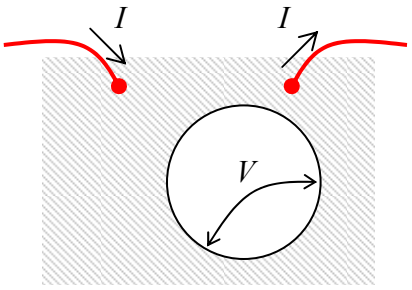
4.7. In two separate experiments, a narrow gap, possibly of irregular width, between two close, perfectly conducting electrodes is filled with some material: in the first case, a uniform linear dielectric with an electric permittivity ε, and in the second case, a uniform conducting material with an Ohmic conductivity σ. Neglecting the fringe effects, calculate the relation between the mutual capacitance C between the electrodes (in the first case) and the dc resistance R between them (in the second case).

4.8. Calculate the voltage V across a uniform, wide resistive slab of thickness t, at distance ρ from the points of injection/pickup of the dc current I passed across the slab – see the figure on the right.

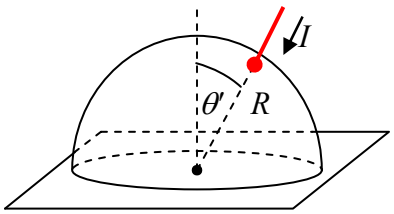


4.9. Calculate the distribution of the dc current's density in a thin, round, uniform resistive disk, if the current is inserted into some point at the disk's rim, and picked up in its center – see the figure on the right.



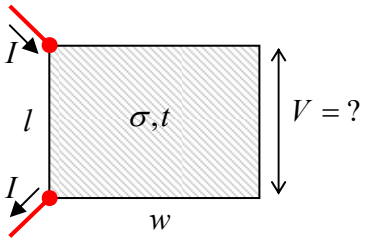


4.10. DC current is passed between two point electrodes connected to a wide, thin, uniform resistive sheet – see the figure on the right. Use the model solution of the previous problem to prove, without much new calculation, that cutting a round hole in the sheet (outside of the current injection/extraction points) doubles the voltage between any two points on its border.



4.11. The rim of a hemispherical thin shell, of radius R and thickness $t << R$, made of a uniform Ohmic conductor, is connected to a plane ground electrode. Calculate the distribution of the electrostatic potential created in the shell by a dc current I injected into it through a small-size electrode located at a polar angle $\theta' < \pi/2$ from the symmetry axis – see the figure on the right.

Hint: You may like to use the variable substitution $\rho \equiv \tan(\theta/2)$ to map the hemisphere onto a unit circle.



4.12. A rectangle of area $l \times w$ is cut from a uniform resistive sheet of thickness $t << l, w$. Use two different approaches to calculate the voltage V between its two adjacent corners, induced by the dc current I passed between the two other corners – see the figure on the right.

Hint: Besides the charge/current image method, you may like to consider using the variable separation method, with due respect to the current injection/extraction points.

4.13.* The simplest reasonable model of a vacuum diode consists of two parallel planar metallic electrodes of area A, separated by a gap of thickness $d << A^{1/2}$: a "cathode" that emits electrons into the gap, and an "anode" that absorbs the electrons arriving from the gap at its surface. Calculate the dc I-V curve of the diode, i.e. the relation between the average current I flowing between the electrodes and the dc voltage V applied between them, using the following simplifying assumptions:

(i) due to the effect of the negative space charge of the emitted electrons, the current I is much lower than the emission ability of the cathode,

(ii) the initial velocity of the emitted electrons is negligible, and

(iii) the direct Coulomb interaction of electrons (besides the space charge effect) is negligible.

4.14.* Calculate the space-charge-limited current in a system with the same geometry as in the previous problem, and using the same assumptions besides that now the emitted charge carriers do not fly ballistically, but rather drift in accordance with the Ohm law, with the conductivity given by Eq. (13): $\sigma = q^2 \mu n$, with a constant mobility μ.[19]

Hint: In order to get a realistic result, assume that the medium in which the charge carriers move has a certain dielectric constant κ unrelated to the carriers.

[19] As was mentioned in Sec. 2, the approximation of a constant (in particular, field- and charge-density-independent) mobility is most suitable for semiconductors.

4.15. Prove that the distribution of dc currents in a uniform Ohmic conductor with a given voltage distribution along its surface corresponds to the minimum of the total energy dissipation rate ("Joule heat").

Chapter 5. Magnetism

Even though this chapter addresses a completely new type of electric charge interaction, its discussion (for the stationary case) will take not too much time/space, because it recycles many ideas and methods of electrostatics, though with a twist or two.

5.1. Magnetic interaction of currents

DC currents in conductors usually leave them *electroneutral*, $\rho(\mathbf{r}) = 0$, with very good precision, because even a minute imbalance of positive and negative charge density results in extremely strong Coulomb forces that restore the electroneutrality by a very fast additional shift of free charge carriers. This is why let us start the discussion of magnetism from the simplest case of two spatially separated, dc-current-carrying, electroneutral conductors (Fig. 1).

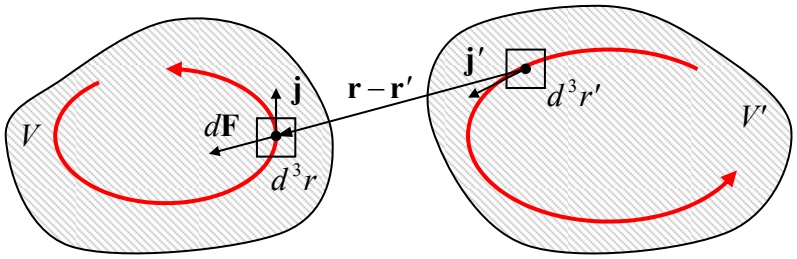


Fig. 5.1. Magnetic interaction of two currents.

According to the Coulomb law, there is no electrostatic force between them. However, several experiments carried out in 1820[1] proved that there is a different, *magnetic* interaction between the currents. In the present-day notation, the results of all such experiments may be summarized with just one formula, in SI units expressed as

Magnetic force

$$\mathbf{F} = -\frac{\mu_0}{4\pi}\int_V d^3r \int_{V'} d^3r' \left[\mathbf{j}(\mathbf{r})\cdot\mathbf{j}'(\mathbf{r}')\right]\frac{\mathbf{r}-\mathbf{r}'}{\left|\mathbf{r}-\mathbf{r}'\right|^3}. \qquad (5.1)$$

Here the coefficient $\mu_0/4\pi$ (where μ_0 is called either the *magnetic constant* or the *free space permeability*) equals *almost exactly* 10^{-7} SI units, with the product $\varepsilon_0\mu_0$ equal to *exactly* $1/c^2$.[2]

Note a close similarity of this expression to the Coulomb law (1.1) rewritten for the interaction of two continuously distributed charges, with the account of the linear superposition principle (1.4):

Electric force

$$\mathbf{F} = \frac{1}{4\pi\varepsilon_0}\int_V d^3r \int_{V'} d^3r'\, \rho(\mathbf{r})\rho'(\mathbf{r}')\frac{\mathbf{r}-\mathbf{r}'}{\left|\mathbf{r}-\mathbf{r}'\right|^3}. \qquad (5.2)$$

[1] Most notably, by Hans Christian Ørsted who discovered the effect of electric currents on magnetic needles, and André-Marie Ampère who extended this work by finding the magnetic interaction between two currents.

[2] For details, see Appendix *UCA: Selected Units and Constants*. In the Gaussian units, the coefficient $\mu_0/4\pi$ in Eq. (1) and beyond is replaced with $1/c^2$.

Besides the different coefficient and a different sign, the "only" difference of Eq. (1) from Eq. (2) is the scalar product of the current densities, evidently necessary because of their vector character. We will see soon that this difference brings certain complications in applying the approaches discussed in the previous chapters, to magnetostatics.

Before going to their discussion, let us have one more glance at the coefficients in Eqs. (1) and (2). To compare them, let us consider two objects with *uncompensated* charge distributions $\rho(\mathbf{r})$ and $\rho'(\mathbf{r})$, moving parallel to each other with certain velocities $\mathbf{v}$ and $\mathbf{v}'$, as measured in the same inertial ("laboratory") reference frame. In this case, $\mathbf{j}(\mathbf{r}) = \rho(\mathbf{r})\mathbf{v}$, so $\mathbf{j}(\mathbf{r})\cdot\mathbf{j}'(\mathbf{r}) = \rho(\mathbf{r})\rho'(\mathbf{r})vv'$, and the integrals in Eqs. (1) and (2) become functionally similar, differing only by the factor

$$\frac{F_{\text{magnetic}}}{F_{\text{electric}}} = -\frac{\mu_0 vv'}{4\pi}\Big/\frac{1}{4\pi\varepsilon_0} \equiv -\frac{vv'}{c^2}. \tag{5.3}$$

(The last expression is valid in any consistent system of units.) We immediately see that magnetism is an essentially relativistic phenomenon, very weak in comparison with the electrostatic interaction at the human scale velocities, $v << c$, and may dominate only if the latter interaction vanishes – as it does in electroneutral systems.[3] The discovery and initial studies[4] of such a subtle, relativistic phenomenon as magnetism were much facilitated by the relative abundance of natural *ferromagnets*: materials with a spontaneous magnetic polarization, whose strong magnetic field is due to relativistic effects (such as spin) inside the constituent atoms – see Sec. 5 below.

Also, Eq. (3) points to an interesting paradox. Consider two electron beams moving parallel to each other, with the same velocity v with respect to a lab reference frame. Then, according to Eq. (3), the net force of their total (electric plus magnetic) interaction is proportional to $(1 - v^2/c^2)$, tending to zero in the limit $v \to c$. However, in the reference frame moving together with the electrons, they are not moving at all, i.e. $v = 0$. Hence, from the point of view of such a moving observer, the electron beams should interact only electrostatically, with a repulsive force independent of the velocity v. Historically, this had been one of several paradoxes that led to the development of special relativity; its resolution will be discussed in Chapter 9 devoted to this theory.

Returning to Eq. (1), in some simple cases the double integration in it may be carried out analytically. First of all, let us simplify this expression for the case of two thin, long conductors ("wires") separated by a distance much larger than their thickness. In this case, we may integrate the products $\mathbf{j}d^3r$ and $\mathbf{j}'d^3r'$ over the wires' cross-sections first, neglecting the corresponding change of the factor $(\mathbf{r} - \mathbf{r}')$. Since the integrals of the current density over the cross-sections of the wires are just the currents I and I' flowing in the wires, and cannot change along their lengths (say, l and l', respectively), they may be taken out of the remaining integrals, reducing Eq. (1) to

$$\mathbf{F} = -\frac{\mu_0 II'}{4\pi}\oint_l\oint_{l'}(d\mathbf{r}\cdot d\mathbf{r}')\frac{\mathbf{r}-\mathbf{r}'}{|\mathbf{r}-\mathbf{r}'|^3}. \tag{5.4}$$

[3] An important case when the electroneutrality may not hold is the motion of electrons in free space. (However, in this case, the electron speed is often comparable with the speed of light, so the magnetic forces may be comparable in strength with electrostatic forces, and hence important.) Minor local violations of electroneutrality also play an important role in some semiconductor devices – see, e.g., SM Chapter 6.

[4] The first detailed book on this subject, *De Magnete* by William Gilbert (a.k.a. Gilberd), was published as early as 1600.

As the simplest example, consider two straight, parallel wires (Fig. 2) separated by distance d, both with length $l >> d$.

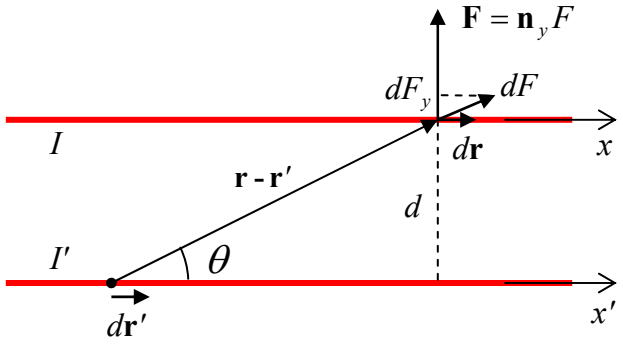


Fig. 5.2. The magnetic force between two straight parallel currents.

Due to the symmetry of this system, the vector of the magnetic interaction force has to:

(i) lie in the same plane as the currents, and
(ii) be normal to the wires – see Fig. 2.

Hence we may limit our calculations to just one component of the force – normal to the wires. Using the fact that with the coordinate choice shown in Fig. 2, the scalar product $d\mathbf{r}\cdot d\mathbf{r}'$ is just $dxdx'$, we get

$$F = -\frac{\mu_0 II'}{4\pi}\int_{-\infty}^{+\infty}dx\int_{-\infty}^{+\infty}dx'\frac{\sin\theta}{d^2+(x-x')^2} = -\frac{\mu_0 II'}{4\pi}\int_{-\infty}^{+\infty}dx\int_{-\infty}^{+\infty}dx'\frac{d}{\left[d^2+(x-x')^2\right]^{3/2}}. \tag{5.5}$$

Now introducing, instead of x', a new, dimensionless variable $\xi \equiv (x - x')/d$, we may reduce the internal integral to a table one, which we have already encountered in this course:

$$F = -\frac{\mu_0 II'}{4\pi d}\int_{-\infty}^{+\infty}dx\int_{-\infty}^{+\infty}\frac{d\xi}{\left(1+\xi^2\right)^{3/2}} = -\frac{\mu_0 II'}{2\pi d}\int_{-\infty}^{+\infty}dx\,. \tag{5.6}$$

The integral over x formally diverges, but it gives a finite interaction force *per unit length* of the wires:

$$\frac{F}{l} = -\frac{\mu_0 II'}{2\pi d}\quad. \tag{5.7}$$

Note that the force drops rather slowly (only as $1/d$) as the distance d between the wires is increased, and is attractive (rather than repulsive as in the Coulomb law) if the currents are of the same sign.

This is an important result,[5] but again, the problems so simply solvable are few and far between, and it is intuitively clear that we would strongly benefit from the same approach as in electrostatics, i.e., from decomposing Eq. (1) into a product of two factors via the introduction of a suitable field. Such decomposition may be done as follows:

Lorentz force: current

$$\mathbf{F} = \int_V \mathbf{j}(\mathbf{r})\times\mathbf{B}(\mathbf{r})d^3r\,, \tag{5.8}$$

[5] In particular, until very recently (2018), Eq. (7) was used for the legal definition of the SI unit of current, the ampere (A), via the SI unit of force (the newton, N), with the coefficient μ_0 considered exactly fixed. (A brief description of the recent changes in legal metrology is given in Appendix UCA.)

where the vector **B** is called the *magnetic field*.[6] In the case when it is induced by the current **j**':

$$\mathbf{B}(\mathbf{r}) \equiv \frac{\mu_0}{4\pi}\int_{V'}\mathbf{j}'(\mathbf{r}')\times\frac{\mathbf{r}-\mathbf{r}'}{|\mathbf{r}-\mathbf{r}'|^3}d^3r'. \qquad (5.9)$$

Biot-Savart law

The last relation is called the *Biot-Savart law*,[7] while the force **F** expressed by Eq. (8) is sometimes called the *Lorentz force*.[8] However, more frequently the latter term is reserved for the full force,

$$\mathbf{F} = q\left(\mathbf{E} + \mathbf{v}\times\mathbf{B}\right), \qquad (5.10)$$

Lorentz force: particle

exerted by electric and magnetic fields field on a point charge q, moving with velocity **v**.[9]

Now we have to prove that the new formulation, given by Eqs. (8)-(9), is equivalent to Eq. (1). At first glance, this seems unlikely. Indeed, first of all, Eqs. (8) and (9) involve vector products, while Eq. (1) is based on a scalar product. More profoundly, in contrast to Eq. (1), Eqs. (8) and (9) do *not* satisfy the 3rd Newton's law applied to elementary current components $\mathbf{j}d^3r$ and $\mathbf{j}'d^3r'$, if these vectors are not parallel to each other. Indeed, consider the situation shown in Fig. 3.

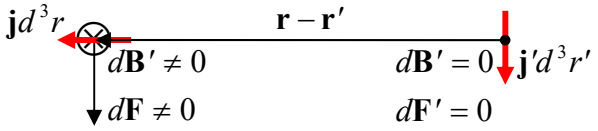


Fig. 5.3. The apparent violation of the 3rd Newton law in magnetism.

Here the vector **j**' is perpendicular to the vector (**r** – **r**'), and hence, according to Eq. (9), produces a non-zero contribution $d\mathbf{B}'$ to the magnetic field directed (in Fig. 3) normally to the plane of the drawing, i.e. perpendicular to the vector **j**. Hence, according to Eq. (8), this field provides a non-zero contribution to **F**. On the other hand, if we calculate the reciprocal force **F**' by swapping the prime indices in Eqs. (8) and (9), the latter equation immediately shows that $d\mathbf{B}(\mathbf{r}') \propto \mathbf{j}\times(\mathbf{r}'-\mathbf{r}) = 0$, because the two operand vectors are parallel – see Fig. 3 again. Hence, the current component $\mathbf{j}'d^3\mathbf{r}'$ does exert a force on its counterpart, while $\mathbf{j}d^3\mathbf{r}$ does not.

---

[6] The SI unit of the magnetic field is called *tesla* (T) – after Nikola Tesla, a pioneer of electrical engineering. In the Gaussian units, the already discussed constant $1/c^2$ in Eq. (1) is equally divided between Eqs. (8) and (9), so in them both, the constant before the integral is $1/c$. The resulting Gaussian unit of the field **B** is called *gauss* (G); taking into account the difference of units of electric charge and length, and hence of the current density, 1 G equals exactly 10^{-4} T. Note also that in some textbooks, especially old ones, **B** is called either the *magnetic induction* or the *magnetic flux density*, while the term "magnetic field" is reserved for the field **H** that will be introduced in Sec. 5 below.

[7] Named after Jean-Baptiste Biot and Félix Savart who made several key contributions to the theory of magnetic interactions – in the same notorious 1820.

[8] Named after Hendrik Antoon Lorentz, famous mostly for his numerous contributions to the development of special relativity – see Chapter 9 below. To be fair, the magnetic part of the Lorentz force was implicitly described in a much earlier (1865) paper by J. C. Maxwell and then spelled out by Oliver Heaviside (another genius of electrical engineering – and mathematics!) in 1889, i.e. also before the 1895 work by H. Lorentz.

[9] From the magnetic part of Eq. (10), Eq. (8) may be derived by the elementary summation of all forces acting on $n >> 1$ particles in a unit volume, with $\mathbf{j} = qn\mathbf{v}$ – see the footnote on Eq. (4.13a). On the other hand, the reciprocal derivation of Eq. (10) from Eq. (8) with $\mathbf{j} = q\mathbf{v}\delta(\mathbf{r}-\mathbf{r}_0)$, where $\mathbf{r}_0$ is the current particle's position (so $d\mathbf{r}_0/dt = \mathbf{v}$), requires certain mathematical care and will be performed in Chapter 9.

Despite this apparent problem, let us still go ahead and plug Eq. (9) into Eq. (8):

$$\mathbf{F} = \frac{\mu_0}{4\pi}\int_V d^3r \int_{V'} d^3r'\, \mathbf{j}(\mathbf{r}) \times \left(\mathbf{j}'(\mathbf{r}') \times \frac{\mathbf{r}-\mathbf{r}'}{\left|\mathbf{r}-\mathbf{r}'\right|^3} \right). \tag{5.11}$$

This double vector product may be transformed into two scalar products, using the vector algebraic identity called the *bac minus cab rule*, $\mathbf{a}\times(\mathbf{b}\times\mathbf{c}) = \mathbf{b}(\mathbf{a}\cdot\mathbf{c}) - \mathbf{c}(\mathbf{a}\cdot\mathbf{b})$.[10] Applying this relation, with $\mathbf{a} = \mathbf{j}$, $\mathbf{b} = \mathbf{j}'$, and $\mathbf{c} = \mathbf{R} \equiv \mathbf{r} - \mathbf{r}'$, to Eq. (11), we get

$$\mathbf{F} = \frac{\mu_0}{4\pi}\int_{V'} d^3r' \mathbf{j}'(\mathbf{r}') \left(\int_V d^3r \frac{\mathbf{j}(\mathbf{r})\cdot\mathbf{R}}{R^3} \right) - \frac{\mu_0}{4\pi}\int_V d^3r \int_{V'} d^3r'\, \mathbf{j}(\mathbf{r})\cdot\mathbf{j}'(\mathbf{r}')\frac{\mathbf{R}}{R^3}. \tag{5.12}$$

The second term on the right-hand side of this equality coincides with the right-hand side of Eq. (1), while the first term equals zero because its internal integral vanishes. Indeed, we may break the volumes V and V' into narrow *current tubes* – the stretched elementary volumes whose walls are not crossed by current lines (so on their walls, $j_n = 0$). As a result, the elementary current in each tube, $dI = jdA = jd^2r$, is the same along its length, and, just as in a thin wire, $\mathbf{j}d^2r$ may be replaced with $dId\mathbf{r}$, with the vector $d\mathbf{r}$ directed along $\mathbf{j}$. Because of this, each tube's contribution to the internal integral in the first term of Eq. (12) may be represented as

$$dI \oint_l d\mathbf{r}\cdot\frac{\mathbf{R}}{R^3} = -dI \oint_l d\mathbf{r}\cdot\nabla\frac{1}{R} = -dI \oint_l dr \frac{\partial}{\partial r}\frac{1}{R}, \tag{5.13}$$

where the operator ∇ acts in the $\mathbf{r}$-space, and the integral is taken along the tube's length l. Due to the current continuity expressed by Eq. (4.6), each loop should follow a closed contour, and an integral of a full differential of some scalar function (in our case, of $1/R$) along such contour equals zero.

So we have recovered Eq. (1). Returning for a minute to the paradox illustrated in Fig. 3, we may conclude that the apparent violation of the 3rd Newton law was the artifact of our interpretation of Eqs. (8) and (9) as the sums of independent elementary components. In reality, due to the dc current continuity, these components are *not* independent. For the whole currents, Eqs. (8)-(9) do obey the 3rd law – as follows from their already proved equivalence to Eq. (1).

Thus it is possible to break the magnetic interaction into two effects: the induction of the magnetic field $\mathbf{B}$ by one current (in our notation, $\mathbf{j}'$), and the effect of this field on the other current ($\mathbf{j}$). Now comes an additional experimental fact: other elementary components $\mathbf{j}d^3r'$ of the current $\mathbf{j}(\mathbf{r})$ also contribute to the magnetic field (9) acting on the component $\mathbf{j}d^3r$.[11] This fact allows us to drop the prime sign after $\mathbf{j}$ in Eq. (9), and rewrite Eqs. (8) and (9) as

$$\mathbf{B}(\mathbf{r}) = \frac{\mu_0}{4\pi}\int_{V'} \mathbf{j}(\mathbf{r}')\times\frac{\mathbf{r}-\mathbf{r}'}{\left|\mathbf{r}-\mathbf{r}'\right|^3} d^3r', \tag{5.14}$$

$$\mathbf{F} = \int_V \mathbf{j}(\mathbf{r})\times\mathbf{B}(\mathbf{r}) d^3r. \tag{5.15}$$

[10] See, e.g., MA Eq. (7.5).

[11] Just as in electrostatics, one needs to exercise due caution in transforming these expressions for the limit of discrete classical particles, and extended wavefunctions in quantum mechanics, to avoid the (non-existing) magnetic interaction of a charged particle with itself.

Again, the field *observation* point **r** and the field *source* point **r**' have to be clearly distinguished. We immediately see that these expressions are close to, but still different from the corresponding relations of the electrostatics, namely Eq. (1.9) and the distributed-charge version of Eq. (1.6):

$$\mathbf{E}(\mathbf{r}) = \frac{1}{4\pi\varepsilon_0}\oint_{V'}\rho(\mathbf{r}')\frac{\mathbf{r}-\mathbf{r}'}{|\mathbf{r}-\mathbf{r}'|^3}d^3r', \tag{5.16}$$

$$\mathbf{F} = \oint_{V}\rho(\mathbf{r})\mathbf{E}(\mathbf{r})d^3r\ . \tag{5.17}$$

(Note that the sign difference has disappeared, at the cost of the replacement of scalar-by-vector multiplications in electrostatics with cross-products of vectors in magnetostatics.)

For the frequent particular case of a thin wire of length *l*', Eq. (14) may be re-written as

$$\mathbf{B}(\mathbf{r}) = \frac{\mu_0 I}{4\pi}\oint_{l'} d\mathbf{r}'\times\frac{\mathbf{r}-\mathbf{r}'}{|\mathbf{r}-\mathbf{r}'|^3}. \tag{5.18}$$

Let us see how this formula works for the simplest case of a straight wire (Fig. 4a). The magnetic field contributions *d***B** due to all small fragments *d***r**' of the wire's length are directed along the same line (perpendicular to both the wire and the shortest distance *d* from the observation point to the wire's line), and its magnitude is

$$dB = \frac{\mu_0 I}{4\pi}\frac{dx'}{|\mathbf{r}-\mathbf{r}'|^2}\sin\theta = \frac{\mu_0 I}{4\pi}\frac{dx'}{\left(d^2+x^2\right)}\frac{d}{\left(d^2+x^2\right)^{1/2}}. \tag{5.19}$$

Summing up all such elementary contributions, we get

$$B = \frac{\mu_0 I\rho}{4\pi}\int_{-\infty}^{\infty}\frac{dx}{(x^2+d^2)^{3/2}} = \frac{\mu_0 I}{2\pi d}. \tag{5.20}$$

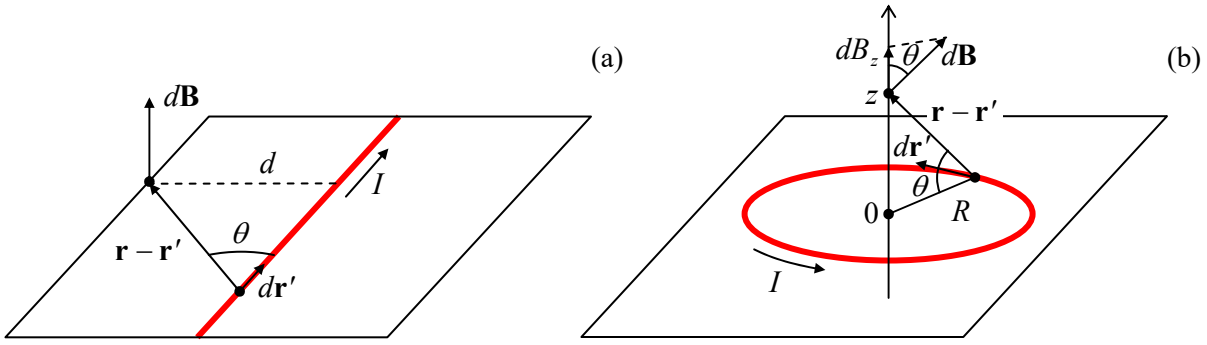


Fig. 5.4. Calculating magnetic fields: (a) of a straight current, and (b) of a current loop.

This is a simple but important result. (Note that it is only valid for very long ($l >> d$), straight wires.) It is especially crucial to note the "vortex" character of the field: its lines go around the wire, forming rings with the centers on the current line. This is in sharp contrast to the electrostatic field lines, which can only begin and end on electric charges and never form closed loops (otherwise the Coulomb force $q\mathbf{E}$ would not be conservative). In the magnetic case, the vortex structure of the *field* may be reconciled with the potential character of the magnetic *forces*, which is evident from Eq. (1), due to the vector products in Eqs. (14)-(15).

Now we may readily use Eq. (15), or rather its thin-wire version

$$\mathbf{F} = I \oint_l d\mathbf{r} \times \mathbf{B}(\mathbf{r}), \tag{5.21}$$

to apply Eq. (20) to the two-wire problem (Fig. 2). Since for the second wire, the vectors d**r** and **B** are perpendicular to each other, we immediately arrive at our previous result (7), which was obtained directly from Eq. (1).

The next important example of the application of the Biot-Savart law (14) is the magnetic field at the axis of a circular current loop (Fig. 4b). Due to the problem's symmetry, the net field **B** has to be directed along the axis, but each of its elementary components d**B** is tilted by the angle $\theta = \tan^{-1}(z/R)$ to this axis, so its axial component is

$$dB_z = dB\cos\theta = \frac{\mu_0 I}{4\pi}\frac{dr'}{R^2+z^2}\frac{R}{\left(R^2+z^2\right)^{1/2}}. \tag{5.22}$$

Since the denominator of this expression remains the same for all wire components *dr'*, the integration over **r'** is easy ($\int dr' = 2\pi R$), giving finally

$$B = \frac{\mu_0 I}{2}\frac{R^2}{\left(R^2+z^2\right)^{3/2}}. \tag{5.23}$$

Note that the magnetic field in the loop's center (i.e., for $z = 0$),

$$B = \frac{\mu_0 I}{2R}, \tag{5.24}$$

is π times higher than that due to a similar current in a straight wire, at the distance $d = R$ from it. This difference is readily understandable, since all elementary components of the loop are at the same distance R from the observation point, while in the case of a straight wire, all its points but one are separated from the observation point by distances larger than d.

Another notable fact is that at large distances ($z^2 >> R^2$), the field (23) is proportional to z^{-3}:

$$B \approx \frac{\mu_0 I}{2}\frac{R^2}{|z|^3} \equiv \frac{\mu_0}{4\pi}\frac{2m}{|z|^3}, \qquad \text{with } m \equiv IA, \tag{5.25}$$

where $A = \pi R^2$ is the loop area. Comparing this expression with Eq. (3.13), for the particular case $\theta = 0$, we see that such field is similar to that of an electric dipole (at least along its direction), with the replacement of the electric dipole moment magnitude p with the m so defined – besides the front factor. Indeed, such a plane current loop is the simplest example of a system whose field, at distances much larger than R, is that of a *magnetic dipole*, with a *dipole moment* **m** – the notions to be discussed in much more detail in Sec. 4 below.

5.2. Vector potential and the Ampère law

The reader could see that the calculations of the magnetic field using Eq. (14) or (18) are still somewhat cumbersome even for the very simple systems we have examined. As we saw in Chapter 1, similar calculations in electrostatics, at least for several important highly symmetric systems, could be

substantially simplified using the Gauss law (1.16). A similar relation exists in magnetostatics as well, but has a different form, due to the vortex character of the magnetic field.

To derive it, let us notice that in an analogy with the scalar case, the vector product under the integral (14) may be transformed as

$$\frac{j(\mathbf{r}')\times(\mathbf{r}-\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|^3}=\nabla\times\frac{\mathbf{j}(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|}, \tag{5.26}$$

where the operator ∇ acts in the $\mathbf{r}$-space. (This equality may be readily verified by Cartesian components, noticing that the current density is a function of $\mathbf{r}'$ and hence its components are independent of $\mathbf{r}$.) Plugging Eq. (26) into Eq. (14), and moving the operator ∇ out of the integral over $\mathbf{r}'$, we see that the magnetic field may be represented as the curl of another vector field – the so-called *vector potential*, defined as:[12]

$$\mathbf{B}(\mathbf{r})\equiv\nabla\times\mathbf{A}(\mathbf{r}), \tag{5.27}$$

and in our current case equal to

Vector potential

$$\mathbf{A}(\mathbf{r})=\frac{\mu_0}{4\pi}\int_{V'}\frac{\mathbf{j}(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|}d^3r'. \tag{5.28}$$

Please note a beautiful analogy between Eqs. (27)-(28) and, respectively, Eqs. (1.33) and (1.38).[13] This analogy implies that the vector potential $\mathbf{A}$ plays, for the magnetic field, essentially the same role as the scalar potential ϕ plays for the electric field (hence the name "potential"), with due respect to the vortex character of $\mathbf{B}$. This notion will be discussed in more detail below.

Now let us see what equations we may get for the spatial derivatives of the magnetic field. First, vector algebra says that the divergence of any curl is zero.[14] In application to Eq. (27), this means that

$$\nabla\cdot\mathbf{B}=0. \tag{5.29}$$

No magnetic monopoles

Comparing this equation with Eq. (1.27), we see that Eq. (29) may be interpreted as the absence of a magnetic analog of an electric charge, on which magnetic field lines could originate or end. Numerous searches for such hypothetical magnetic charges, called *magnetic monopoles*, using very sensitive and sophisticated experimental setups,[15] have not given any reliable evidence of their existence in Nature.

Proceeding to the alternative, vector derivative of the magnetic field, i.e. to its curl, and using Eq. (28), we obtain

$$\nabla\times\mathbf{B}(\mathbf{r})=\frac{\mu_0}{4\pi}\nabla\times\left(\nabla\times\int_{V'}\frac{\mathbf{j}(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|}d^3r'\right). \tag{5.30}$$

This expression may be simplified by using the following general vector identity:[16]

$$\nabla\times(\nabla\times\mathbf{c})=\nabla(\nabla\cdot\mathbf{c})-\nabla^2\mathbf{c}, \tag{5.31}$$

applied to vector $\mathbf{c}(\mathbf{r})\equiv\mathbf{j}(\mathbf{r}')/|\mathbf{r}-\mathbf{r}'|$:

[12] In the Gaussian units, Eq. (27) remains the same, and hence in Eq. (28), $\mu_0/4\pi$ is replaced with $1/c$.
[13] In Eq. (1.38), there was no real need for the additional clarification provided by the integration volume label V'.
[14] See, e.g., MA Eq. (11.2).
[15] For a recent example, see B. Acharya *et al.*, *Nature* **602**, 63 (2022).
[16] See, e.g., MA Eq. (11.3).

$$\nabla\times\mathbf{B}=\frac{\mu_0}{4\pi}\nabla\int_{V'}\mathbf{j}(\mathbf{r}')\cdot\nabla\frac{1}{|\mathbf{r}-\mathbf{r}'|}d^3r'-\frac{\mu_0}{4\pi}\int_{V'}\mathbf{j}(\mathbf{r}')\nabla^2\frac{1}{|\mathbf{r}-\mathbf{r}'|}d^3r'. \tag{5.32}$$

As was already discussed during our study of electrostatics in Sec. 3.1,

$$\nabla^2\frac{1}{|\mathbf{r}-\mathbf{r}'|}=-4\pi\delta(\mathbf{r}-\mathbf{r}'), \tag{5.33}$$

so the last term of Eq. (32) is just $\mu_0\mathbf{j}(\mathbf{r})$. On the other hand, inside the first integral, we can replace ∇ with $(-\nabla')$, where prime means the differentiation in the space of the radius vectors $\mathbf{r}'$. Integrating that term by parts, we get

$$\nabla\times\mathbf{B}=-\frac{\mu_0}{4\pi}\nabla\oint_{S'}j_n(\mathbf{r}')\frac{1}{|\mathbf{r}-\mathbf{r}'|}d^2r'+\nabla\int_{V'}\frac{\nabla'\cdot\mathbf{j}(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|}d^3r'+\mu_0\mathbf{j}(\mathbf{r}). \tag{5.34}$$

Applying this equality to the volume V' limited by a surface S' either sufficiently distant from the field concentration or with no current crossing it, we may neglect the first term on the right-hand side of Eq. (34), while the second term always equals zero in statics, due to the dc charge continuity – see Eq. (4.6). As a result, we arrive at a very simple differential equation[17]

$$\nabla\times\mathbf{B}=\mu_0\mathbf{j}. \tag{5.35}$$

This is (the dc form of) the inhomogeneous Maxwell equation – which in magnetostatics plays a role similar to Eq. (1.27) in electrostatics. Let me display, for the first time in this course, this fundamental system of equations (at this stage, for statics only), and give the reader a minute to stare, in silence, at their beautiful symmetry – which has inspired so much of the later development of physics:

Maxwell equations: statics

$$\boxed{\begin{aligned}&\nabla\times\mathbf{E}=0, \qquad \nabla\times\mathbf{B}=\mu_0\mathbf{j},\\ &\nabla\cdot\mathbf{E}=\frac{\rho}{\varepsilon_0}, \qquad \nabla\cdot\mathbf{B}=0.\end{aligned}} \tag{5.36}$$

Their only asymmetry, two zeros on the right-hand sides (for the magnetic field's divergence and electric field's curl), is due to the absence in the Nature of magnetic monopoles and their currents. I will discuss these equations in more detail in Sec. 6.7, after the first two equations (for the fields' curls) have been generalized to their full, time-dependent versions.

Returning now to our current, more mundane but important task of calculating the magnetic field induced by simple current configurations, we can benefit from an integral form of Eq. (35). For that, let us integrate this equation over an arbitrary surface S limited by a closed contour C, and apply to the result the Stokes theorem.[18] The resulting expression,

Ampère law

$$\boxed{\oint_C\mathbf{B}\cdot d\mathbf{r}=\mu_0\oint_S j_n d^2r\equiv\mu_0 I}, \tag{5.37}$$

where I is the net electric current crossing surface S, is called the *Ampère law*.

[17] As in all earlier formulas for the magnetic field, in the Gaussian units, the coefficient μ_0 in this relation is replaced with $4\pi/c$.

[18] See, e.g., MA Eq. (12.1) with $\mathbf{f}=\mathbf{B}$.

As the first example of its application, let us return to the current in a straight wire (Fig. 4a). With the Ampère law in our arsenal, we can readily pursue an even more ambitious goal than that achieved in the previous section, namely to calculate the magnetic field both outside and inside of a wire of an arbitrary radius R, with an arbitrary (albeit axially-symmetric) current distribution $j(\rho)$ – see Fig. 5.

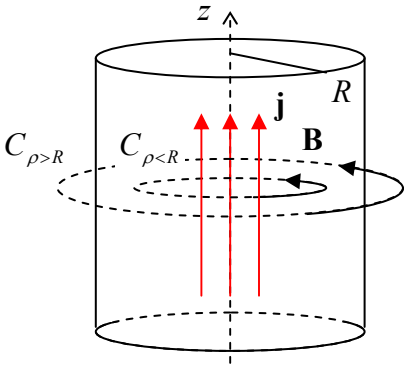


Fig. 5.5. The simplest application of the Ampère law: the magnetic field of a straight current.

Selecting the Ampère-law contour C in the form of a ring of some radius ρ in the plane normal to the wire's axis z, we have $\mathbf{B}\cdot d\mathbf{r} = B\rho\, d\varphi$, where φ is the azimuthal angle, so Eq. (37) yields:

$$2\pi\rho B(\rho) = \mu_0 \times \begin{cases} 2\pi\int_0^{\rho} j(\rho')\rho' d\rho', & \text{for } \rho \le R, \\ 2\pi\int_0^{R} j(\rho')\rho' d\rho' \equiv I, & \text{for } \rho \ge R. \end{cases} \tag{5.38}$$

Thus we have not only recovered our previous result (20), with the notation replacement $d \to \rho$, in a much simpler way but also have been able to calculate the magnetic field's distribution inside the wire. In the most common particular case when the current is uniformly distributed along its cross-section, $j(\rho)$ = const, the first of Eqs. (38) immediately yields $B \propto \rho$ for $\rho \le R$.

Another important system is a straight, long *solenoid* (Fig. 6a), with dense winding: $n^2A >> 1$, where n is the number of wire turns per unit length, and A is the area of the solenoid's cross-section.

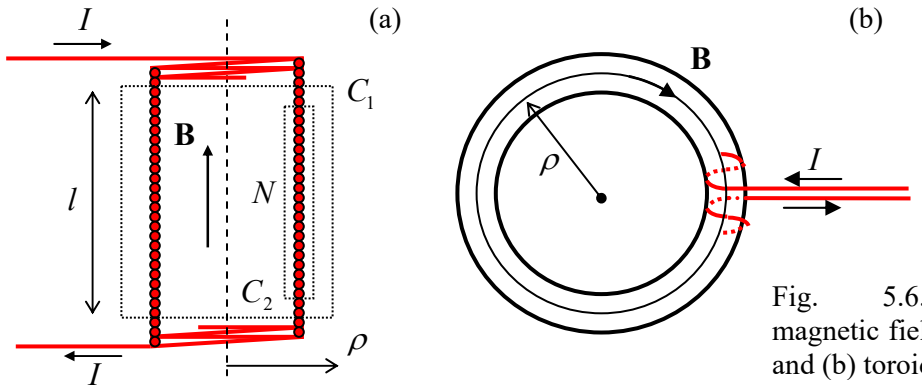


Fig. 5.6. Calculating magnetic fields of (a) straight and (b) toroidal solenoids.

From the symmetry of this problem, the longitudinal (in Fig. 6a, vertical) component B_z of the magnetic field may only depend on the distance ρ of the observation point from the solenoid's axis. First taking a plane Ampère contour C_1, with both long sides outside the solenoid, we get $B_z(\rho_2) - B_z(\rho_1) = 0$,

because the total current piercing the contour equals zero. This is only possible if $B_z = 0$ at any ρ outside of the solenoid, provided that it is infinitely long.[19] With this result on hand, from the Ampère law applied to the contour C_2, we get the following relation for the only (z-) component of the internal field:

$$Bl = \mu_0 NI \,, \tag{5.39}$$

where N is the number of wire turns passing through the contour of length l. This means that regardless of the exact position of the internal side of the contour, the result is the same:

$$B = \mu_0 \frac{N}{l} I \equiv \mu_0 nI \,. \tag{5.40}$$

Thus, the field inside an infinitely long solenoid (with an arbitrary shape of its cross-section) is uniform; in this sense, a long solenoid is a magnetic analog of a wide plane capacitor, explaining why this system is so widely used in physical experiment.

As should be clear from its derivation, the obtained results, especially that the field outside of the solenoid equals zero, are conditional on the solenoid length being very large in comparison with its lateral size. (From Eq. (25), we may predict that for a solenoid of a finite length l, the close-range external field is a factor of $\sim A/l^2$ lower than the internal one.) A much better suppression of such "fringe" fields may be obtained using *toroidal solenoids* (Fig. 6b). The application of the Ampère law to this geometry shows that in the limit of dense winding ($N >> 1$), there is no fringe field at all (for any relation between the two radii of the torus), while inside the solenoid, at distance ρ from the system's axis,

$$B = \frac{\mu_0 NI}{2\pi\rho} \,. \tag{5.41}$$

We see that a possible drawback of this system for practical applications is that the internal field does depend on ρ, i.e. is not quite uniform; however, if the torus is relatively thin, this deficiency is minor.

Next let us discuss a very important question: how can we solve the problems of magnetostatics for systems whose low symmetry does not allow getting easy results from the Ampère law? (The examples are of course too numerous to list; for example, we cannot use this approach even to reproduce Eq. (23) for a round current loop.) From the deep analogy with electrostatics, we may expect that in this case, we could calculate the magnetic field by solving a certain boundary problem for the field's potential – in our current case, the vector potential $\mathbf{A}$ defined by Eq. (28). However, despite the similarity of this formula and Eq. (1.38) for ϕ, which was noticed above, there is an additional issue we should tackle in the magnetic case – besides the obvious fact that calculating the vector potential distribution means determining three scalar functions (say, A_x, A_y, and A_z), rather than just one (ϕ).

To reveal the issue, let us plug Eq. (27) into Eq. (35):

$$\nabla \times \left(\nabla \times \mathbf{A} \right) = \mu_0 \mathbf{j} \,, \tag{5.42}$$

and then apply to the left-hand side of this equation the same identity (31). The result is

[19] Applying the Ampère law to a circular contour of radius ρ, coaxial with the solenoid, we see that the field outside (but not inside!) it has an azimuthal component B_φ, similar to that of the straight wire (see Eq. (38) above) and hence (at $N >> 1$) much weaker than the longitudinal field inside the solenoid – see Eq. (40).

$$\nabla(\nabla\cdot\mathbf{A})-\nabla^2\mathbf{A}=\mu_0\mathbf{j}\,. \tag{5.43}$$

On the other hand, as we know from electrostatics (please compare Eqs. (1.38) and (1.41)), the vector potential $\mathbf{A}(\mathbf{r})$ given by Eq. (28) has to satisfy a simpler ("vector-Poisson") equation

$$\boxed{\nabla^2\mathbf{A}=-\mu_0\mathbf{j}\,,} \tag{5.44}$$

Poisson equation for **A**

which is just a set of three usual Poisson equations for each Cartesian component of **A**.

To resolve the difference between these results, let us note that Eq. (43) is reduced to Eq. (44) if $\nabla\cdot\mathbf{A}=0$. In this context, let us discuss what discretion we have in the choice of the potential. In electrostatics, we may add, to the scalar function ϕ' that satisfies Eq. (1.33) for the given field **E**, not only an arbitrary constant but even an arbitrary function of time:

$$-\nabla\left[\phi'+f(t)\right]=-\nabla\phi'=\mathbf{E}\,. \tag{5.45}$$

Similarly, using the fact that the curl of the gradient of any scalar function equals zero,[20] we may add to any vector function $\mathbf{A}'$ that satisfies Eq. (27) for the given field **B**, not only any constant but even a gradient of an arbitrary scalar function $\chi(\mathbf{r},t)$, because

$$\nabla\times(\mathbf{A}'+\nabla\chi)=\nabla\times\mathbf{A}'+\nabla\times(\nabla\chi)=\nabla\times\mathbf{A}'=\mathbf{B}\,. \tag{5.46}$$

Such additions, which keep the fields intact, are called *gauge transformations*.[21] Let us see what such a transformation does to $\nabla\cdot\mathbf{A}'$:

$$\nabla\cdot(\mathbf{A}'+\nabla\chi)=\nabla\cdot\mathbf{A}'+\nabla^2\chi\,. \tag{5.47}$$

For any choice of such a function $\mathbf{A}'$, we can always choose the function χ in such a way that it satisfies the Poisson equation $\nabla^2\chi=-\nabla\cdot\mathbf{A}'$, and hence makes the divergence of the transformed vector potential, $\mathbf{A}\equiv\mathbf{A}'+\nabla\chi$, equal to zero everywhere,

$$\boxed{\nabla\cdot\mathbf{A}=0\,,} \tag{5.48}$$

Coulomb gauge

thus reducing Eq. (43) to Eq. (44).

To summarize, the set of distributions $\mathbf{A}'(\mathbf{r})$ that satisfy Eq. (27) for a given field $\mathbf{B}(\mathbf{r})$, is not limited to the vector potential $\mathbf{A}(\mathbf{r})$ given by Eq. (44), but is reduced to it upon the additional *Coulomb gauge condition* (48). However, as we will see in a minute, even this condition still leaves some degrees of freedom in the choice of the vector potential. To illustrate this fact, and also to get a better gut feeling of the vector potential's distribution in space, let us calculate $\mathbf{A}(\mathbf{r})$ for two very basic cases.

First, let us revisit the straight wire problem shown in Fig. 5. As Eq. (28) shows, in this case the vector potential **A** has just one component (along the axis z). Moreover, due to the problem's axial symmetry, its magnitude may only depend on the distance from the axis: $\mathbf{A}=\mathbf{n}_z A(\rho)$. Hence, the gradient of **A** is directed across the z-axis, so Eq. (48) is satisfied at all points. For our symmetry ($\partial/\partial\varphi=\partial/\partial z=0$), the Laplace operator, written in cylindrical coordinates, has just one term,[22] reducing Eq. (44) to

[20] See, e.g., MA Eq. (11.1).
[21] The use of the term "gauge" (originally meaning "a measure" or "a scale") in this context is purely historic, so the reader should not try to find too much hidden sense in it.
[22] See, e.g., MA Eq. (10.3).

$$\frac{1}{\rho}\frac{d}{d\rho}\left(\rho\frac{dA}{d\rho}\right) = -\mu_0 j(\rho)\,. \tag{5.49}$$

Multiplying both sides of this equation by ρ and integrating them over the coordinate once, we get

$$\rho\frac{dA}{d\rho} = -\mu_0\int_0^{\rho} j(\rho')\rho' d\rho' + \text{const}\,. \tag{5.50}$$

Since in the cylindrical coordinates, for our symmetry, $B = -dA/d\rho$,[23] Eq. (50) is nothing else than our old result (38) for the magnetic field.[24] However, let us continue the integration, at least for the region outside the wire, where the function $A(\rho)$ depends only on the full current I rather than on the current distribution. Dividing both parts of Eq. (50) by ρ, and integrating them over this argument again, we get

$$A(\rho) = -\frac{\mu_0 I}{2\pi}\ln\rho + \text{const}, \qquad \text{where } I = 2\pi\int_0^{R} j(\rho)\rho d\rho, \quad \text{for } \rho \geq R\,. \tag{5.51}$$

As a reminder, we had similar logarithmic behavior for the electrostatic potential outside a uniformly charged straight line. This is natural because the Poisson equations for both cases are similar.

Now let us find the vector potential for the long solenoid (Fig. 6a), with its uniform magnetic field. Since Eq. (28) tells us that the vector $\mathbf{A}$ should follow the direction of the inducing current, we may start by looking for it in the form $\mathbf{A} = \mathbf{n}_\varphi A(\rho)$. (This is especially natural if the solenoid's cross-section is circular.) With this orientation of $\mathbf{A}$, the same general expression for the curl operator in cylindrical coordinates yields $\nabla\times\mathbf{A} = \mathbf{n}_z(1/\rho)d(\rho A)/d\rho$. According to Eq. (27), this expression should be equal to $\mathbf{B}$ – in our current case to $\mathbf{n}_z B$, with a constant B – see Eq. (40). Integrating this equality, and selecting such integration constant that $A(0)$ is finite, we get

$$A(\rho) = \frac{B\rho}{2}, \qquad \text{i.e. } \mathbf{A} = \frac{B\rho}{2}\mathbf{n}_\varphi\,. \tag{5.52}$$

Plugging this result into the general expression for the Laplace operator in the cylindrical coordinates,[25] we see that the Poisson equation (44) with $\mathbf{j} = 0$ (i.e. the Laplace equation) is satisfied again – which is natural since, for this distribution, the Coulomb gauge condition (48) is satisfied: $\nabla\cdot\mathbf{A} = 0$.

However, Eq. (52) is not the unique (or even the simplest) vector potential that gives the same uniform field $\mathbf{B} = \mathbf{n}_z B$. Indeed, using the well-known expression for the curl operator in Cartesian coordinates,[26] it is straightforward to check that each of the vector functions $\mathbf{A}' = \mathbf{n}_y Bx$ and $\mathbf{A}'' = -\mathbf{n}_x By$ also has the same curl, and also satisfies the Coulomb gauge condition (48).[27] If such solutions do not look very natural because of their anisotropy in the $[x, y]$ plane, please consider the fact that they represent the uniform magnetic field regardless of its source – for example, regardless of the shape of the long solenoid's cross-section. Such choices of the vector potential may be very convenient for some

[23] See, e.g., MA Eq. (10.5) with $\partial/\partial\varphi = \partial/\partial z = 0$.

[24] Since the magnetic field at the wire's axis has to be zero (otherwise, being normal to the axis, where would it be directed?), the integration constant in Eq. (50) has to equal zero.

[25] See, e.g., MA Eq. (10.6).

[26] See, e.g., MA Eq. (8.5).

[27] The axially symmetric vector potential (52) is just a weighed sum of these two functions: $\mathbf{A} = (\mathbf{A}' + \mathbf{A}'')/2$.

problems, for example for the quantum-mechanical analysis of the 2D motion of a charged particle in the perpendicular magnetic field, giving the famous Landau energy levels.[28]

5.3. Magnetic energy, flux, and inductance

Considering the currents flowing in a system as generalized coordinates, the magnetic forces (1) between them are their unique functions, and in this sense, the energy U of their magnetic interaction may be considered the potential energy of the system. The apparent (but somewhat deceptive) way to derive an expression for this energy is to use the analogy between Eq. (1) and its electrostatic analog, Eq. (2). Indeed, Eq. (2) may be transformed into Eq. (1) with just three replacements:

(i) $\rho(\mathbf{r})\rho'(\mathbf{r}')$ should be replaced with $[\mathbf{j}(\mathbf{r})\cdot\mathbf{j}'(\mathbf{r}')]$,

(ii) ε_0 should be replaced with $1/\mu_0$, and

(iii) the sign before the double integral has to be replaced with the opposite one.

Hence we may avoid repeating the calculation made in Chapter 1, by making these replacements in Eq. (1.59), which gives the electrostatic potential energy of the system with $\rho(\mathbf{r})$ and $\rho'(\mathbf{r}')$ describing the same charge distribution, i.e. with $\rho'(\mathbf{r}) = \rho(\mathbf{r})$, to get the following expression for the magnetic potential energy in the system with, similarly, $\mathbf{j}'(\mathbf{r}) = \mathbf{j}(\mathbf{r})$:[29]

$$U_j = -\frac{\mu_0}{4\pi}\frac{1}{2}\int d^3r \int d^3r' \frac{\mathbf{j}(\mathbf{r})\cdot\mathbf{j}(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|}. \tag{5.53}$$

But this is *not* the unique answer! Indeed, Eq. (53) describes the proper potential energy of the system (in particular, giving the correct result for the current interaction forces) only in the case when the interacting currents are fixed – just as Eq. (1.59) is adequate when the interacting charges are fixed. Here comes a substantial difference between electrostatics and magnetostatics: due to the fundamental fact of electric charge conservation (already discussed in Secs. 1.1 and 4.1), keeping electric charges fixed does not require external work, while the maintenance of currents generally does. As a result, Eq. (53) describes the energy of the magnetic interaction *plus* of the system keeping the currents constant – or rather of its part depending on the system under our consideration. In this situation, using the terminology already used in Sec. 3.5 (see also a general discussion in CM Sec. 1.4.), U_j may be called the Gibbs potential energy of our magnetic system.

Now to exclude from U_j the contribution due to the interaction with the current-supporting system(s), i.e. calculate the potential energy U of our system as such, we need to know this contribution. The simplest way to do this is to use the *Faraday induction law* that describes this interaction and will be discussed at the beginning of the next chapter. This is why let me postpone the derivation until that point, and for now, ask the reader to believe me that the removal of the interaction leads to an expression similar to Eq. (53), but with the opposite sign:

$$\boxed{U = \frac{\mu_0}{4\pi}\frac{1}{2}\int d^3r \int d^3r' \frac{\mathbf{j}(\mathbf{r})\cdot\mathbf{j}(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|},} \tag{5.54}$$

Magnetic interaction energy

[28] See, e.g., QM Sec. 3.2.

[29] Just as in electrostatics, for the interaction of two *independent* current distributions $\mathbf{j}(\mathbf{r})$ and $\mathbf{j}'(\mathbf{r}')$, the factor ½ should be dropped.

I will prove this result in Sec. 6.2, but actually, this sign dichotomy should not be quite surprising to the attentive reader, in the context of a similar duality of Eqs. (3.73) and (3.81) for the electrostatic energies including and excluding the interaction with the field source.

Due to the importance of Eq. (54), let us rewrite it in several other forms, convenient for various applications. First of all, just as in electrostatics, it may be recast into a potential-based form. Indeed, with the definition (28) of the vector potential $\mathbf{A}(\mathbf{r})$, Eq. (54) becomes

$$U = \frac{1}{2}\int \mathbf{j}(\mathbf{r})\cdot\mathbf{A}(\mathbf{r})d^3r\,. \tag{5.55}$$

This formula, which is a clear magnetic analog of Eq. (1.60) of electrostatics, is very popular among field theorists, because it is very handy for their manipulations; it is also useful for some practical applications. However, for many calculations, it is more convenient to have a direct expression of the energy via the magnetic field. Again, this may be done very similarly to what had been done for electrostatics in Sec. 1.3, i.e. by plugging into Eq. (55) the current density expressed from Eq. (35) and then transforming it as[30]

$$U = \frac{1}{2}\int \mathbf{j}\cdot\mathbf{A}d^3r = \frac{1}{2\mu_0}\int \mathbf{A}\cdot(\nabla\times\mathbf{B})d^3r = \frac{1}{2\mu_0}\int \mathbf{B}\cdot(\nabla\times\mathbf{A})d^3r - \frac{1}{2\mu_0}\int \nabla\cdot(\mathbf{A}\times\mathbf{B})d^3r\,. \tag{5.56}$$

Now using the divergence theorem, the second integral may be transformed into a surface integral of $(\mathbf{A}\times\mathbf{B})_n$. According to Eqs. (27)-(28) if the current distribution $\mathbf{j}(\mathbf{r})$ is localized, this vector product drops, at large distances, faster than $1/r^2$, so if the integration volume is large enough, the surface integral is negligible. In the remaining first integral in Eq. (56), we may use Eq. (27) to rewrite $\nabla\times\mathbf{A}$ as $\mathbf{B}$. As a result, we get a very simple and fundamental formula.

$$U = \frac{1}{2\mu_0}\int B^2 d^3r\,. \tag{5.57a}$$

Just as with the electric field, this expression may be interpreted as a volume integral of the *magnetic energy density u*:

Magnetic field energy

$$U = \int u(\mathbf{r})d^3r, \quad \text{with } u(\mathbf{r}) \equiv \frac{1}{2\mu_0}\mathbf{B}^2(\mathbf{r}), \tag{5.57b}$$

clearly similar to Eq. (1.65).[31] Again, the conceptual choice between the spatial localization of magnetic energy – either at the location of electric currents only, as implied by Eqs. (54) and (55), or in all regions where the magnetic field exists, as apparent from Eq. (57b), cannot be done within the framework of magnetostatics, and only the electrodynamics gives a decisive preference for the latter choice.

For the practically important case of currents flowing in several thin wires, Eq. (54) may be first integrated over the cross-section of each wire, just as was done at the derivation of Eq. (4). As before, since the integral of the current density over the k^{th} wire's cross-section is just the current I_k in the wire, and cannot change along its length, it may be taken from the remaining integrals, giving

[30] For that, we may use MA Eq. (11.7) with $\mathbf{f} = \mathbf{A}$ and $\mathbf{g} = \mathbf{B}$, giving $\mathbf{A}\cdot(\nabla\times\mathbf{B}) = \mathbf{B}\cdot(\nabla\times\mathbf{A}) - \nabla\cdot(\mathbf{A}\times\mathbf{B})$.

[31] The transfer to the Gaussian units in Eqs. (57) may be accomplished by the usual replacement $\mu_0 \to 4\pi$, thus giving, in particular, $u = B^2/8\pi$.

$$U = \frac{\mu_0}{4\pi}\frac{1}{2}\sum_{k,k'} I_k I_{k'} \oint_{l_k}\oint_{l_{k'}} \frac{d\mathbf{r}_k \cdot d\mathbf{r}_{k'}}{|\mathbf{r}_k - \mathbf{r}_{k'}|}, \tag{5.58}$$

where l_k is the full length of the k^{th} wire loop. Note that Eq. (58) is valid if all currents I_k are independent of each other, because the double sum counts each current pair twice, compensating the coefficient ½ in front of the sum. It is useful to decompose this relation as

$$U = \frac{1}{2}\sum_{k,k'} I_k I_{k'} L_{kk'}, \tag{5.59}$$

where the coefficients $L_{kk'}$ are independent of the currents:

$$L_{kk'} \equiv \frac{\mu_0}{4\pi}\oint_{l_k}\oint_{l_{k'}} \frac{d\mathbf{r}_k \cdot d\mathbf{r}_{k'}}{|\mathbf{r}_k - \mathbf{r}_{k'}|}, \tag{5.60}$$

Mutual inductance coefficients

The coefficient $L_{kk'}$ with $k \neq k'$, is called the *mutual inductance* between current the k^{th} and k'^{th} loops, while the diagonal coefficient $L_k \equiv L_{kk}$ is called the *self-inductance* (or just *inductance*) of the k^{th} loop.[32] From the symmetry of Eq. (60) with respect to the index swap, $k \leftrightarrow k'$, it is evident that the matrix of coefficients $L_{kk'}$ is symmetric:[33]

$$L_{kk'} = L_{k'k}, \tag{5.61}$$

so for the practically most important case of two interacting currents I_1 and I_2, Eq. (59) reads

$$U = \frac{1}{2}L_1 I_1^2 + M I_1 I_2 + \frac{1}{2}L_2 I_2^2, \tag{5.62}$$

where $M \equiv L_{12} = L_{21}$ is the *mutual inductance coefficient*.

These formulas clearly show the importance of the self- and mutual inductances, so I will demonstrate their calculation for at least a few basic geometries. Before doing that, however, let me recast Eq. (58) into one more form that may facilitate such calculations. Namely, let us notice that for the magnetic field induced by current I_k in a thin wire, Eq. (28) is reduced to

$$\mathbf{A}_k(\mathbf{r}) = \frac{\mu_0}{4\pi} I_k \int_{l'} \frac{d\mathbf{r}_k}{|\mathbf{r} - \mathbf{r}_k|}, \tag{5.63}$$

so Eq. (58) may be rewritten as

$$U = \frac{1}{2}\sum_{k,k'} I_k \oint_{l_k} \mathbf{A}_{k'}(\mathbf{r}_k)\cdot d\mathbf{r}_{k'}. \tag{5.64}$$

But according to the same Stokes theorem that was used earlier in this chapter to derive the Ampère law, and Eq. (27), the integral in Eq. (64) is nothing else than the *magnetic field's flux* Φ (more frequently called just the *magnetic flux*) through a surface S limited by the contour l :

[32] As evident from Eq. (60), these coefficients depend only on the geometry of the system. Moreover, in the Gaussian units, in which Eq. (60) is valid without the factor $\mu_0/4\pi$, the inductance coefficients have the dimension of length (centimeters). The SI unit of inductance is called the *henry*, abbreviated H – after Joseph Henry, who in particular discovered the effect of electromagnetic induction (see Sec. 6.1) independently of Michael Faraday.

[33] Note that the matrix of the mutual inductances $L_{jj'}$ is very similar to the matrix of *reciprocal* capacitance coefficients $p_{kk'}$ – for example, compare Eq. (62) with Eq. (2.21).

Magnetic flux

$$\oint_l \mathbf{A}(\mathbf{r})\cdot d\mathbf{r} = \int_S (\nabla \times A)_n d^2 r = \int_S B_n d^2 r \equiv \Phi \tag{5.65}$$

– in the particular case of Eq. (64), the flux $\Phi_{kk'}$ of the field induced by the k'^{th} current through the loop of the k^{th} current.[34] As a result, Eq. (64) may be rewritten as

$$U = \frac{1}{2}\sum_{k,k'} I_k \Phi_{kk'}. \tag{5.66}$$

Comparing this expression with Eq. (59), we see that

$$\Phi_{kk'} \equiv \int_{S_k} (\mathbf{B}_{k'})_n d^2 r = L_{kk'} I_{k'}, \tag{5.67}$$

This expression not only gives us one more means for calculating the coefficients $L_{kk'}$, but also shows their physical sense: the mutual inductance characterizes what part of the magnetic flux (colloquially, "what fraction of field lines") induced by the current $I_{k'}$ pierces the k^{th} loop's area S_k – see Fig. 7.

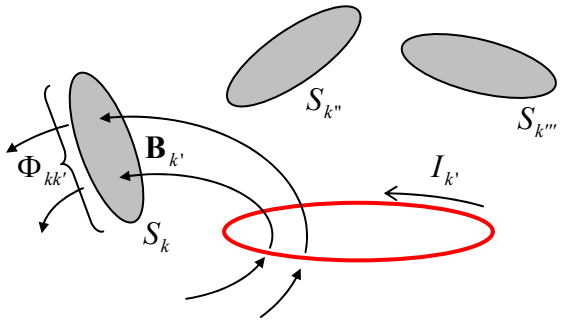


Fig. 5.7. The physical sense of the mutual inductance coefficient $L_{kk'} \equiv \Phi_{kk'}/I_{k'}$ – schematically.

Due to the linear superposition principle, the total flux piercing the k^{th} loop may be represented as

Magnetic flux from currents

$$\Phi_k \equiv \sum_{k'} \Phi_{kk'} = \sum_{k'} L_{kk'} I_{k'} \tag{5.68}$$

For example, for the system of two currents, this expression is reduced to a clear analog of Eqs. (2.19):

$$\begin{aligned} \Phi_1 &= L_1 I_1 + M I_2, \\ \Phi_2 &= M I_1 + L_2 I_2 . \end{aligned} \tag{5.69}$$

For the even simpler case of a single current,

Φ of a single current

$$\Phi = LI, \tag{5.70}$$

so the magnetic energy of the current may be represented in several equivalent forms:

[34] The SI unit of magnetic flux is called *weber*, abbreviated Wb – after Wilhelm Edward Weber (1804-1891), who in particular co-invented (with Carl Gauss) the electromagnetic telegraph. More importantly for this course, in 1856 he was the first (together with Rudolf Kohlrausch) to notice that the value of (in modern terms) $1/(\varepsilon_0\mu_0)^{1/2}$, derived from electrostatic and magnetostatic measurements, coincides with the independently measured speed of light c. This observation gave an important motivation for Maxwell's theory.

$$U = \frac{L}{2} I^2 = \frac{1}{2} I\Phi = \frac{1}{2L}\Phi^2 . \tag{5.71}$$

U of a single current

These relations, similar to Eqs. (2.14)-(2.15) of electrostatics, show that the self-inductance L of a current loop may be considered a measure of the system's magnetic energy. However, as we will see in Sec. 6.1, this measure is adequate only if the flux Φ, rather than the current I, is fixed.

Now we are well equipped for the calculation of inductance coefficients for particular systems, having three options. The first one is to use Eq. (60) directly.[35] The second one is to calculate the magnetic field energy from Eq. (57) as the function of all currents I_k in the system, and then use Eq. (59) to find all coefficients $L_{kk'}$. For example, for a system with just one current, Eq. (71) yields

$$L = \frac{U}{I^2/2} . \tag{5.72}$$

Finally, if the system consists of thin wires, so the loop areas S_k and hence the fluxes $\Phi_{kk'}$ are well defined, we may calculate them from Eq. (65), and then use Eq. (67) to find the inductances.

Usually, the third option is simpler, but the first two may be very useful even for thin-wire systems, especially if the notion of magnetic flux in them is not quite apparent. As an important example, let us find the self-inductance of a long solenoid – see Fig. 6a again. We have already calculated the magnetic field inside it – see Eq. (40) – so, due to the field uniformity, the magnetic flux piercing each turn of the wire is just

$$\Phi_1 = BA = \mu_0 nIA , \tag{5.73}$$

where A is the area of the solenoid's cross-section – for example πR^2 for a round solenoid, though Eq. (40), and hence Eq. (73) are valid for cross-sections of any shape. Comparing Eqs. (73) with Eq. (70), one might wrongly conclude that $L = \Phi_1/I = \mu_0 nA$ (WRONG!), i.e. that the solenoid's inductance is independent of its length. Actually, the magnetic flux Φ_1 pierces *each* wire turn, so the total flux through the *whole* current loop, consisting of N turns, is

$$\Phi = N\Phi_1 = \mu_0 n^2 lAI , \tag{5.74}$$

and the correct expression for the long solenoid's self-inductance is

$$L = \frac{\Phi}{I} = \mu_0 n^2 lA \equiv \frac{\mu_0 N^2 A}{l} , \tag{5.75}$$

L of a solenoid

i.e. at fixed A and l, the inductance scales as N^2, not as N. Since this reasoning may seem not quite evident, it is prudent to verify this result by using Eq. (72), with the full magnetic energy inside the solenoid (neglecting minor fringe field contributions), given by Eq. (57) with $\mathbf{B}$ = const within the internal volume $V = lA$, and zero outside of it:

$$U = \frac{1}{2\mu_0} B^2 Al = \frac{1}{2\mu_0} (\mu_0 nI)^2 Al \equiv \mu_0 n^2 lA \frac{I^2}{2} . \tag{5.76}$$

Plugging this relation into Eq. (72) immediately confirms the result (75).

[35] Numerous applications of that *Neumann formula* (derived in 1845 by F. Neumann) to electrical engineering problems may be found, for example, in the classical text by F. Grover, *Inductance Calculations*, Dover, 1946.

This energy-based approach becomes virtually inevitable for continuously distributed currents. As an example, let us calculate the self-inductance L of a long coaxial cable with the cross-section shown in Fig. 8, [36] and the full current in the outer conductor equal and opposite to that (I) in the inner conductor.

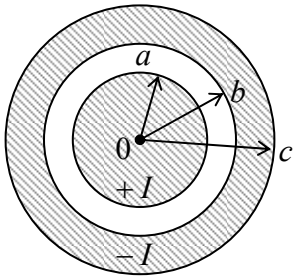


Fig. 5.8. The cross-section of a coaxial cable.

Let us assume that the current is uniformly distributed over the cross-sections of both conductors. (As we know from the previous chapter, this is indeed the case if both the internal and external conductors are made of a uniform resistive material.) First, we should calculate the radial distribution of the magnetic field – which has only one, azimuthal component because of the axial symmetry of the problem. This distribution may be immediately found by applying the Ampère law (37) to circular contours of radii ρ within four different ranges:

$$2\pi\rho B = \mu_0 I\Big|_{\text{piercing the contour}} = \mu_0 I \times \begin{cases} \rho^2/a^2, & \text{for } \rho < a, \\ 1, & \text{for } a < \rho < b, \\ \left(c^2-\rho^2\right)/\left(c^2-b^2\right), & \text{for } b < \rho < c, \\ 0, & \text{for } c < \rho. \end{cases} \tag{5.77}$$

Now, an easy integration yields the magnetic energy per unit length of the cable:

$$\begin{aligned} \frac{U}{l} &= \frac{1}{2\mu_0}\int B^2 d^2 r = \frac{\pi}{\mu_0}\int_0^\infty B^2 \rho d\rho = \frac{\mu_0 I^2}{4\pi}\left[\int_0^a \left(\frac{\rho}{a^2}\right)^2 \rho d\rho + \int_a^b \left(\frac{1}{\rho}\right)^2 \rho d\rho + \int_b^c \left(\frac{c^2-\rho^2}{\rho(c^2-b^2)}\right)^2 \rho d\rho\right] \\ &= \frac{\mu_0}{2\pi}\left[\ln\frac{b}{a} + \frac{c^2}{c^2-b^2}\left(\frac{c^2}{c^2-b^2}\ln\frac{c}{b} - \frac{1}{2}\right)\right]\frac{I^2}{2}. \end{aligned} \tag{5.78}$$

From here, and Eq. (72), we get the final answer:

$$\frac{L}{l} = \frac{\mu_0}{2\pi}\left[\ln\frac{b}{a} + \frac{c^2}{c^2-b^2}\left(\frac{c^2}{c^2-b^2}\ln\frac{c}{b} - \frac{1}{2}\right)\right]. \tag{5.79}$$

Note that for the particular case of a thin outer conductor, $c - b << b$, this expression reduces to

$$\frac{L}{l} \approx \frac{\mu_0}{2\pi}\left(\ln\frac{b}{a} + \frac{1}{4}\right), \tag{5.80}$$

where the first term in the parentheses is due to the contribution of the magnetic field energy in the free space between the conductors. This distinction is important for some applications because in

[36] As a reminder, the mutual capacitance C between the conductors of such a system was calculated in Sec. 2.3.

superconductor cables, as well as the normal-metal cables at high frequencies (to be discussed in the next chapter), the field does not penetrate the conductor's bulk, so Eq. (80) is valid without the last term ¼ in the parentheses – for any $b < c$.

As the last example, let us calculate the *mutual* inductance between a long straight wire and a round wire loop adjacent to it (Fig. 9), neglecting the thickness of both wires.

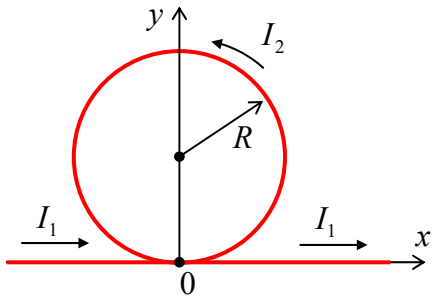


Fig. 5.9. An example of the mutual inductance calculation.

Here there is no problem with using the last approach discussed above, based on the direct calculation of the magnetic flux. Indeed, as was discussed in Sec. 1, the field $\mathbf{B}_1$ induced by the current I_1 at any point of the round loop is normal to its plane – e.g., to the plane of the drawing of Fig. 9. In the Cartesian coordinates shown in that figure, Eq. (20) reads $B_1 = \mu_0 I_1/2\pi y$, giving the following magnetic flux through the loop:

$$\Phi_{21} = \frac{\mu_0 I_1}{2\pi}\int_{-R}^{+R} dx \int_{R-\left(R^2-x^2\right)^{1/2}}^{R+\left(R^2-x^2\right)^{1/2}} \frac{dy}{y} = \frac{\mu_0 I_1}{\pi}\int_0^R \ln\frac{R+\left(R^2-x^2\right)^{1/2}}{R-\left(R^2-x^2\right)^{1/2}}\,dx \equiv \frac{\mu_0 I_1 R}{\pi}\int_0^1 \ln\frac{1+\left(1-\xi^2\right)^{1/2}}{1+\left(1-\xi^2\right)^{1/2}}\,d\xi. \quad (5.81)$$

This is a table integral equal to π,[37] so $\Phi_{21} = \mu_0 I_1 R$, and the final answer for the mutual inductance $M \equiv L_{12} = L_{21} = \Phi_{21}/I_1$ is finite (and very simple):

$$M = \mu_0 R\,, \quad (5.82)$$

despite the magnetic field's divergence at the lowest point of the loop ($y = 0$).

Note that in contrast with the finite *mutual* inductance of this system, the *self*-inductances of both its wires are formally infinite in the thin-wire limit – see, e.g., Eq. (80), which, in the limit $b/a >> 1$, describes a thin straight wire. However, since this divergence is very weak (logarithmic), it is quenched by any deviation from this perfectly axial geometry. For example, a fair estimate of the inductance of a wire of a large but finite length $l >> a$ may be obtained from Eq. (80) by the replacement of b with l:

$$L \approx \frac{\mu_0 l}{2\pi}\ln\frac{l}{a}\,. \quad (5.83)$$

(Note, however, that the exact result depends on where from/to the current flows beyond that segment.) It turns out that a similar approximate result, with l replaced with $2\pi R$ in the front factor, and with R under the logarithm, is valid for the self-inductance of a round loop with $a << R$. (A proof of this fact is a very useful exercise, highly recommended to the reader.)

[37] See, e.g., MA Eq. (6.13), with $a = 1$.

5.4. Magnetic dipole moment, and magnetic dipole media

The most natural way of the magnetic media description parallels that for dielectrics in Chapter 3, and is based on the properties of *magnetic dipoles* – the notion close (but not identical!) to that of the electric dipoles discussed in Sec. 3.1. To introduce this notion quantitatively, let us consider, just as in Sec. 3.1, a spatially-localized system with a current distribution $\mathbf{j}(\mathbf{r}')$, whose magnetic field is measured at relatively large distances $r >> r'$ (Fig. 10).

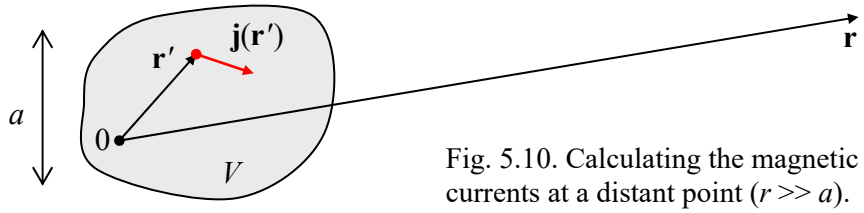


Fig. 5.10. Calculating the magnetic field of localized currents at a distant point ($r >> a$).

Applying the truncated Taylor expansion (3.5) of the fraction $1/|\mathbf{r} - \mathbf{r}'|$ to the vector potential given by Eq. (28), we get

$$\mathbf{A}(\mathbf{r}) \approx \frac{\mu_0}{4\pi}\left[\frac{1}{r}\int_V \mathbf{j}(\mathbf{r}')d^3r' + \frac{1}{r^3}\int_V (\mathbf{r}\cdot\mathbf{r}')\mathbf{j}(\mathbf{r}')d^3r'\right]. \tag{5.84}$$

Now, due to the vector character of this potential, we have to depart somewhat from the approach of Sec. 3.1 and use the following vector algebra identity:[38]

$$\int_V \left[f(\mathbf{j}\cdot\nabla g) + g(\mathbf{j}\cdot\nabla f)\right]d^3r = 0, \tag{5.85}$$

that is valid for any pair of smooth (differentiable) scalar functions $f(\mathbf{r})$ and $g(\mathbf{r})$, and any vector function $\mathbf{j}(\mathbf{r})$ that, as the dc current density, satisfies the continuity condition $\nabla\cdot\mathbf{j} = 0$ and whose normal component vanishes on the surface of the volume V. First, let us use Eq. (85) with f equal to 1, and g equal to any Cartesian component of the radius-vector $\mathbf{r}$: $g = r_l$ ($l = 1, 2, 3$). Then it yields

$$\int_V (\mathbf{j}\cdot\mathbf{n}_l)d^3r = \int_V j_l d^3r = 0, \tag{5.86}$$

so for the vector as the whole

$$\int_V \mathbf{j}(\mathbf{r})d^3r = 0, \tag{5.87}$$

showing that the first term on the right-hand side of Eq. (84) equals zero. Next, let us use Eq. (85) again, but now with $f = r_l$, $g = r_{l'}$ (where $l, l' = 1, 2, 3$); then it yields

$$\int_V \left(r_l j_{l'} + r_{l'} j_l\right) d^3r = 0, \tag{5.88}$$

so the l^{th} Cartesian component of the second integral in Eq. (84) may be transformed as

[38] See, e.g., MA Eq. (12.3) with the additional condition $j_n|_S = 0$, pertinent to space-restricted currents.

$$\int_V (\mathbf{r}\cdot\mathbf{r}')j_l d^3r' = \int_V \sum_{l'=1}^{3} r_{l'} r'_{l'} j_l d^3r' = \frac{1}{2}\sum_{l'=1}^{3} r_{l'}\int_V (r'_{l'} j_l + r'_{l'} j_l) d^3r'$$
$$= \frac{1}{2}\sum_{l'=1}^{3} r_{l'}\int_V (r'_{l'} j_l - r'_{l'} j_{l'}) d^3r' = -\frac{1}{2}\left[\mathbf{r}\times\int_V (\mathbf{r}'\times\mathbf{j}) d^3r'\right]_l . \quad (5.89)$$

As a result, Eq. (84) may be rewritten as

$$\mathbf{A}(\mathbf{r}) = \frac{\mu_0}{4\pi}\frac{\mathbf{m}\times\mathbf{r}}{r^3}, \quad (5.90)$$

Magnetic dipole and its potential

where the vector **m**, defined as[39]

$$\mathbf{m} \equiv \frac{1}{2}\int_V \mathbf{r}\times\mathbf{j}(\mathbf{r})\, d^3r , \quad (5.91)$$

is called the *magnetic dipole moment* of a field source – which itself, within the long-range approximation (90), is called the *magnetic dipole*.

Note a close analogy between the **m** defined by Eq. (91), and the orbital[40] angular momentum of a non-relativistic particle with mass m_k:

$$\mathbf{L}_k \equiv \mathbf{r}_k \times \mathbf{p}_k = \mathbf{r}_k \times m_k \mathbf{v}_k , \quad (5.92)$$

where $\mathbf{p}_k = m_k\mathbf{v}_k$ is its linear momentum. Indeed, for a continuum of such particles with equal electric charges q, distributed with spatial density n, we have $\mathbf{j} = qn\mathbf{v}$, and Eq. (91) yields

$$\mathbf{m} = \int_V \frac{1}{2}\mathbf{r}\times\mathbf{j} d^3r = \int_V \frac{nq}{2}\mathbf{r}\times\mathbf{v}\, d^3r , \quad (5.93)$$

while the total angular momentum of such a system of particles of equal masses m_0, is

$$\mathbf{L} = \int_V nm_0 \mathbf{r}\times\mathbf{v} d^3r ,$$

so we get a very straightforward relation

$$\mathbf{m} = \frac{q}{2m_0}\mathbf{L} . \quad (5.95)$$

m vs. **L**

For the orbital motion, this classical relation survives in quantum mechanics for linear operators, and hence for eigenvalues of the observables. Since the orbital angular momentum is quantized in the units of the Planck constant $\hbar$, the orbital magnetic moment of an electron is always a multiple of the so-called *Bohr magneton*

$$\mu_B \equiv \frac{e\hbar}{2m_e} , \quad (5.96)$$

Bohr magneton

where m_e is the free electron mass.[41] However, for particles with spin, such a universal relation between the vectors **m** and **L** is no longer valid. For example, the electron's spin $s = ½$ gives a contribution of $\hbar/2$ to its mechanical angular momentum, but a contribution very close to μ_B to its magnetic moment.

[39] In the Gaussian units, the definition (91) is kept valid "as is", so Eq. (90) is stripped of the factor $\mu_0/4\pi$.
[40] This adjective is used, especially in quantum mechanics, to distinguish the motion of a particle as a whole (not necessarily along a closed orbit!) from its intrinsic angular momentum, the spin – see, e.g., QM Chapters 3-6.
[41] In the SI units, $m_e \approx 0.91\times10^{-30}$ kg, so $\mu_B \approx 0.93\times10^{-23}$ J/T.

The next important example of a magnetic dipole is a *planar* thin-wire loop, limiting area A (of arbitrary shape), and carrying current I, for which $\mathbf{m}$ has a surprisingly simple form,

$$\mathbf{m} = I\mathbf{A}, \tag{5.97}$$

where the modulus of the vector $\mathbf{A}$ equals the loop's area A, and its direction is normal to the loop's plane. This formula may be readily proved by noticing that if we select the coordinate frame origin on the plane of the loop (Fig. 11), then the elementary component of the magnitude of the integral (91),

$$dm = \frac{1}{2}\left|\oint_C \mathbf{r}\times I\,d\mathbf{r}\right| \equiv I\oint_C\left|\frac{1}{2}\mathbf{r}\times d\mathbf{r}\right| = I\oint_C \frac{1}{2}r^2 d\varphi, \tag{5.98}$$

is just the elementary area $dA = (1/2)rd(r\varphi) = r^2 d\varphi/2$ – the equality already used in CM Eq. (3.40).

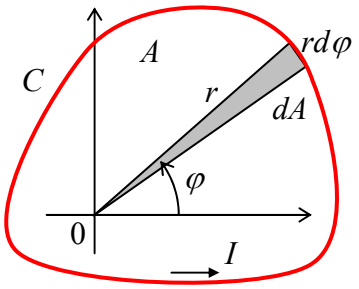


Fig. 5.11. Calculating the magnetic dipole moment of a planar current loop.

The comparison of Eqs. (96) and (97) allows a useful estimate of atomic currents, by finding what current I should flow in a circular loop of the atomic size scale (the Bohr radius) $r_B \sim 0.5\times10^{-10}$ m, i.e. of area $A \sim 10^{-20}$ m^2, to produce a magnetic moment of the order of μ_B.[42] The result is surprisingly macroscopic: $I \sim 1$ mA – quite comparable to the current driving the sound in your phone's earbuds. Though due to the quantum-mechanical spread of electron wavefunctions, this estimate should not be taken too literally, it is very useful for getting a gut feeling of how significant the atomic magnetism is, and hence why ferromagnets may provide such strong magnetic fields.

After these illustrations, let us return to the discussion of the general Eq. (90). Plugging it into (also general) Eq. (27), we may calculate the magnetic field of a magnetic dipole: [43]

---

[42] Another way to arrive at the same estimate is to take $I \sim ef = e\omega/2\pi$ with $\omega \sim 10^{16}$ s^{-1} being the typical frequency of radiation due to atomic interlevel quantum transitions.

[43] Similarly to the situation with the electric dipoles (see Eq. (3.24) and its discussion), it may be shown that the magnetic field of any closed current loop (or any system of such loops) satisfies the following equality:

$$\int_V \mathbf{B}(\mathbf{r})d^3r = (2/3)\mu_0\mathbf{m},$$

where the integral is over any sphere confining all the currents. On the other hand, as we know from Sec. 3.1, for a field with the structure (99), derived from the long-range approximation (90), such an integral vanishes. As a result, to get a coarse-grain description of the magnetic field of a small system located at $r = 0$, that would give the correct average value of the magnetic field, Eq. (99) should be modified as follows:

$$\mathbf{B}_{cg}(\mathbf{r}) = \frac{\mu_0}{4\pi}\left(\frac{3\mathbf{r}(\mathbf{r}\cdot\mathbf{m}) - \mathbf{m}r^2}{r^5} + \frac{8\pi}{3}\mathbf{m}\delta(\mathbf{r})\right),$$

in a conceptual (though not quantitative) similarity to Eq. (3.25).

$$\mathbf{B}(\mathbf{r}) = \frac{\mu_0}{4\pi}\frac{3\mathbf{r}(\mathbf{r}\cdot\mathbf{m}) - \mathbf{m}r^2}{r^5}. \tag{5.99}$$

Magnetic dipole's field

The structure of this formula *exactly* replicates that of Eq. (3.13) for the electric dipole field – including the sign). Because of this similarity, the energy of a dipole of a fixed magnitude m in an external field, and hence the torque and the force exerted on it by a fixed external field, are given by expressions fully similar to those for an electric dipole – see Eqs. (3.15)-(3.19):[44]

$$U = -\mathbf{m}\cdot\mathbf{B}_{\text{ext}}, \tag{5.100}$$

Magnetic dipole in external field

and as a result,

$$\boldsymbol{\tau} = \mathbf{m}\times\mathbf{B}_{\text{ext}}, \tag{5.101}$$

$$\mathbf{F} = \nabla(\mathbf{m}\cdot\mathbf{B}_{\text{ext}}). \tag{5.102}$$

Now let us consider a system of many magnetic dipoles (e.g., atoms or molecules), distributed in space with an atomic-scale-averaged density n. Then we can use Eq. (90) generalized in an evident way for an arbitrary position $\mathbf{r}'$ of the dipole, and the linear superposition principle, to calculate the macroscopic vector potential $\mathbf{A}$:

$$\mathbf{A}(\mathbf{r}) = \frac{\mu_0}{4\pi}\int\frac{\mathbf{M}(\mathbf{r}')\times(\mathbf{r}-\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|^3}d^3r', \tag{5.103}$$

where $\mathbf{M} \equiv n\mathbf{m}$ is the *magnetization*: the average magnetic moment per unit volume. Transforming this integral absolutely similarly to how Eq. (3.27) had been transformed into Eq. (3.29), we get:

$$\mathbf{A}(\mathbf{r}) = \frac{\mu_0}{4\pi}\int\frac{\nabla'\times\mathbf{M}(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|}d^3r'. \tag{5.104}$$

Comparing this result with Eq. (28), we see that $\nabla\times\mathbf{M}$ is equivalent, in its magnetic effect, to the density $\mathbf{j}_{\text{ef}}$ of a certain effective "magnetization current". Just as the electric-polarization charge ρ_{ef} discussed in Sec. 3.2 (see Fig. 3.4), the vector $\mathbf{j}_{\text{ef}} = \nabla\times\mathbf{M}$ may be interpreted as the uncompensated part of the loop currents representing single magnetic dipoles $\mathbf{m}$ – see Fig. 12. Note, however, that since the atomic magnetic dipoles may be due to particles' spins, rather than the actual electric currents due to the orbital motion, the magnetization current's nature is not as direct as that of the polarization charge.

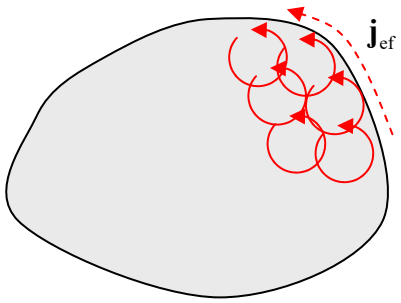


Fig. 5.12. A cartoon illustrating the physical nature of the effective magnetization current $\mathbf{j}_{\text{ef}} = \nabla\times\mathbf{M}$.

[44] Note that the fixation of m and $\mathbf{B}_{\text{ext}}$ effectively means that the currents producing them are fixed – please have one more look at Eqs. (35) and (97). As a result, Eq. (100) is a particular case of Eq. (53) rather than (54) – hence the minus sign.

Now, using Eq. (28) to add the possible contribution from the *stand-alone currents* **j** not included in the currents of microscopic magnetic dipoles, we get the general expression for the vector potential of the macroscopic field:

$$\mathbf{A}(\mathbf{r}) = \frac{\mu_0}{4\pi}\int \frac{\left[\mathbf{j}(\mathbf{r}') + \nabla' \times \mathbf{M}(\mathbf{r}')\right]}{|\mathbf{r}-\mathbf{r}'|} d^3 r' . \tag{5.105}$$

Repeating the calculations that have led us from Eq. (28) to the Maxwell equation (35), with the account of the magnetization current term, for the macroscopic magnetic field **B** we get

$$\nabla \times \mathbf{B} = \mu_0 (\mathbf{j} + \nabla \times \mathbf{M}). \tag{5.106}$$

Following the same reasoning as in Sec. 3.2, we may recast this equation as

$$\nabla \times \mathbf{H} = \mathbf{j}, \tag{5.107}$$

where the field defined as

Magnetic field **H**

$$\boxed{\mathbf{H} \equiv \frac{\mathbf{B}}{\mu_0} - \mathbf{M},} \tag{5.108}$$

for historic reasons (and very unfortunately) is also called the *magnetic field*.[45] This is why it is crucial to remember that the physical sense of field **H** is very much different from field **B**. To understand this difference better, let us use Eq. (107) to bring Eqs. (3.32), (3.36), (29), and (107) together, writing them as the following system of *macroscopic Maxwell equations* (again, so far for the stationary case $\partial/\partial t = 0$):[46]

Stationary macroscopic Maxwell equations

$$\boxed{\begin{aligned} \nabla \times \mathbf{E} = 0, \quad & \nabla \times \mathbf{H} = \mathbf{j}, \\ \nabla \cdot \mathbf{D} = \rho, \quad & \nabla \cdot \mathbf{B} = 0. \end{aligned}} \tag{5.109}$$

These equations clearly show that the roles of the vector fields **D** and **H** are very similar: they both may be called "would-be fields" – meaning the fields that *would be* induced by the stand-alone charges ρ and currents **j**, if the medium had not modified them by its dielectric and magnetic polarization.

Despite this similarity, let me note an important difference of signs in the relation (3.33) between **E, D**, and **P**, on one hand, and the relation (108) between **B**, **H**, and **M**, on the other hand. This is *not* just a matter of definition. Indeed, due to the similarity of Eqs. (3.15) and (100), including similar signs, the electric and magnetic fields both try to orient the corresponding dipole moments along the field. Hence, in the media that allow such an orientation (and as we will see momentarily, for magnetic media it is not always the case), the induced polarizations **P** and **M** are directed along, respectively, the vectors **E** and **B** of the genuine (though macroscopic, i.e. atomic-scale-averaged) fields. According to Eq. (3.33), if the would-be field **D** is fixed – say, by a fixed stand-alone charge distribution $\rho(\mathbf{r})$ – such a polarization *reduces* the electric field $\mathbf{E} = (\mathbf{D} - \mathbf{P})/\varepsilon_0$. On the other hand, Eq. (108) shows that in a magnetic media with a fixed would-be field **H**, the magnetic polarization making **M** parallel to **B**,

[45] This confusion is exacerbated by the fact that in Gaussian units, Eq. (108) has the form $\mathbf{H} = \mathbf{B} - 4\pi\mathbf{M}$, and hence the fields **B** and **H** have the same dimensionality (and are formally equal in free space!) – though the unit of **H** has a different name (*oersted*, abbreviated as Oe). Mercifully, in the SI units, the dimensionality of **B** and **H** is different, with the unit of **H** called the *ampere per meter*.

[46] Let me remind the reader once again that in contrast with the system (36) of the Maxwell equations for the genuine (microscopic) fields, the right-hand sides of Eqs. (109) represent only the stand-alone charges and currents, not included in the microscopic electric and magnetic dipoles.

enhances the magnetic field $\mathbf{B} = \mu_0(\mathbf{H} + \mathbf{M})$. This difference may be traced back to the sign difference in the basic relations (1) and (2), i.e. to the fundamental fact that the electric charges of the same sign repulse, while the currents of the same direction attract each other.

5.5. Magnetic materials

In order to form a complete system, sufficient for the calculation of all fields from given $\rho(\mathbf{r})$ and $\mathbf{j}(\mathbf{r})$, the macroscopic Maxwell equations (109) have to be complemented with the constitutive relations describing the medium: $\mathbf{D} \leftrightarrow \mathbf{E}$, $\mathbf{j} \leftrightarrow \mathbf{E}$, and $\mathbf{B} \leftrightarrow \mathbf{H}$. The first two of them were discussed, in brief, in the last two chapters; let us proceed to the last one.

A major difference between the dielectric and magnetic constitutive relations $\mathbf{D}(\mathbf{E})$ and $\mathbf{B}(\mathbf{H})$ is that while a dielectric medium always *reduces* the external field, magnetic media may *either reduce or enhance* it. To quantify this fact, let us consider the most common case – *linear magnetic materials* in that $\mathbf{M}$ (and hence $\mathbf{H}$) is proportional to $\mathbf{B}$. For isotropic materials, this proportionality is characterized by a scalar – either the *magnetic permeability* μ defined by the following relation:

$$\mathbf{B} \equiv \mu\mathbf{H}, \tag{5.110}$$

Magnetic permeability

or the *magnetic susceptibility*[47] defined as

$$\mathbf{M} = \chi_m \mathbf{H}. \tag{5.111}$$

Magnetic susceptibility

Plugging these relations into Eq. (108), we see that these two parameters are not independent, but are related as

$$\mu = (1 + \chi_m)\mu_0. \tag{5.112}$$

χ_m vs. μ

Note that despite the superficial similarity between Eqs. (110)-(112) and the corresponding relations (3.43)-(3.47) for linear dielectrics:

$$\mathbf{D} = \varepsilon\mathbf{E}, \qquad \mathbf{P} = \chi_e\varepsilon_0\mathbf{E}, \qquad \varepsilon = (1 + \chi_e)\varepsilon_0, \tag{5.113}$$

there is an important conceptual difference between them. Namely, while the vector $\mathbf{E}$ on the right-hand sides of Eqs. (113) is the actual (though macroscopic) electric field, the vector $\mathbf{H}$ on the right-hand side of Eqs. (110)-(111) represents a "would-be" magnetic field, in most aspects similar to $\mathbf{D}$ rather than $\mathbf{E}$ – see, for example, Eqs. (109). This historic difference in the traditional form of the constitutive relations for the electric and magnetic fields is not without its physical reasons. Most experiments with electric and magnetic materials are performed by placing their samples into nearly uniform electric and magnetic fields, and the simplest systems for their implementation are, respectively, plane capacitors (Fig. 2.3) and long solenoids (Fig. 6). The field in the former system may be most conveniently

[47] According to Eqs. (110) and (112), i.e. in the SI units, χ_m is dimensionless, while μ has the same dimensionality as μ_0. In the Gaussian units, μ is dimensionless: $(\mu)_{\text{Gaussian}} = (\mu)_{\text{SI}}/\mu_0$, and χ_m is also introduced differently, as $\mu = 1 + 4\pi\chi_m$, Hence, just as for the electric susceptibilities, these dimensionless coefficients are different in the two systems: $(\chi_m)_{\text{SI}} = 4\pi(\chi_m)_{\text{Gaussian}}$. Note also that χ_m is formally called the *volumic* magnetic susceptibility, to distinguish it from the *atomic* (or "molecular") susceptibility χ defined by a similar relation, $\langle\mathbf{m}\rangle \equiv \chi\mathbf{H}$, where $\mathbf{m}$ is the induced magnetic moment of a single dipole – e.g., an atom. (χ is an analog of the electric atomic polarizability α – see Eq. (3.48) and its discussion.) In a dilute medium, i.e. in the absence of substantial dipole-dipole interactions, $\chi_m = n\chi$, where n is the dipole density.

controlled by fixing the voltage V between its plates, which is proportional to the electric field **E**. On the other hand, the field provided by the solenoid may be fixed by the current I in it, and according to Eq. (107), the field proportional to this stand-alone current is **H**, rather than **B**.[48]

Table 1 lists the approximate magnetic susceptibility values for several materials. It shows that in contrast to linear dielectrics whose susceptibility χ_e is always positive, i.e. the dielectric constant $\kappa = \chi_e + 1$ is always larger than 1 (see Table 3.1), linear magnetic materials may be either *paramagnets* (with $\chi_m > 0$, i. e. $\mu > \mu_0$) or *diamagnets* (with $\chi_m < 0$, i.e. $\mu < \mu_0$).

Table 5.1. Susceptibility $(\chi_m)_{SI}$ of a few representative and/or important magnetic materials[(a)]

"Mu-metal" (75% Ni + 15% Fe + a few %% of Cu and Mo)	~20,000[(b)]
Permalloy (80% Ni + 20% Fe)	~8,000[(b)]
"Electrical" (or "transformer") steel (Fe + a few %% of Si)	~4,000[(b)]
Nickel	~100
Aluminum	$+2\times10^{-5}$
Oxygen (at ambient conditions)	$+0.2\times10^{-5}$
Water	-0.9×10^{-5}
Diamond	-2×10^{-5}
Copper	-7×10^{-5}
Bismuth (the strongest non-superconducting diamagnet)	-17×10^{-5}

[(a)] The table does not include bulk superconductors, which may be described, in a so-called *coarse-scale approximation*, as perfect diamagnets (with **B** = 0, i.e. formally with $\chi_m = -1$ and $\mu = 0$), though the actual physics of this phenomenon is different – see Sec. 6.3 below.

[(b)] The exact values of $\chi_m >> 1$ for soft ferromagnetic materials (see, e.g., the upper three rows of the table) depend not only on their composition but also on their thermal processing ("annealing"). Moreover, due to unintentional vibrations, the extremely high values of χ_m of such materials may decay with time, though they may be restored to the original values by new annealing. The reason for such behavior is discussed in the text below.

The reason for this difference is that in dielectrics, two different polarization mechanisms (schematically illustrated by Fig. 3.7) lead to the same sign of the average polarization – see the discussion in Sec. 3.3. One of these mechanisms, illustrated by Fig. 3.7b, i.e. the ordering of spontaneous dipoles by the applied field, is also possible for magnetization – for the atoms and molecules with spontaneous internal magnetic dipoles of magnitude $m_0 \sim \mu_B$, due to their net spins. Again, in the absence of an external magnetic field the spins, and hence the dipole moments $\mathbf{m}_0$ may be disordered, but according to Eq. (100), the external magnetic field tends to align the dipoles along its direction. As a result, the average direction of the spontaneous elementary moments $\mathbf{m}_0$, and hence the direction of the arising magnetization **M**, is the same as that of the microscopic field **B** at the points of the dipole location (i.e., for a diluted media, of $\mathbf{H} \approx \mathbf{B}/\mu_0$), resulting in a positive susceptibility χ_m, i.e. in the paramagnetism, such as that of oxygen and aluminum – see Table 1.

[48] This fact also explains the misleading term "magnetic field" for **H**.

However, in contrast to the electric polarization of atoms/molecules with no spontaneous electric dipoles, which gives the same sign of $\chi_e \equiv \kappa - 1$ (see Fig. 3.7a and its discussion), the magnetic materials with no spontaneous atomic magnetic dipole moments have $\chi_m < 0$ – the effect called the *orbital* (or "Larmor"[49]) *diamagnetism*. As the simplest model of this effect, let us consider the orbital motion of an atomic electron about an atomic nucleus as that of a classical particle of mass m_0, with an electric charge q, about an immobile attracting center. As classical mechanics tells us, the central attractive force does not change the particle's angular momentum $\mathbf{L} \equiv m_0 \mathbf{r} \times \mathbf{v}$, but the applied magnetic field $\mathbf{B}$ (that may be taken uniform on the atomic scale) does, due to the torque (101) it exerts on the magnetic moment (95):

$$\frac{d\mathbf{L}}{dt} = \boldsymbol{\tau} = \mathbf{m} \times \mathbf{B} = \frac{q}{2m_0} \mathbf{L} \times \mathbf{B} . \tag{5.114}$$

The diagram in Fig. 13 shows that in the limit of a relatively weak field, when the magnitude of the angular momentum $\mathbf{L}$ may be considered constant, this equation describes the rotation (called the *torque-induced precession*[50]) of the vector $\mathbf{L}$ about the direction of the vector $\mathbf{B}$, with the angular frequency $\boldsymbol{\Omega} = -q\mathbf{B}/2m_0$, independent of the angle θ. According to Eqs. (91) and (114), the resulting additional (field-induced) magnetic moment $\Delta\mathbf{m} \propto q\boldsymbol{\Omega} \propto -q^2\mathbf{B}/m_0$ has, irrespectively of the sign of q, a direction *opposite* to the field. Hence, according to Eq. (111) with $\mathbf{H} \approx \mathbf{B}/\mu_0$, the susceptibility $\chi_m \propto \chi \equiv \Delta\mathbf{m}/\mathbf{H}$ is indeed negative. (Let me leave its quantitative estimate within this classical model for the reader's exercise.) The quantum-mechanical treatment confirms this qualitative picture of the Larmor diamagnetism, giving only quantitative corrections to the classical result for χ_m.[51]

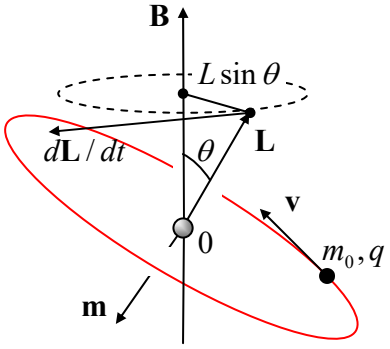


Fig. 5.13. The torque-induced precession of a classical charged particle in a magnetic field.

A simple estimate (also left for the reader's exercise) shows that in atoms with spontaneous non-zero net spins, the magnetic dipole orientation mechanism prevails over the orbital diamagnetism, so the materials incorporating such atoms usually exhibit net paramagnetism – see Table 1. Due to possible strong quantum interaction between the spin dipole moments, the magnetism of such materials is rather complex, with numerous interesting phenomena and elaborate theories. Unfortunately, all this physics is well outside the framework of this course, and I have to refer the interested reader to special literature,[52] but still will mention some key facts.

[49] Named after Sir Joseph Larmor who was the first (in 1897) to describe this effect mathematically.

[50] For a detailed discussion of this effect see, e.g., CM Sec. 4.5.

[51] See, e.g., QM Sec. 6.4. Quantum mechanics also explains why in most common (*s*-) ground states, the average contribution (95) of the orbital angular momentum $\mathbf{L}$ to the net vector $\mathbf{m}$ vanishes.

[52] See, e.g., D. J. Jiles, *Introduction to Magnetism and Magnetic Materials*, 2nd ed., CRC Press, 1998, or R. C. O'Handley, *Modern Magnetic Materials*, Wiley, 1999.

Most importantly, a sufficiently strong magnetic dipole-dipole interaction may lead to their spontaneous ordering, even in the absence of the applied field. This ordering may correspond to either parallel alignment of the dipoles (*ferromagnetism*) or anti-parallel alignment of the adjacent dipoles (*antiferromagnetism*). Evidently, the external effects of ferromagnetism are stronger, because this phase corresponds to a substantial spontaneous magnetization **M** even in the absence of an external magnetic field. (The corresponding magnitude of $\mathbf{B} = \mu_0\mathbf{M}$ is called the *remanence field*, B_R.) The direction of the vector $\mathbf{B}_R$ may be switched by the application of an external magnetic field, with a magnitude above a certain value H_C called *coercivity*, leading to the well-known hysteretic loops on the [H, B] plane (see Fig. 14 for a typical example) – similar to those in ferroelectrics, already discussed in Sec. 3.3.

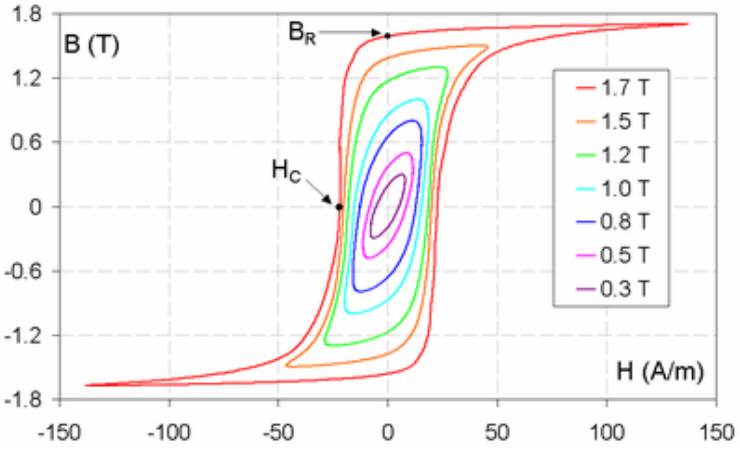


Fig. 5.14. Experimental magnetization curves of specially processed (cold-rolled) electrical steel – a solid solution of ~10% C and ~6% Si in Fe. (Reproduced from www.thefullwiki.org/Hysteresis under the Creative Commons BY-SA 3.0 license.)

Just as the ferroelectrics, the ferromagnets may also be *hard* or *soft* – in the magnetic rather than mechanical sense. In hard ferromagnets (also called *permanent magnets*), the dipole interaction is so strong that B stays close to B_R in all applied fields below H_C, so the hysteretic loops are virtually rectangular. Hence, in lower fields, the magnetization **M** of a permanent magnet may be considered constant, with the magnitude B_R/μ_0. Such hard ferromagnetic materials (notably, rare-earth compounds such as $SmCo_5$, Sm_2Co_{17}, and especially $Nd_2Fe_{14}B$), with high remanence fields (~1 T) and high coercivity (~10^6 A/m), have numerous practical applications.[53] Let me give just two, most important examples.

First, permanent magnets are the core components of most *electric motors*. By the way, this venerable (~150-years-old) technology is currently experiencing a quiet revolution, driven mostly by the electric car development. In the most advanced type of motors, called *permanent-magnet synchronous machines* (PMSM), the remanence magnetic field B_R of a permanent-magnet rotating part of the machine (called the *rotor*) interacts with the magnetic field of ac currents passed through wire windings in the external, static part of the motor (called the *stator*). The resulting torque may drive the rotor to extremely high speeds, exceeding 10,000 rotations per minute, enabling the motor to deliver several kilowatts of mechanical power from each kilogram of its mass.

As the second important example, despite the decades of the exponential (*Moore's-law*) progress of semiconductor electronics, most computer data storage systems (e.g., in data centers) are still based

[53] Currently, the neodymium-iron-boron compound holds nearly 95% percent of the world permanent-magnet application market, due to its combination of high B_R and H_C with lower fabrication costs.

on *hard disk drives* whose active media are submicron-thin layers of hard ferromagnets, with the data bits stored in the form of the direction of the remanent magnetization of small film spots. This technology has reached fantastic sophistication, with the recorded data density of the order of 10^{12} bits per square inch.[54] (Only recently it started to be seriously challenged by *solid-state drives* based on the floating-gate semiconductor memories already mentioned in Chapter 3.) [55]

In contrast, in soft ferromagnets, with their lower magnetic dipole interactions, the magnetization is constant only inside each of the spontaneously formed magnetic domains, while the volume and shape of the domains are affected by the applied magnetic field. As a result, the hysteresis loop's shape of soft ferromagnets is dependent on the cycled field's amplitude and cycling history – see Fig. 14. At high fields, their **B** (and hence **M**) is driven into saturation, with $B \approx B_R$, but at low fields, they behave essentially as linear magnetics with very high values of χ_m and hence μ – see the top rows of Table 1. (The magnetic domain interaction, and hence the low-field susceptibility of such soft ferromagnets are highly dependent on the material's fabrication technology and its post-fabrication thermal and mechanical treatments.) Due to these high values of μ, soft ferromagnets, especially iron and its alloys (e.g., various special steels), are extensively used in electrical engineering – for example in the cores of transformers – see the next section.

Due to the relative weakness of the magnetic dipole interaction in some materials, their ferromagnetic ordering may be destroyed by thermal fluctuations, if the temperature is increased above some value called the *Curie temperature* T_C, specific for each material. The transition between the ferromagnetic and paramagnetic phases at $T = T_C$ is a classical example of a *continuous phase transition*, with the average polarization **M** playing the role of the so-called *order parameter* that (in the absence of external fields) becomes different from zero only at $T < T_C$, increasing gradually at the further temperature reduction.[56]

5.6. Systems with magnetic materials

Just as the electrostatics of linear dielectrics, the magnetostatics is very simple in the particular case when all essential stand-alone currents are embedded into a linear magnetic medium with a constant permeability μ. Indeed, let us assume that we know the solution $\mathbf{B}_0(\mathbf{r})$ of the magnetic pair of

[54] "*A magnetic head slider* [the read/write head – KKL] *flying over a disk surface with a flying height of 25 nm with a relative speed of 20 meters/second* [all realistic parameters – KKL] *is equivalent to an aircraft flying at a physical spacing of 0.2 μm at 900 kilometers/hour.*" B. Bhushan, as quoted in the (generally good) book by G. Hadjipanayis, *Magnetic Storage Systems Beyond 2000*, Springer, 2001.

[55] The high-frequency properties of hard ferromagnets are also very non-trivial. For example, according to Eq. (101), an external magnetic field $\mathbf{B}_{ext}$ exerts torque $\boldsymbol{\tau} = \mathbf{M}\times\mathbf{B}_{ext}$ on the spontaneous magnetic moment **M** of a unit volume of a ferromagnet. In some nearly-isotropic, mechanically fixed ferromagnetic samples, this torque causes the precession, around the direction of $\mathbf{B}_{ext}$ (very similar to that illustrated in Fig. 13), of not the sample as such, but of the magnetization **M** inside it, with a certain frequency ω_r. If the frequency ω of an additional ac field becomes very close to ω_r, its absorption sharply increases – the so-called *ferromagnetic resonance*. Moreover, if ω is somewhat higher than ω_r, the effective magnetic permeability $\mu(\omega)$ of the material for the ac field may become *negative*, enabling a series of interesting effects and practical applications. Very unfortunately, I could not find time for their discussion in this series and have to refer the interested reader to literature, for example the monograph by A. Gurevich and G. Melkov, *Magnetization Oscillations and Waves*, CRC Press, 1996.

[56] In this series, a quantitative discussion of such transitions is given in SM Chapter 4.

the genuine ("microscopic") Maxwell equations (36) in free space, i.e. when the genuine current density **j** coincides with that of stand-alone currents. Then the macroscopic Maxwell equations (109) and the linear constitutive equation (110) are satisfied with the pair of functions

$$\mathbf{H}(\mathbf{r}) = \frac{\mathbf{B}_0(\mathbf{r})}{\mu_0}, \qquad \mathbf{B}(\mathbf{r}) = \mu\mathbf{H}(\mathbf{r}) = \frac{\mu}{\mu_0}\mathbf{B}_0(\mathbf{r}). \tag{5.115}$$

Hence the only effect of the complete filling of a fixed-current system with a uniform, linear magnetic medium is the change of the magnetic field **B** at all points by the same constant factor $\mu/\mu_0 \equiv 1 + \chi_m$, which may be either larger or smaller than 1. (As a reminder, a similar filling of a system of fixed stand-alone charges with a uniform, linear dielectric always leads to a reduction of the electric field **E** by a factor of $\varepsilon/\varepsilon_0 \equiv 1 + \chi_e$ – the difference whose physics was already discussed at the end of Sec. 4.)

However, this simple result is generally invalid in the case of nonuniform (or piecewise-uniform) magnetic samples. To analyze this case, let us first integrate the macroscopic Maxwell equation (107) along a closed contour *C* limiting a smooth surface *S*. Now using the Stokes theorem just as at the derivation of Eq. (37), we get the macroscopic version of the Ampère law (37):

Macroscopic Ampère law

$$\oint_C \mathbf{H}\cdot d\mathbf{r} = I. \tag{5.116}$$

Let us apply this relation to a sharp boundary between two regions with different magnetic materials, with no stand-alone currents on the interface, similarly to how this was done for the field **E** in Sec. 3.4 – see Fig. 3.5. The result is similar as well:

$$H_\tau = \text{const}. \tag{5.117}$$

On the other hand, the integration of the Maxwell equation (29) over a Gaussian pillbox enclosing a border fragment (again just as shown in Fig. 3.5 for the field **D**) yields a result similar to Eq. (3.35):

$$B_n = \text{const}. \tag{5.118}$$

For linear magnetic media, with $\mathbf{B} = \mu\mathbf{H}$, the latter boundary condition is reduced to

$$\mu H_n = \text{const}. \tag{5.119}$$

Let us use these boundary conditions, first of all, to see what happens with a long cylindrical sample of a uniform magnetic material, placed parallel to a uniform external magnetic field $\mathbf{B}_0$ – see Fig. 15. Such a sample cannot noticeably disturb the field in the free space outside it, at most of its length: $\mathbf{B}_{ext} = \mathbf{B}_0$, $\mathbf{H}_{ext} = \mu_0\mathbf{B}_{ext} = \mu_0\mathbf{B}_0$. Now applying Eq. (117) to the dominating surfaces of the sample, we get $\mathbf{H}_{int} = \mathbf{H}_0$.[57] For a linear magnetic material, these relations yield $\mathbf{B}_{int} = \mu\mathbf{H}_{int} = (\mu/\mu_0)\,\mathbf{B}_0$.[58] For the high-$\mu$ media, this means that $B_{int} >> B_0$. This effect may be vividly represented as the concentration of the magnetic field lines in high-μ samples – see Fig. 15 again. (The concentration affects the external field

[57] The independence of **H** on magnetic properties of the sample in this geometry explains why this field's magnitude is commonly used as the argument in the plots like Fig. 14: such measurements are typically carried out by placing an elongated sample of the material under study into a long solenoid with a controllable current *I*, so according to Eq. (116), $H_0 = nI$, regardless of the sample.

[58] The reader is highly encouraged to carry out a similar analysis of the fields inside narrow gaps cut in a linear magnetic material, similar to that carried in Sec. 3.3 out for linear dielectrics – see Fig. 3.6 and its discussion.

distribution only at distances of the order of $(\mu/\mu_0)\, t << l$ near the sample's ends.) Such concentration is widely used in such practically important devices as transformers, in which two multi-turn coils are wound on a ring-shaped (e.g., toroidal, see Fig. 6b) core made of a soft ferromagnetic material (such as the *transformer steel*, see Table 1) with $\mu >> \mu_0$. This minimizes the number of "stray" field lines, and makes the magnetic flux Φ piercing each wire turn of either coil virtually the same – the equality important for the secondary voltage induction – see the next chapter.

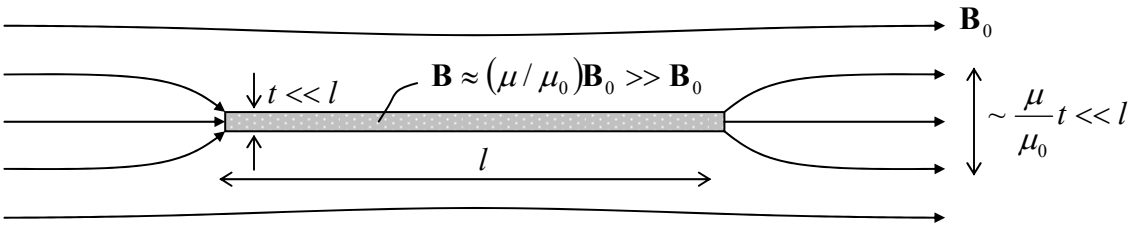


Fig. 5. 15. Magnetic field concentration in long, high-μ magnetic samples (schematically).

Samples of other geometries may create strong perturbations of the external field, extended to distances of the order of the sample's dimensions. To analyze such problems, we may benefit from a simple, partial differential equation for a scalar function, e.g., the Laplace equation, because in Chapter 2 we have learned how to solve it for many simple geometries. In magnetostatics, the introduction of a scalar potential is generally impossible due to the vortex-like magnetic field lines. However, if there are no stand-alone currents within the region we are interested in, then the macroscopic Maxwell equation (107) for the field $\mathbf{H}$ is reduced to $\nabla \times \mathbf{H} = 0$, similar to Eq. (1.28) for the electric field, showing that we may introduce the scalar potential of the magnetic field, $\phi_{\rm m}$, using a relation similar to Eq. (1.33):

$$\mathbf{H} = -\nabla \phi_{\rm m}. \tag{5.120}$$

Combining it with the homogenous Maxwell equation (29) for the magnetic field, $\nabla \cdot \mathbf{B} = 0$, and Eq. (110) for a linear magnetic material, we arrive at a single differential equation, $\nabla \cdot (\mu \nabla \phi_{\rm m}) = 0$. For a uniform medium ($\mu(\mathbf{r}) = \text{const}$), it is reduced to our beloved Laplace equation:

$$\nabla^2 \phi_{\rm m} = 0. \tag{5.121}$$

Moreover, Eqs. (117) and (119) give us very familiar boundary conditions: the first of them

$$\frac{\partial \phi_{\rm m}}{\partial \tau} = \text{const}, \tag{5.122a}$$

being equivalent to

$$\phi_{\rm m} = \text{const}, \tag{5.122b}$$

while the second one giving

$$\mu \frac{\partial \phi_{\rm m}}{\partial n} = \text{const}. \tag{5.123}$$

Indeed, these boundary conditions are absolutely similar for (3.37) and (3.56) of electrostatics, with the replacement $\varepsilon \to \mu$.[59]

[59] This similarity may seem strange because earlier we have seen that the parameter μ is physically more similar to $1/\varepsilon$. The reason for this paradox is that in magnetostatics, the magnetic potential $\phi_{\rm m}$ is traditionally used to

Let us analyze the geometric effects on magnetization, first using the (too?) familiar structure: a sphere, made of a linear magnetic material, placed into a uniform external field $\mathbf{H}_0 \equiv \mathbf{B}_0/\mu_0$. Since the differential equation and the boundary conditions are similar to those of the corresponding electrostatics problem (see Fig. 3.11 and its discussion), we can use the above analogy to reuse the solution we already have – see Eqs. (3.63). Just as in the electric case, the field outside the sphere, with

$$(\phi_{\rm m})_{r>R} = H_0\left(-r + \frac{\mu-\mu_0}{\mu+2\mu_0}\frac{R^3}{r^2}\right)\cos\theta, \tag{5.124}$$

is a sum of the uniform external field $\mathbf{H}_0$, with the potential $-H_0 r\cos\theta \equiv -H_0 z$, and the dipole field (99) with the following induced magnetic dipole moment of the sphere:[60]

$$\mathbf{m} = 4\pi\frac{\mu-\mu_0}{\mu+2\mu_0}R^3\mathbf{H}_0. \tag{5.125}$$

On the contrary, the internal field is perfectly uniform, and directed along the external one:

$$(\phi_{\rm m})_{r<R} = -H_0\frac{3\mu_0}{\mu+2\mu_0}r\cos\theta, \quad \text{so that} \quad \frac{H_{\rm int}}{H_0} = \frac{3\mu_0}{\mu+2\mu_0}, \qquad \frac{B_{\rm int}}{B_0} = \frac{\mu H_{\rm int}}{\mu_0 H_0} = \frac{3\mu}{\mu+2\mu_0}. \tag{5.126}$$

Note that the field $\mathbf{H}_{\rm int}$ inside the sphere is *not* equal to the applied external field $\mathbf{H}_0$. This example shows that the interpretation of $\mathbf{H}$ as the "would-be" magnetic field generated by external stand-alone currents $\mathbf{j}$ should not be exaggerated by saying that its distribution is independent of the magnetic bodies in the system. In the limit $\mu >> \mu_0$, Eqs. (126) yield $H_{\rm int}/H_0 << 1$, $B_{\rm int}/H_0 = 3\mu_0$, the factor 3 being specific for the particular geometry of the sphere. If a sample is strongly stretched along the applied field, with its length l much larger than the scale t of its cross-section, this geometric effect is gradually decreased, and $B_{\rm int}$ tends to its value $\mu H_0 >> B_0$, as was discussed above – see Fig. 15.

Now let us calculate the field distribution in a similar, but slightly more complex (and practically important) system: a round cylindrical shell, made of a linear magnetic material, placed into a uniform external field $\mathbf{H}_0$ normal to its axis – see Fig. 16.

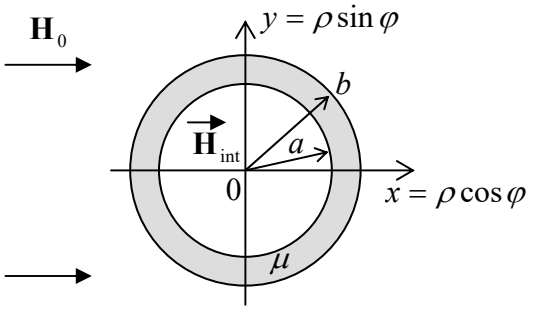


Fig. 5.16. Cylindrical magnetic shield.

describe the "would-be field" $\mathbf{H}$, while in electrostatics, the potential ϕ describes the actual electric field $\mathbf{E}$. (This tradition persists from the days when $\mathbf{H}$ was perceived as a genuine magnetic field.)

[60] To derive Eq. (125), we may either calculate the gradient of the $\phi_{\rm m}$ given by Eq. (124), or use the similarity of Eqs. (3.13) and (99), to derive from Eq. (3.17) a similar expression for the magnetic dipole's potential:

$$\phi_{\rm m} = \frac{1}{4\pi}\frac{m\cos\theta}{r^2}.$$

Now comparing this formula with the second term of Eq. (124), we immediately get Eq. (125).

Since there are no stand-alone currents in the region of our interest, we can again represent the field $\mathbf{H}(\mathbf{r})$ by the gradient of the magnetic potential ϕ_m – see Eq. (120). Inside each of three constant-μ regions, i.e. at $\rho < b$, $a < \rho < b$, and $b < \rho$ (where ρ is the 2D distance from the cylinder's axis), the potential obeys the Laplace equation (121). In the convenient, polar coordinates (see Fig. 16), we may, guided by the general solution (2.112) of the Laplace equation and our experience in its application to axially-symmetric geometries, look for ϕ_m in the following form:

$$\phi_m = \begin{cases} \left(-H_0\rho + b_1'/\rho\right)\cos\varphi, & \text{for } b \le \rho, \\ \left(a_1\rho + b_1/\rho\right)\cos\varphi, & \text{for } a \le \rho \le b, \\ -H_{\text{int}}\rho\cos\varphi, & \text{for } \rho \le a\,. \end{cases} \tag{5.127}$$

Plugging this solution into the boundary conditions (122)-(123) at both interfaces ($\rho = b$ and $\rho = a$), we get the following system of four equations:

$$\begin{aligned} -H_0 b + b_1'/b &= a_1 b + b_1/b, & \left(a_1 a + b_1/a\right) &= -H_{\text{int}} a, \\ \mu_0\left(-H_0 - b_1'/b^2\right)H_0 &= \mu\left(a_1 - b_1/b^2\right), & \mu\left(a_1 - b_1/a^2\right) &= -\mu_0 H_{\text{int}}, \end{aligned} \tag{5.128}$$

for four unknown coefficients a_1, b_1, b_1', and H_{int}. Solving the system, we get, in particular:

$$\frac{H_{\text{int}}}{H_0} = \frac{\alpha_c - 1}{\alpha_c - (a/b)^2}, \qquad \text{with } \alpha_c \equiv \left(\frac{\mu + \mu_0}{\mu - \mu_0}\right)^2. \tag{5.129}$$

According to these formulas, at $\mu > \mu_0$, the field in the free space inside the cylinder is lower than the external field. This fact allows using such structures, made of high-μ materials such as permalloy (see Table 1), for passive shielding[61] from unintentional magnetic fields (e.g., the Earth's field) – the task very important for the design of many physical experiments. As Eq. (129) shows, the larger is μ, the closer is α_c to 1, and the smaller is the ratio H_{int}/H_0, i.e. the better is the shielding, for the same a/b ratio. On the other hand, for a given magnetic material, i.e. for a fixed parameter α_c, the shielding is improved by making the ratio $a/b < 1$ smaller, i.e. the shield thicker. On the other hand, as Fig. 16 shows, smaller a leaves less space for the shielded samples, calling for a compromise.

Note that in the limit $\mu/\mu_0 \to \infty$, both Eq. (126) and Eq. (129), describing different geometries, yield $H_{\text{int}}/H_0 \to 0$. Indeed, as it follows from Eq. (119), in this limit, the field $\mathbf{H}$ tends to zero inside magnetic samples of virtually any geometry. (The formal exception is the longitudinal cylindrical geometry shown in Fig. 15, with $t/l \to 0$, where $\mathbf{H}_{\text{int}} = \mathbf{H}_0$ for any finite μ, but even in it, the last equality holds only if $t/l << \mu_0/\mu$.)

Now let us discuss a curious (and practically important) approach to systems with relatively thin, closed magnetic cores made of several sections of high-μ magnetic materials, with the cross-section areas A_k much smaller than the squared lengths l_k of the sections – see Fig. 17. If all $\mu_k >> \mu_0$, virtually all field lines are confined to the interior of the core. Then, applying the macroscopic Ampère law (116) to a contour C that follows a magnetic field line inside the core (see, for example, the dashed line in Fig. 17), we get the following approximate expression (exactly valid only in the limit μ_k/μ_0, $l_k^2/A_k \to \infty$):

[61] Another approach to the undesirable magnetic fields' reduction is the "active shielding" – the external field's compensation with the counter-field induced by controlled currents in specially designed wire coils.

$$\oint_C H_l dl \approx \sum_k l_k H_k \equiv \sum_k l_k \frac{B_k}{\mu_k} = NI\,. \tag{5.130}$$

However, since the magnetic field lines stay in the core, the magnetic flux $\Phi_k \approx B_k A_k$ should be the same ($\equiv \Phi$) for each section, so $B_k = \Phi/A_k$. Plugging this condition into Eq. (130), we get

Magnetic Ohm law and reluctance

$$\Phi = \frac{NI}{\sum_k \mathscr{R}_k}, \qquad \text{where } \mathscr{R}_k \equiv \frac{l_k}{\mu_k A_k}\,. \tag{5.131}$$

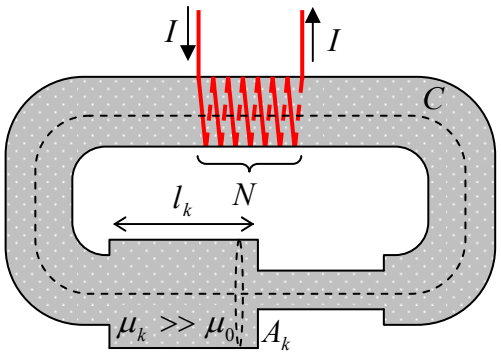


Fig. 5.17. Deriving the "magnetic Ohm law" (131).

Note a close analogy of the first of these equations with the usual Ohm law for several resistors connected in series, with the magnetic flux playing the role of electric current, while the product NI, the role of the voltage applied to the chain of resistors. This analogy is fortified by the fact that the second of Eqs. (131) is similar to the expression for the resistance $R = l/\sigma A$ of a long, uniform conductor, with the magnetic permeability μ playing the role of the electric conductivity σ. (To sound similar, but still different from the resistance R, the parameter $\mathscr{R}$ is called *reluctance*.) This is why Eq. (131) is called the *magnetic Ohm law*; it is very useful for approximate analyses of systems like ac transformers, magnetic energy storage systems, etc.

Now let me proceed to a brief discussion of systems with permanent magnets. First of all, using the definition (108) of the field **H**, we may rewrite the Maxwell equation (29) for the field **B** as

$$\nabla \cdot \mathbf{B} \equiv \mu_0 \nabla \cdot (\mathbf{H} + \mathbf{M}) = 0, \qquad \text{i.e. as } \nabla \cdot \mathbf{H} = -\nabla \cdot \mathbf{M}\,, \tag{5.132}$$

While this relation is general, it is especially convenient in permanent magnets, where the magnetization vector **M** may be approximately considered field-independent.[62] In this case, Eq. (132) for **H** is an exact analog of Eq. (1.27) for **E**, with the fixed term $-\nabla\cdot\mathbf{M}$ playing the role of the fixed charge density (more exactly, of ρ/ε_0). For the scalar potential ϕ_m, defined by Eq. (120), this gives the Poisson equation

$$\nabla^2 \phi_m = \nabla \cdot \mathbf{M}\,, \tag{5.133}$$

similar to those solved, for quite a few geometries, in the previous chapters.

In the case when **M** is not only field-independent but also uniform inside a permanent magnet's volume, then the right-hand sides of Eqs. (132) and (133) vanish both inside the volume and in the

[62] Note that in this approximation, there is no difference between the *remanence magnetization* $M_R \equiv B_R/\mu_0$, of the magnet and its *saturation magnetization* $M_S \equiv \lim_{H\to\infty}[B(H)/\mu_0 - H]$.

surrounding free space, and give a non-zero effective charge only on the magnet's surface. Integrating Eq. (132) along a short path normal to the surface and crossing it, we get the following boundary conditions:

$$\Delta H_n \equiv (H_n)_{\text{in free space}} - (H_n)_{\text{in magnet}} = M_n \equiv M\cos\theta, \tag{5.134}$$

where θ is the angle between the magnetization vector and the outer normal to the magnet's surface. This relation is an exact analog of Eq. (1.24) for the normal component of the field **E**, with the effective surface charge density (or rather σ/ε_0) equal to $M\cos\theta$.

This analogy between the magnetic field induced by a fixed, constant magnetization and the electric field induced by surface electric charges enables one to reuse the solutions of quite a few problems considered in Chapters 1-3. Leaving a few such problems for the reader's exercise (see Sec. 7), let me demonstrate the power of this analogy on just two examples specific to magnetic systems. First, let us calculate the force necessary to detach the flat ends of two long, uniform rod magnets, of length l and cross-section area $A << l^2$, with the saturated remanent magnetization $\mathbf{M}_0$ directed along their length – see Fig. 18.

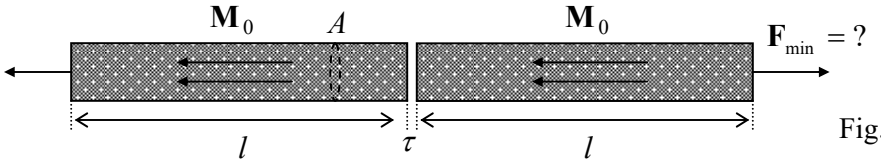


Fig. 5.18. Detaching two magnets.

Let us assume we have succeeded in separating the magnets by an infinitesimal distance $\tau << A^{1/2}, l$. Then, according to Eqs. (133)-(134), the distribution of the magnetic field near this small gap should be similar to that of the electric field in a system of two equal by opposite surface charges with the surface density σ proportional to M_0. From Chapters 1-3, we know the properties of such a system very well: within the gap, the field is virtually constant, uniform, proportional to σ, and independent of τ. For its magnitude in the magnetic case, Eq. (134) gives simply $H = M_0$, and hence $B = \mu_0 M_0$. (Just outside of the gap, the field is very low, because due to the condition $A << l^2$, the effect of the similar effective charges at the "outer" ends of the rods on the field near the gap t is negligible.)

From here, we can readily calculate $F_{\min}$ as the force exerted by this field on the effective surface "charges". However, it is even easier to find it from the following energy argument. Since the magnetic field energy localized inside the magnets and near their outer ends cannot depend on τ, this small detachment may only alter the energy inside the gap. For this part of the energy, Eq. (57) yields:

$$\Delta U = \frac{B^2}{2\mu_0} V = \frac{(\mu_0 M_0)^2}{2\mu_0} A\tau\,. \tag{5.135}$$

The gradient of this potential energy is equal to the attraction force $\mathbf{F} = -\nabla(\Delta U)$, trying to reduce ΔU by decreasing the gap, with the following magnitude:

$$|F| = \frac{\partial(\Delta U)}{\partial\tau} = \frac{\mu_0 M_0^2 A}{2}\,. \tag{5.136}$$

The magnet detachment requires an equal and opposite external force. For a typical permanent magnet, with $\mu_0 M_0 \approx B_R \sim 1$T, the force corresponds to a ratio $|F|/A$ close to 4×10^5 Pa, a few times the normal atmospheric pressure.

Now let us consider the situation when similar long permanent magnets (such as the *magnetic needles* used in magnetic compasses) are separated, in otherwise free space, by a larger distance $d >> A^{1/2}$ – see Fig. 19. For each needle (Fig. 19a), of a length $l >> A^{1/2}$, the right-hand side of Eq. (133) is substantially different from zero only in two relatively small areas at the needle's ends. Integrating the equation over each area, we see that at distances $r >> A^{1/2}$ from each end, we may reduce Eq. (132) to

$$\nabla \cdot \mathbf{H} = \frac{q_m}{\mu_0}\left[\delta(\mathbf{r}-\mathbf{r}_+) - \delta(\mathbf{r}-\mathbf{r}_-)\right], \qquad \text{i.e. } \nabla \cdot \mathbf{B} = q_m\left[\delta(\mathbf{r}-\mathbf{r}_+) - \delta(\mathbf{r}-\mathbf{r}_-)\right], \tag{5.137}$$

where $\mathbf{r}_\pm$ are the ends' positions, and $q_m \equiv \mu_0 M_0 A$, with A being the needle's cross-section area.[63] This expression for **B** is completely similar to Eq. (3.32) for the electric displacement **D**, for the particular case of two equal and opposite point charges, i.e. with $\rho = q[\delta(\mathbf{r} - \mathbf{r}_+) - \delta(\mathbf{r} - \mathbf{r}_+)]$, with the only replacement $q \to q_m$. Since we know the resulting electric field all too well (see, e.g., Eq. (1.7) for $\mathbf{E} \equiv \mathbf{D}/\varepsilon_0$), we may immediately write a similar expression for the field **H**:

$$\mathbf{H}(\mathbf{r}) = \frac{q_m}{4\pi\mu_0}\left(\frac{\mathbf{r}-\mathbf{r}_+}{\left|\mathbf{r}-\mathbf{r}_+\right|^3} - \frac{\mathbf{r}-\mathbf{r}_-}{\left|\mathbf{r}-\mathbf{r}_-\right|^3}\right). \tag{5.138}$$

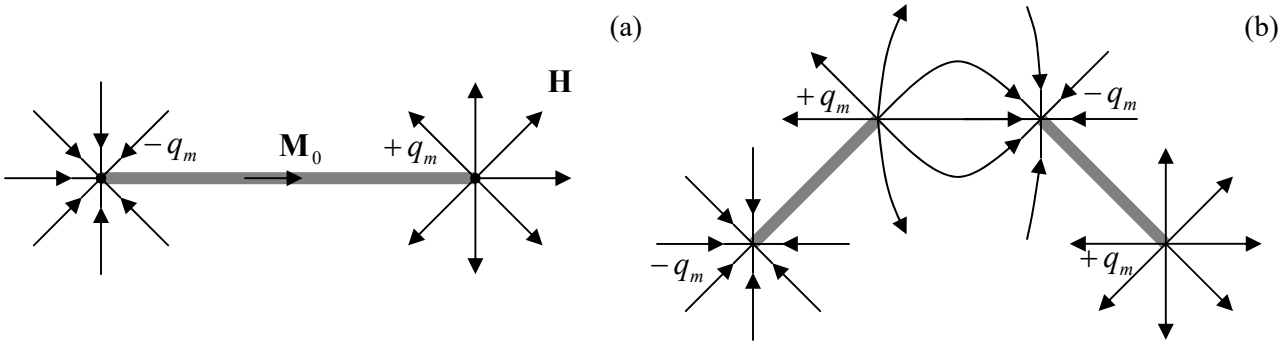


Fig. 5.19. (a) "Magnetic charges" at the ends of a thin permanent-magnet needle and (b) the result of its breaking into two parts (schematically).

The resulting magnetic field $\mathbf{H}(\mathbf{r})$ exerts on another "magnetic charge" q'_m, located at some point $\mathbf{r}'$, the force $\mathbf{F} = q'_m\mathbf{H}(\mathbf{r}')$.[64] Hence if two ends of different needles are separated by an intermediate distance R ($A^{1/2} << R << l$, see Fig. 19b), we may neglect one term in Eq. (138), and get the following "magnetic Coulomb law" for the interaction of the nearest ends:

$$\mathbf{F} = \pm\frac{q_m q'_m}{4\pi\mu_0}\frac{\mathbf{R}}{R^3}. \tag{5.139}$$

The "only" (but conceptually, crucial!) difference between this interaction and that of the electric point charges is that the two "magnetic charges" (quasi-monopoles) of a magnetic needle cannot be fully separated. For example, if we break a needle in the middle in an attempt to bring its two ends further apart, two new "point charges" appear – see Fig. 19b.

[63] Note that the constant coefficient in the definition of q_m, and hence in Eqs. (138)-(139), is the matter of convention. The above choice makes the free-space Maxwell equations $\nabla\cdot\mathbf{D} = \rho$ and $\nabla\cdot\mathbf{B} = \rho_m$ (where ρ and ρ_m are the volumic densities of the electric and magnetic charges) pleasantly symmetric.

[64] This expression is the magnetic analog of the basic equation $\mathbf{F} = q'_e\mathbf{E}(\mathbf{r}')$ for the electric charges.

There are several solid-state systems where more flexible structures, similar in their magnetostatics to the needles, may be implemented. First of all, certain ("type-II") superconductors may carry so-called *Abrikosov vortices* – flexible tubes with field-suppressed superconductivity inside, each carrying one quantum $\Phi_0 = \pi\hbar/e \approx 2\times10^{-15}$ Wb of the magnetic flux. Ending on superconductor's surfaces, these tubes let their magnetic field lines spread into the surrounding free space, essentially forming magnetic monopole analogs – of course, with equal and opposite "magnetic charges" q_m on each end of the tube – just as Fig. 19a shows. Such flux tubes are not only flexible but also stretchable, resulting in several peculiar effects – see Sec. 6.4 for more detail. Another recently found example of such paired quasi-monopoles is *spin chains* in the so-called *spin ices* – crystals with paramagnetic ions arranged into a specific (pyrochlore) lattice – such as dysprosium titanate $Dy_2Ti_2O_7$.[65] Let me emphasize again that any reference to magnetic monopoles in such systems should not be taken literally.

In order to complete this section (and this chapter), let me briefly discuss the magnetic field energy U, for the simplest case of systems with linear magnetic materials. In this case, we still may use Eq. (55), but if we want to operate only with macroscopic fields, and hence only stand-alone currents, we should repeat the manipulations that have led us to Eq. (57), using **j** not from Eq. (35), but from Eq. (107). As a result, instead of Eq. (57) we get

$$U = \int_V u(\mathbf{r})d^3r, \quad \text{with } u = \frac{\mathbf{B}\cdot\mathbf{H}}{2} = \frac{B^2}{2\mu} = \frac{\mu H^2}{2}, \tag{5.140}$$

Magnetic field energy: linear medium

This result is evidently similar to Eq. (3.73) of electrostatics.

As a simple but important example of its application, let us again consider a long solenoid (Fig. 6a), but now filled with a linear magnetic material with permeability μ. Using the macroscopic Ampère law (116), just as we used Eq. (37) for the derivation of Eq. (40), we get

$$H = In, \quad \text{and hence } B = \mu In, \tag{5.141}$$

where $n \equiv N/l$, just as in Eq. (40), is the winding density, i.e. the number of wire turns per unit length. (At $\mu = \mu_0$, we immediately return to that old result.) Now we may plug Eq. (141) into Eq. (140) to calculate the magnetic energy stored in the solenoid:

$$U = uV = \frac{\mu H^2}{2} lA = \frac{\mu(nI)^2 lA}{2}, \tag{5.142}$$

and then use Eq. (72) to calculate its self-inductance:[66]

$$L = \frac{U}{I^2/2} = \mu n^2 lA \tag{5.143}$$

We see that $L \propto \mu V$, so filling a solenoid with a high-μ material may allow making it more compact while preserving the same value of inductance. In addition, as the discussion of Fig. 15 has shown, such filling reduces the fringe fields near the solenoid's ends, which may be detrimental for some applications, especially in physical experiments striving for high measurement precision.

[65] See, e.g., L. Jaubert and P. Holdworth, *J. Phys. – Cond. Matt.* **23**, 164222 (2011), and references therein.

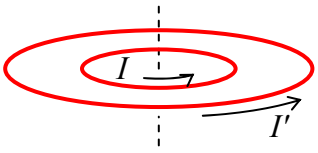
[66] Admittedly, we could get the same result simpler, just by arguing that since the magnetic material fills the whole volume of a substantial magnetic field in this system, the filling simply increases the vector **B** at all points, and hence its flux Φ, and hence $L \equiv \Phi/I$ by the factor μ/μ_0 in comparison with the free-space value (75).

However, we still need to explore the issue of magnetic energy beyond Eq. (140), not only to get a general expression for it in materials with an arbitrary dependence **B**(**H**), but also to finally prove Eq. (54) and explore its relation with Eq. (53). I will do this at the beginning of the next chapter.

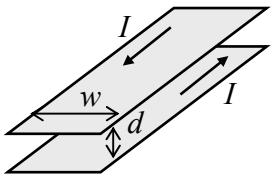
5.7. Exercise problems

5.1. DC current I flows around a thin wire loop bent into the form of a plane equilateral triangle with side a. Calculate the magnetic field in the center of the loop.

5.2. A circular wire loop, carrying a fixed dc current, is placed inside a similar but larger loop, carrying a fixed current in the same direction – see the figure on the right. Use semi-quantitative arguments to analyze the mechanical stability of the coaxial and coplanar position of the inner loop with respect to its possible angular, axial, and lateral displacements relative to the outer loop.



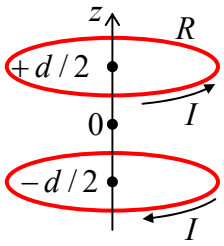
5.3. Two planar, parallel, long, thin conducting strips of width w, separated by distance d, carry equal but oppositely directed currents I – see the figure on the right. Calculate the magnetic field in the plane located in the middle between the strips, assuming that the flowing currents are uniformly distributed across the strip widths.



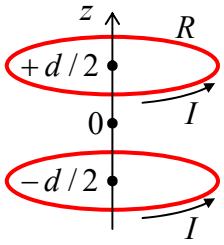
5.4. For the system studied in the previous problem, but now only in the limit $d << w$, calculate:

(i) the distribution of the magnetic field in space,
(ii) the vector potential of the field,
(iii) the magnetic force (per unit length) exerted on each strip, and
(iv) the magnetic energy and self-inductance of the loop formed by the strips (per unit length).

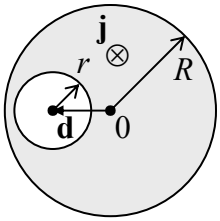
5.5. Calculate the magnetic field distribution near the center of the system of two similar, plane, round, coaxial wire coils, carrying equal but oppositely directed currents – see the figure on the right.



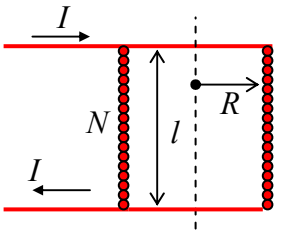
5.6. The two-coil-system considered in the previous problem now carries equal and *similarly* directed currents – see the figure on the right.[67] Calculate what should be the ratio d/R for the second derivative $\partial^2 B_z/\partial z^2$ to equal zero at $z = 0$.



[67] This *Helmholtz coils* system, producing a highly uniform field near its center, is broadly used in physical experiment.



5.7. DC current of a constant density j flows along a round cylindrical wire of radius R, with a round cylindrical cavity of radius r cut in it. The cavity's axis is parallel to that of the wire but offset from it by a distance $d < R - r$ (see the figure on the right). Calculate the magnetic field inside the cavity.



5.8. Calculate the magnetic field's distribution along the axis of a straight solenoid (see Fig. 6a, partly reproduced on the right) of a finite length l, and a round cross-section of radius R. Assume that the solenoid has many ($N >> 1$, l/R) wire turns, uniformly distributed along its length.

5.9. A thin round disk of radius R, carrying an electric charge of a constant areal density σ, rotates about its axis with a constant angular velocity ω. Calculate:

(i) the magnetic field on the disk's axis,

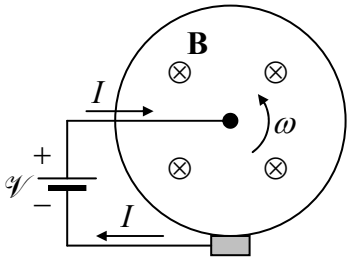
(ii) the magnetic moment of the disk,

and relate these results.

5.10. A thin spherical shell of radius R, with charge Q uniformly distributed over its surface, rotates about its diameter with a constant angular velocity ω. Calculate the distribution of the magnetic field everywhere in space.

5.11. A sphere of radius R, made of an insulating material with a uniform electric charge density ρ, rotates about its diameter with a constant angular velocity ω. Calculate the magnetic field distribution inside the sphere and outside it.

5.12. A conducting sphere with no total electric charge is rotated about its diameter with a constant angular velocity ω, in a uniform constant external magnetic field **B** directed along the rotation axis. Assuming that the sphere's contribution to the magnetic field is negligibly small, calculate the stationary distribution of the electric charge density inside the sphere and on its surface, and the electrostatic potential both inside and outside the sphere. Quantify the above assumption.



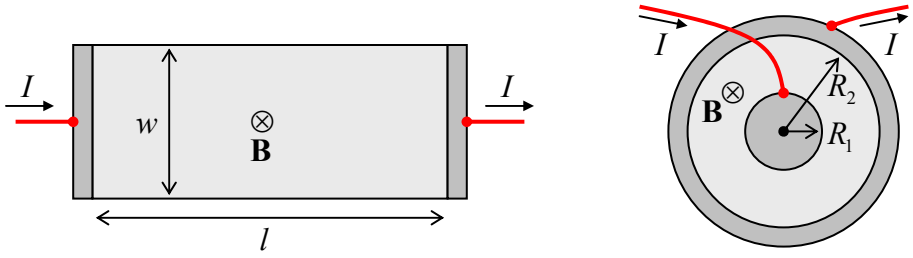
5.13.* The simplest version of the famous *homopolar* (or "unipolar") *motor* is a thin round conducting disk, placed into a uniform magnetic field normal to its plane, with dc current passed between the disk's center and a sliding electrode ("brush") on its rim – see the figure on the right.

(i) Express the torque rotating the disk via its radius R, the magnetic field **B**, and the current I.

(ii) If the disk is allowed to rotate about its axis, and the motor is driven by a battery with e.m.f. $\mathscr{V}$, calculate its stationary angular velocity ω, neglecting friction and the electric circuit's resistance.

(iii) Now assuming that the current's path (battery + wires + contacts + disk itself) has a non-zero resistance $\mathscr{R}$, derive and solve the equation for the time evolution of ω, and analyze the solution.

5.14. The reader is hopefully familiar with the classical Hall effect in the usual rectangular *Hall bar* geometry – see the left panel of the figure below. However, the effect takes a different form in the so-called *Corbino disk* – see the right panel of the figure below. (Dark shading shows electrodes, with no appreciable resistance.) Analyze the effect in both geometries, assuming that in both cases, the conductors are thin and planar, have a constant Ohmic conductivity σ and charge carrier density n, and that the applied magnetic field **B** is uniform and normal to conductors' planes.



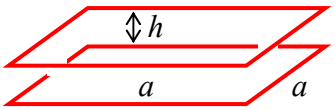
5.15. A wire with a round cross-section of radius a has been bent into a round loop of radius $R >> a$. Prove the formula for its self-inductance, which was mentioned at the end of Sec. 5.3 of the lecture notes: $L = \mu_0 R \ln(cR/a)$, with $c \sim 1$.

5.16. Prove that:

(i) the self-inductance L of a current loop cannot be negative, and

(ii) any inductance coefficient $L_{kk'}$, defined by Eq. (60), cannot be larger than $(L_{kk}L_{k'k'})^{1/2}$.

5.17. Calculate the mutual inductance of two similar thin-wire square-shaped loops offset by distance h in the direction normal to their planes – see the figure on the right.



5.18.* Estimate the values of magnetic susceptibility due to

(i) orbital diamagnetism, and

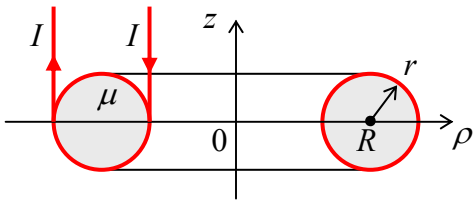
(ii) spin paramagnetism,

for a medium with negligible interactions between the induced molecular dipoles. Compare the results.

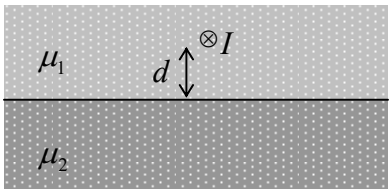
Hints: For Task (i), you may use the classical model described by Eq. (114) – see Fig. 13. For Task (ii), assume the ordering of spontaneous magnetic dipoles $\mathbf{m}_0$, with a fixed magnitude m_0 of the order of the Bohr magneton μ_B, similar to the one sketched for electric dipoles in Fig. 3.7a.

5.19.* Use the classical picture of the orbital ("Larmor") diamagnetism, discussed in Sec. 5, to calculate its (small) contribution $\Delta\mathbf{B}(0)$ to the magnetic field **B** felt by an atomic nucleus, treating the electrons of the atom as a spherically symmetric cloud with an electric charge density $\rho(r)$. Express the result via the value $\phi(0)$ of the electrostatic potential of the electron cloud, and use this expression for a crude numerical estimate of the ratio $\Delta B(0)/B$ for the hydrogen atom.

5.20. Calculate the self-inductance of a toroidal solenoid with a round cross-section of radius $r \sim R$, with $N >> 1$, R/r wire turns uniformly distributed along the perimeter, and filled with a linear magnetic material of permeability μ.



5.21. A long, straight, thin wire carrying current I runs parallel to the plane boundary between two uniform, linear magnetic media – see the figure on the right. Calculate the magnetic field everywhere in the system, and the force (per unit length) exerted on the wire.

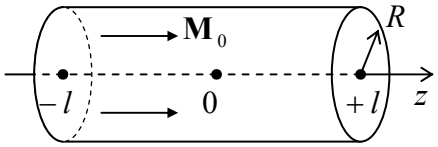


5.22. Solve the magnetic shielding problem similar to that discussed in Sec. 5.6 of the lecture notes, but for a spherical rather than cylindrical shell, with the same central cross-section as shown in Fig. 16. Compare the efficiency of those two shields, for the same shell's permeability μ, and the same b/a ratio.

5.23. Calculate the magnetic field's distribution around a spherical permanent magnet with uniform magnetization $\mathbf{M}_0 = \text{const}$.

5.24. A limited volume V is filled with a magnetic material with field-independent magnetization $\mathbf{M}(\mathbf{r})$. Write explicit expressions for the magnetic field induced by the magnetization and its potential, and recast these expressions into the forms that are more convenient when $\mathbf{M}(\mathbf{r}) = \mathbf{M}_0 = \text{const}$ throughout the volume.

5.25. Use the results of the previous problem to calculate the distribution of the magnetic field $\mathbf{H}$ along the axis of a straight permanent magnet of length $2l$ and a round cross-section of radius R, with a uniform magnetization $\mathbf{M}_0$ parallel to the axis – see the figure on the right.



5.26. A flat end of a long straight permanent magnet, similar to that considered in the previous problem but with an arbitrary cross-section of area A, is stuck to a flat surface of a large sample of a linear magnetic material with a very high permeability $\mu >> \mu_0$. Calculate the normally directed force needed to detach them.

5.27. A permanent magnet with a uniform magnetization $\mathbf{M}_0$ has the form of a spherical shell with an internal radius R_1 and an external radius $R_2 > R_1$. Calculate the magnetic field inside the shell.

5.28. A very broad film of thickness $2t$ is permanently magnetized normally to its plane, with a periodic checkerboard pattern, with the square of area $a \times a$:

$$\mathbf{M}\Big|_{|z|<t} = \mathbf{n}_z M(x,y), \qquad \text{with } M(x,y) = M_0 \,\mathrm{sgn}\left(\cos\frac{\pi x}{a}\cos\frac{\pi y}{a}\right).$$

Calculate the magnetic field's distribution in space.

5.29.* Based on the discussion of the quadrupole electrostatic lens in Sec. 2.4, suggest the permanent-magnet systems that may similarly focus particles moving close to the system's axis, for the cases when each particle carries:

(i) an electric charge,

(ii) no net electric charge, but a spontaneous magnetic dipole moment $\mathbf{m}$ of a certain orientation.

Chapter 6. Electromagnetism

This chapter discusses two major effects that arise when electric and magnetic fields change over time: the "electromagnetic induction" of an additional electric field by changing the magnetic field, and the reciprocal effect of the "displacement currents"– actually, the induction of an additional magnetic field by changing electric field. These two phenomena, which make time-dependent electric and magnetic fields inseparable (hence the term "electromagnetism"[1]), are reflected in the full system of Maxwell equations, valid for an arbitrary electromagnetic process. On the way toward this system, I will make a brief detour to review the electrodynamics of superconductivity, which (besides its own significance), provides a perfect platform for discussion of the important general issue of gauge invariance.

6.1. Electromagnetic induction

As Eqs. (5.36) show, in static situations ($\partial/\partial t = 0$) the Maxwell equations describing the electric and magnetic fields are independent – more exactly, coupled only implicitly, via the continuity equation (4.5) relating their right-hand sides ρ and $\mathbf{j}$. In dynamics, when the fields change in time, the situation is different.

Historically, the first discovered explicit coupling between the electric and magnetic fields was the effect of electromagnetic induction. Although this effect was discovered independently by Joseph Henry, it was a brilliant series of experiments by Michael Faraday, carried out mostly in 1831, that resulted in the first general formulation of the induction law. The summary of Faraday's numerous experiments has turned out to be very simple: if the magnetic flux defined by Eq. (5.65),

$$\Phi \equiv \int_S B_n d^2 r \,, \tag{6.1}$$

through a surface S limited by a closed contour C, changes in time by whatever reason (e.g., either due to a change of the magnetic field $\mathbf{B}$ (as in Fig.1), or the contour's motion, or its deformation, or any combination of the above), it induces an additional, vortex-like electric field $\mathbf{E}_{\text{ind}}$ directed along the contour – see Fig. 1.

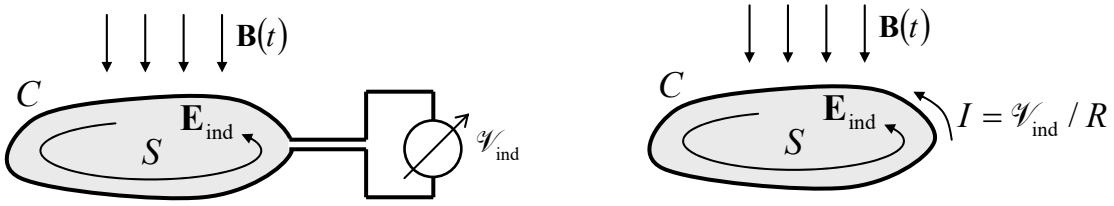


Fig. 6.1. Two simplest ways to observe the Faraday electromagnetic induction.

The exact distribution of $\mathbf{E}_{\text{ind}}$ in space depends on the system's details, but its integral along the contour C, called the *inductive electromotive force* (e.m.f.), obeys a very simple *Faraday induction law*:

[1] It was coined by H. Ørsted in 1820 in the context of his experiments – see the previous chapter.

$$\mathscr{V}_{\rm ind} \equiv \oint_C \mathbf{E}_{\rm ind} \cdot d\mathbf{r} = -\frac{d\Phi}{dt}. \qquad (6.2)$$

Faraday induction law

(In the Gaussian units, the right-hand side of this formula has an additional coefficient of $1/c$.)

It is straightforward (and hence left for the reader's exercise) to show that this e.m.f. may be measured, for example, either by inserting a voltmeter into a conducting loop following the contour C or by measuring the small current $I = \mathscr{V}_{\rm ind}/R$ it induces in a thin wire with a sufficiently large Ohmic resistance R,[2] whose shape follows that contour – see Fig. 1. (Actually, these methods are not entirely different, because a typical voltmeter measures voltage by the small Ohmic current it drives through the pre-calibrated high internal resistance of the device.) In the context of the latter approach, the minus sign in Eq. (2) may be described by the following *Lenz rule*: the magnetic field of the induced current I provides a partial compensation of the *change* of the original flux $\Phi(t)$ with time.[3]

In order to recast Eq. (2) in a differential form, more convenient in many cases, let us apply to the contour integral in it the Stokes theorem, which was repeatedly used in Chapter 5. The result is

$$\mathscr{V}_{\rm ind} = \int_S \left(\nabla \times \mathbf{E}_{\rm ind}\right)_n d^2 r\,. \qquad (6.3)$$

Now combining Eqs. (1)-(3), for a contour C whose shape does not change in time (so that the integration along it is interchangeable with the time derivative), we get

$$\int_S \left(\nabla \times \mathbf{E}_{\rm ind} + \frac{\partial \mathbf{B}}{\partial t}\right)_n d^2 r = 0\,. \qquad (6.4)$$

Since the induced electric field is an addition to the gradient field (1.33) created by electric charges, for the net field we may write $\mathbf{E} = \mathbf{E}_{\rm ind} - \nabla\phi$. However, since the curl of any gradient field is zero,[4] $\nabla\times(\nabla\phi) = 0$, Eq. (4) remains valid even for the net field $\mathbf{E}$. Since this equation should be correct for *any* closed area S, we may conclude that

$$\nabla \times \mathbf{E} + \frac{\partial \mathbf{B}}{\partial t} = 0 \qquad (6.5)$$

Faraday law: differential form

at any point. This is the final (time-dependent) form of this Maxwell equation. Superficially, it may look that Eq. (5) is less general than Eq. (2); for example, it does not describe any electric field, and hence any e.m.f. in a moving loop, if the field $\mathbf{B}$ is constant in time, even if the magnetic flux (1) through the loop does change in time. However, this is not true; in Chapter 9 we will see that in the reference frame moving with the loop, the e.m.f. does appear.[5]

---

[2] Such induced current is sometimes called the *eddy current*, though most often this term is reserved for the distributed currents induced by changing magnetic fields in bulk conductors – see Sec. 3 below.

[3] Let me also hope that the reader is familiar with the paradox arising at attempts to measure $\mathscr{V}_{\rm ind}$ with a voltmeter without its insertion into the wire loop; if not, I would highly recommend them to solve the offered Problem 2.

[4] See, e.g., MA Eq. (11.1).

[5] I have to admit that from the beginning of the course, I was carefully sweeping under the rug a very important question: in what exactly reference frame(s) all the equations of electrodynamics are valid? I promise to discuss this issue in detail later in the course (in Chapter 9), and for now would like to get away with a very short answer: all the formulas discussed so far are valid in *any inertial* reference frame, as defined in classical mechanics – see, e.g., CM Sec. 1.3; however, the fields $\mathbf{E}$ and $\mathbf{B}$ have to be measured *in the same* frame.

Now let us reformulate Eq. (5) in terms of the vector potential **A**. Since the induction effect does not alter the fundamental relation $\nabla \cdot \mathbf{B} = 0$, we still may represent the magnetic field as prescribed by Eq. (5.27), i.e. as $\mathbf{B} = \nabla \times \mathbf{A}$. Plugging this expression into Eq. (5), and changing the order of the temporal and spatial differentiation, we get

$$\nabla \times \left(\mathbf{E} + \frac{\partial \mathbf{A}}{\partial t} \right) = 0 . \tag{6.6}$$

Hence we can use the same argumentation as in Sec. 1.3 (there applied to the vector **E** alone) to represent the expression in the parentheses as $-\nabla\phi$, so we get

Fields via potentials

$$\mathbf{E} = -\frac{\partial \mathbf{A}}{\partial t} - \nabla\phi, \qquad \mathbf{B} = \nabla \times \mathbf{A} . \tag{6.7}$$

It is very tempting to interpret the first term of the right-hand side of the expression for **E** as the one describing the electromagnetic induction alone, and the second term as representing a purely electrostatic field induced by electric charges. However, the separation of these two terms is, to a certain extent, conditional. Indeed, let us consider the gauge transformation already mentioned in Sec. 5.2,

$$\mathbf{A} \to \mathbf{A} + \nabla\chi , \tag{6.8}$$

that, as we already know, does not change the magnetic field. According to Eq. (7), to keep the full electric field intact (*gauge-invariant*) as well, the scalar electric potential has to be transformed simultaneously, as

$$\phi \to \phi - \frac{\partial \chi}{\partial t} , \tag{6.9}$$

leaving the choice of an addition to ϕ restricted only by the Laplace equation – since the full ϕ should satisfy the Poisson equation (1.41) with a gauge-invariant right-hand side. We will return to the discussion of the gauge invariance in Sec. 4.

6.2. Magnetic energy revisited

Now we are sufficiently equipped to revisit the issue of magnetic energy, in particular, to finally prove Eqs. (5.57) and (5.140), and discuss the dichotomy of the signs in Eqs. (5.53) and (5.54). For that, let us consider a sufficiently slow and small magnetic field variation $\delta\mathbf{B}$. If we want to neglect the kinetic energy of the system of electric currents under consideration, as well as the wave radiation effects, we need to prevent its significant acceleration by the arising induction field $\mathbf{E}_{\text{ind}}$. Let us suppose that we do this by virtual balancing of this field by an external electric field $\mathbf{E}_{\text{ext}} = -\mathbf{E}_{\text{ind}}$. According to Eq. (4.38), the work of that field[6] on the stand-alone currents of the system during a small time interval δt, and hence the change of the potential energy of the system, is

$$\delta U = \delta t \int_V \mathbf{j} \cdot \mathbf{E}_{\text{ext}} d^3 r, \qquad \text{so that } \delta U = -\delta t \int_V \mathbf{j} \cdot \mathbf{E}_{\text{ind}} d^3 r , \tag{6.10}$$

[6] As a reminder, the magnetic component of the Lorentz force (5.10), $\mathbf{v}\times\mathbf{B}$, is always perpendicular to the particle velocity **v**, so the magnetic field **B** itself cannot perform any work on moving charges, i.e. on currents.

where the integral is over the volume of the system. Now expressing the current density **j** from the macroscopic Maxwell equation (5.107), $\mathbf{j} = \nabla \times \mathbf{H}$, and then applying the vector algebra identity[7]

$$(\nabla \times \mathbf{H}) \cdot \mathbf{E}_{\text{ind}} \equiv \mathbf{H} \cdot (\nabla \times \mathbf{E}_{\text{ind}}) - \nabla \cdot (\mathbf{E}_{\text{ind}} \times \mathbf{H}), \tag{6.11}$$

we get

$$\delta U = -\delta t \int_V \mathbf{H} \cdot (\nabla \times \mathbf{E}) d^3 r + \delta t \int_V \nabla \cdot (\mathbf{E} \times \mathbf{H}) d^3 r . \tag{6.12}$$

According to the divergence theorem, the second integral in the right-hand of this equality is equal to the flux of the so-called *Poynting vector* $\mathbf{S} \equiv \mathbf{E} \times \mathbf{H}$ through the surface limiting the considered volume *V*. Later in the course we will see that this flux represents, in particular, the power of electromagnetic radiation through the surface. If such radiation is negligible (as it always is if the field variation is sufficiently slow), the surface may be selected sufficiently far, so that the flux of **S** vanishes. In this case, we may express $\nabla \times \mathbf{E}$ from the Faraday induction law (5) to get

$$\delta U = -\delta t \int_V \left(-\frac{\partial B}{\partial t} \right) \cdot \mathbf{H} d^3 r = \int_V \mathbf{H} \cdot \delta \mathbf{B} d^3 r . \tag{6.13}$$

Just as in the electrostatics (see Eqs. (1.65) and (3.73), and their discussion), this relation may be interpreted as the variation of the magnetic field energy *U* of the system, and represented in the form

$$\boxed{\delta U = \int_V \delta u(\mathbf{r}) d^3 r, \qquad \text{with } \delta u \equiv \mathbf{H} \cdot \delta \mathbf{B} .} \tag{6.14}$$

Magnetic energy's variation

This is a keystone result; let us discuss it in some detail.

First of all, for a system filled with a linear and isotropic magnetic material, we may use Eq. (14) together with Eq. (5.110): $\mathbf{B} = \mu \mathbf{H}$. Integrating the result over the variation of the field from 0 to a certain final value **B**, we get Eq. (5.140) – so important that it deserves rewriting again:

$$U = \int_V u(\mathbf{r}) d^3 r, \quad \text{with } u = \frac{B^2}{2\mu} . \tag{6.15}$$

In the simplest case of free space (no magnetics at all, so **j** above is the complete current density), we may take $\mu = \mu_0$, and reduce Eq. (15) to Eq. (5.57). Now performing backward the transformations that took us, in Sec. 5.3, to derive that relation from Eq. (5.54), we finally have the latter formula proved – as was promised in the last chapter.

It is very important, however, to understand the limitations of Eq. (15). For example, let us try to apply it to a very simple problem, which was already analyzed in Sec. 5.6 (see Fig. 5.15): a very long cylindrical sample of a linear magnetic material placed into a fixed external field $\mathbf{H}_{\text{ext}}$ parallel to the sample's length. It is evident that in this simple geometry, the field **H** and hence the field $\mathbf{B} = \mu \mathbf{H}$ have to be uniform inside the sample, besides negligible regions near its ends, so Eq. (15) is reduced to

$$U = \frac{B^2}{2\mu} V , \tag{6.16}$$

[7] See, e.g., MA Eq. (11.7) with $\mathbf{f} = \mathbf{E}_{\text{ind}}$ and $\mathbf{g} = \mathbf{H}$.

where $V = Al$ is the cylinder's volume. Now if we try to calculate the static (equilibrium) value of the field from the minimum of this potential energy, we get evident nonsense: $\mathbf{B} = 0$ (WRONG!).[8]

The situation may be readily rectified by using the notion of the Gibbs potential energy, just as it was done for the electric field in Sec. 3.5 (and implicitly in the end of Sec. 1.3). According to Eq. (14), in magnetostatics, the Cartesian components of the field $\mathbf{H}(\mathbf{r})$ play the role of the generalized forces, while those of the field $\mathbf{B}(\mathbf{r})$, of the generalized coordinates (per unit volume).[9] As the result, the Gibbs potential energy, whose minimum corresponds to the stable equilibrium of the system under the effect of a fixed generalized force (in our current case, of the fixed external field $\mathbf{H}_{\text{ext}}$), is

Gibbs potential energy

$$U_{\text{G}} = \int_V u_{\text{G}}(\mathbf{r})\, d^3 r, \qquad \text{with } u_{\text{G}}(\mathbf{r}) \equiv u(\mathbf{r}) - \mathbf{H}_{\text{ext}}(\mathbf{r}) \cdot \mathbf{B}(\mathbf{r}), \tag{6.17}$$

– the expression parallel to Eq. (3.78). For a system with linear magnetics, we may use, for the energy density $u(\mathbf{r})$, our result (15), getting the following Gibbs energy's density:

$$u_{\text{G}}(\mathbf{r}) = \frac{1}{2\mu}\mathbf{B}\cdot\mathbf{B} - \mathbf{H}_{\text{ext}}\cdot\mathbf{B} \equiv \frac{1}{2\mu}\left(\mathbf{B} - \mu\mathbf{H}_{\text{ext}}\right)^2 + \text{const}, \tag{6.18}$$

where "const" means a term independent of the field $\mathbf{B}$ inside the sample. For our simple cylindrical system, with its uniform fields, Eqs. (17)-(18) gives the following full Gibbs energy of the sample:

$$U_{\text{G}} = \frac{\left(\mathbf{B}_{\text{int}} - \mu\mathbf{H}_{\text{ext}}\right)^2}{2\mu} V + \text{const}, \tag{6.19}$$

whose minimum immediately gives the correct stationary value $\mathbf{B}_{\text{int}} = \mu\mathbf{H}_{\text{ext}}$, i.e. $\mathbf{H}_{\text{int}} \equiv \mathbf{B}_{\text{int}}/\mu = \mathbf{H}_{\text{ext}}$, which was already obtained in Sec. 5.6 in a different way, from the boundary condition (5.117).

Now notice that with this result on hand, Eq. (18) may be rewritten in a different form:

$$u_{\text{G}}(\mathbf{r}) = \frac{1}{2\mu}\mathbf{B}\cdot\mathbf{B} - \frac{\mathbf{B}}{\mu}\cdot\mathbf{B} \equiv -\frac{B^2}{2\mu}, \tag{6.20}$$

similar to Eq. (15) for $u(\mathbf{r})$, but with an opposite sign. This sign dichotomy explains that of Eqs. (5.53) and Eq. (5.54); indeed, as was already noted in Sec. 5.3, the former of these expressions gives the potential energy whose minimum corresponds to the equilibrium of a system with fixed currents. (In our current example, these are the external stand-alone currents inducing the field $\mathbf{H}_{\text{ext}}$.) So, the energy U_j given by Eq. (5.53) is essentially the Gibbs energy U_{G} defined by Eqs. (17) and (for the equilibrium state of linear magnetic media) by Eq. (20), while Eq. (5.54) is just another form of Eq. (15) – as was explicitly shown in Sec. 5.3.[10]

---

[8] This erroneous result cannot be corrected by just adding the energy of the field outside the cylinder because in the limit $A \to 0$, this field is not affected by the internal field $\mathbf{B}$.

[9] Note an aspect in that the analogy with electrostatics is not quite complete. Indeed, according to Eq. (3.76), in electrostatics, the role of a generalized coordinate is played by the "would-be" field $\mathbf{D}$, and that of the generalized force, by the actual (if macroscopic) electric field $\mathbf{E}$. This difference may be traced back to the fact that the electric field $\mathbf{E}$ may perform work on a moving charged particle, while the magnetic field cannot. However, this difference does not affect the full analogy of the expressions (3.73) and (15) for the field energy density in *linear* media.

[10] As was already noted in Sec. 5.4, one more example of the energy U_j (i.e. U_{G}) is given by Eq. (5.100).

Let me complete this section by stating that the difference between the energies U and U_G is not properly emphasized (or even left obscure) in some textbooks, so the reader is advised to get additional clarity by solving a few additional simple problems – for example, by spelling out these energies for a long straight solenoid (Fig. 5.6a), and then using the results to calculate the pressure exerted by the magnetic field on the solenoid's walls (windings) and the longitudinal forces exerted on its ends.

6.3. Quasistatic approximation and skin effect

Perhaps the most surprising experimental fact concerning the time-dependent electromagnetic phenomena is that unless they are so fast that one more new effect of the *displacement currents* (to be discussed in Sec. 7 below) becomes noticeable, all formulas of electrostatics and magnetostatics remain valid, with the only exception: the generalization of Eq. (3.36) to Eq. (5), describing the Faraday induction. As a result, the system of macroscopic Maxwell equations (5.109) is generalized to

$$\nabla\times\mathbf{E}+\frac{\partial\mathbf{B}}{\partial t}=0,\qquad \nabla\times\mathbf{H}=\mathbf{j},$$
$$\nabla\cdot\mathbf{D}=\rho,\qquad \nabla\cdot\mathbf{B}=0. \tag{6.21}$$

Quasistatic approximation

(As it follows from the discussions in chapters 3 and 5, the corresponding system of microscopic Maxwell equations for the genuine, "microscopic" fields **E** and **B** may be obtained from Eq. (21) by the formal substitutions $\mathbf{D}=\varepsilon_0\mathbf{E}$ and $\mathbf{H}=\mathbf{B}/\mu_0$, and the replacement of the stand-alone charge and current densities ρ and **j** with their full densities.[11]) These equations, whose range of validity will be quantified in Sec. 7, define the so-called *quasistatic approximation* of electromagnetism and are sufficient for an adequate description of a broad range of physical effects.

In order to form a complete system of equations, Eqs. (21) should be augmented by constituent equations describing the medium under consideration. For a linear isotropic material, they may be taken in the simplest (and simultaneously, most common) linear and isotropic forms already discussed in Chapters 4 and 5:

$$\mathbf{j}=\sigma\mathbf{E},\qquad \mathbf{B}=\mu\mathbf{H}. \tag{6.22}$$

If the conductor is uniform, i.e. the coefficients σ and μ are constant inside it, the whole system of Eqs. (21)-(22) may be reduced to just one simple equation. Indeed, a sequential substitution of these equations into each other, using a well-known vector-algebra identity[12] in the middle, yields:

$$\frac{\partial\mathbf{B}}{\partial t}=-\nabla\times\mathbf{E}=-\frac{1}{\sigma}\nabla\times\mathbf{j}=-\frac{1}{\sigma}\nabla\times(\nabla\times\mathbf{H})=-\frac{1}{\sigma\mu}\nabla\times(\nabla\times\mathbf{B})\equiv-\frac{1}{\sigma\mu}\left[\nabla(\nabla\cdot\mathbf{B})-\nabla^2\mathbf{B}\right]$$
$$=\frac{1}{\sigma\mu}\nabla^2\mathbf{B}. \tag{6.23}$$

Thus we have arrived, without any further assumptions, at a rather simple partial differential equation. Let us use it for an analysis of the so-called *skin effect*, the phenomenon of an Ohmic conductor's self-shielding from the alternating (*ac*) magnetic field. In its simplest geometry (Fig. 2a), an

[11] Obviously, in free space, the last replacement is unnecessary, because all charges and currents may be treated as "stand-alone" ones.

[12] See, e.g., MA Eq. (11.3).

external source (which, at this point, does not need to be specified) produces, near a plane surface of a bulk conductor, a spatially-uniform ac magnetic field $\mathbf{H}^{(0)}(t)$ parallel to the surface.[13]

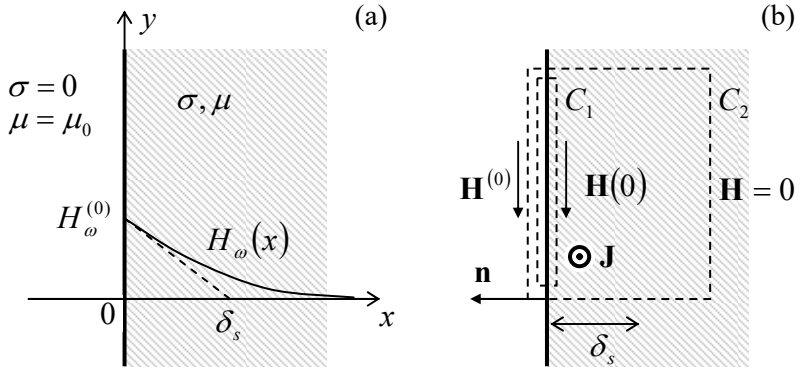


Fig. 6.2. (a) The skin effect in the simplest, planar geometry, and (b) two Ampère contours, C_1 and C_2, for deriving the "macroscopic" (C_1) and the "coarse-grain" (C_2) boundary conditions for **H**.

Selecting the coordinate system as shown in Fig. 2a, we may express this condition as

$$\mathbf{H}\Big|_{x=-0} = H^{(0)}(t)\mathbf{n}_y \,. \tag{6.24}$$

The translational symmetry of our simple problem within the surface plane [y, z] implies that inside the conductor, $\partial/\partial y = \partial/\partial z = 0$ as well, and $\mathbf{H} = H(x, t)\mathbf{n}_y$ even at $x \geq 0$, so Eq. (23) for the conductor's interior is reduced to a differential equation for just one scalar function $H(x, t) = B(x, t)/\mu$:

$$\frac{\partial H}{\partial t} = \frac{1}{\sigma\mu}\frac{\partial^2 H}{\partial x^2}, \qquad \text{for } x \geq 0\,. \tag{6.25}$$

This equation may be further simplified by noticing that due to its linearity, we may use the linear superposition principle for the time dependence of the field,[14] via expanding it, as well as the external field (24), into the Fourier series:

$$\begin{aligned} H(x,t) &= \sum_{\omega} H_{\omega}(x)e^{-i\omega t}, \quad \text{for } x. \geq 0, \\ H^{(0)}(t) &= \sum_{\omega} H_{\omega}^{(0)}e^{-i\omega t}, \qquad \text{for } x = -0, \end{aligned} \tag{6.26}$$

and arguing that if we know the solution for each frequency component of the series, the whole field may be found through the straightforward summation (26) of these solutions.

For each single-frequency component, Eq. (25) is immediately reduced to an ordinary differential equation for the complex amplitude $H_\omega(x)$:[15]

[13] Due to the simple linear relation $\mathbf{B} = \mu\mathbf{H}$ between the fields **B** and **H**, it does not matter too much which of them is used for the solution of this problem, with a slight preference for **H**, due to the simplicity of Eq. (5.117) – the only boundary condition relevant for this simple geometry.

[14] Another way to exploit the linearity of Eq. (6.25) is to use the *spatial-temporal Green's function* approach to explore the dependence of its solutions on various initial conditions. Unfortunately, because of a lack of time, I have to leave an analysis of this opportunity for the reader's exercise.

[15] Let me hope that the reader is not intimidated by the (very convenient) use of such complex variables for describing real fields; their imaginary parts always disappear at the final summation (26). For example, if the

$$-i\omega H_\omega = \frac{1}{\sigma\mu}\frac{d^2}{dx^2}H_\omega . \tag{6.27}$$

From the theory of linear ordinary differential equations, we know that Eq. (27) has the following general solution:

$$H_\omega(x) = H_+ e^{\kappa_+ x} + H_- e^{\kappa_- x}, \tag{6.28}$$

where the constants $\kappa_\pm$ are the roots of the characteristic equation that may be obtained by the substitution of any of these two exponents into the initial differential equation. For our particular case, the characteristic equation following from Eq. (27) is simply

$$-i\omega = \frac{\kappa^2}{\sigma\mu} \tag{6.29}$$

and its roots are, obviously,

$$\kappa_\pm = \left(-i\mu\omega\sigma\right)^{1/2} \equiv \pm\frac{1-i}{\sqrt{2}}\left(\mu\omega\sigma\right)^{1/2}. \tag{6.30}$$

For our problem, the field cannot grow exponentially at $x \to +\infty$, so only one of the coefficients, namely the H_- corresponding to the decaying exponent, with Re $\kappa_- < 0$, may be different from zero, i.e. $H_\omega(x) = H_\omega(0)\exp\{\kappa_- x\}$. To find the constant factor $H_\omega(0)$, we can integrate the macroscopic Maxwell equation $\nabla \times \mathbf{H} = \mathbf{j}$ along a pre-surface contour – say, the contour C_1 shown in Fig. 2b. The right-hand side's integral is negligible because the stand-alone current density $\mathbf{j}$ does not include the "genuinely-surface" currents responsible for the magnetic permeability μ – see Fig. 5.12. As a result, we get the boundary condition similar to Eq. (5.117) for the stationary magnetic field: H_τ = const at $x = 0$, giving us

$$H(0,t) = H^{(0)}(t), \qquad \text{i.e. } H_\omega(0) = H_\omega^{(0)}, \tag{6.31}$$

so the final solution of our boundary problem may be represented as

$$H_\omega(x) = H_\omega^{(0)}\exp\{\kappa_- x\} = H_\omega^{(0)}\exp\left\{-\frac{x}{\delta_s}\right\}\exp\left\{-i\left(\omega t - \frac{x}{\delta_s}\right)\right\}, \tag{6.32}$$

where the constant δ_s, with the dimension of length, is called the *skin depth*:

$$\delta_s \equiv -\frac{1}{\mathrm{Re}\,\kappa_-} = \left(\frac{2}{\mu\sigma\omega}\right)^{1/2}. \tag{6.33}$$

Skin depth

This solution describes the *skin effect*: the penetration of the ac magnetic field, and the eddy currents $\mathbf{j}$, into a conductor only to a finite depth of the order of δ_s. Let me give a few numerical examples of this depth: for copper at room temperature, $\delta_s \approx 1$ cm at the usual ac power distribution frequency of 60 Hz, and is of the order of just 1 μm at a few GHz, i.e. at typical frequencies of cell phone signals and kitchen microwave magnetrons. On the other hand, for lightly salted water, δ_s is close to 250 m at just 1 Hz (with significant implications for radio communications with submarines), and of

external field is purely sinusoidal, with the actual (positive) frequency ω, each sum in Eq. (26) has just two terms, with complex amplitudes H_ω and $H_{-\omega} = H_\omega^*$, so their sum is always real. (For a more detailed discussion of this issue, see, e.g., CM Sec. 5.1.)

the order of 1 cm at a few GHz (explaining, in particular, the nonuniform heating of a soup bowl in a microwave oven).[16]

Let me hope that the equality chain (23) makes the physics of this effect very clear: the external electric field **E**, which is Faraday-induced by an external ac magnetic field, drives the eddy currents **j**, which in turn induce their own magnetic field that eventually (at $x \sim \delta_s$) compensates the external one. Let us quantify these **E** and **j**. Since we have used, in particular, relations $\mathbf{j} = \nabla \times \mathbf{H} = \nabla \times \mathbf{B}/\mu$, and $\mathbf{E} = \mathbf{j}/\sigma$, and spatial differentiation of an exponent yields a similar exponent, the electric field and current density have the same spatial dependence as the magnetic field, i.e. penetrate the conductor only by distances of the order of $\delta_s(\omega)$. Their vectors are directed normally to **B**, while still being parallel to the conductor's surface:[17]

$$\mathbf{j}_\omega(x) = \kappa_- H_\omega(x)\,\mathbf{n}_z, \qquad \mathbf{E}_\omega(x) = \frac{\kappa_-}{\sigma} H_\omega(x)\,\mathbf{n}_z. \tag{6.34}$$

We may use these expressions, in particular, to calculate the time-averaged power density (4.39) of the energy dissipation, for the important case of a sinusoidal ("monochromatic") field $H(x, t) = |H_\omega(x)| \cos(\omega t + \varphi)$, and hence sinusoidal eddy currents: $j(x, t) = |j_\omega(x)| \cos(\omega t + \varphi')$:

$$\bar{\mathcal{P}}(x) = \frac{\overline{j^2(x,t)}}{\sigma} = \frac{|j_\omega(x)|^2 \overline{\cos^2(\omega t + \varphi')}}{\sigma} = \frac{|j_\omega(x)|^2}{2\sigma} = \frac{|\kappa_-|^2 |H_\omega(x)|^2}{2\sigma} = \frac{|H_\omega(x)|^2}{\delta_s^2 \sigma}. \tag{6.35}$$

Now the (elementary) integration of this expression along the *x*-axis (through all the skin depth), using the exponential law (6.32), gives us the following average power of the energy loss per unit area:

Energy loss at skin effect

$$\frac{d\bar{\mathcal{P}}}{dA} \equiv \int_0^\infty \bar{\mathcal{P}}(x)dx = \frac{1}{2\delta_s \sigma}\left|H_\omega^{(0)}\right|^2 \equiv \frac{\mu\omega\delta_s}{4}\left|H_\omega^{(0)}\right|^2. \tag{6.36}$$

We will extensively use this expression in the next chapter to calculate the energy losses in microwave waveguides and resonators with conducting (practically, metallic) walls, and for now let me note only that according to Eqs. (33) and (36), for a fixed magnetic field amplitude, the losses grow with frequency as $\omega^{1/2}$.

One more important remark concerning Eqs. (34): integrating the first of them over *x*, with the help of Eq. (32), we may see that the *linear density* **J** of the surface currents (measured in A/m), is simply and fundamentally related to the applied magnetic field:

$$\mathbf{J}_\omega \equiv \int_0^\infty \mathbf{j}_\omega(x)dx = H_\omega^{(0)}\mathbf{n}_z. \tag{6.37}$$

Since this relation does not have any frequency-dependent factors, we may sum it up for all frequency components, and get a universal relation

$$\mathbf{J}(t) = H^{(0)}(t)\mathbf{n}_z \equiv H^{(0)}(t)(-\mathbf{n}_y \times \mathbf{n}_x) = \mathbf{H}^{(0)}(t) \times (-\mathbf{n}_x) = \mathbf{H}^{(0)}(t) \times \mathbf{n}, \tag{6.38a}$$

[16] Let me hope that the reader's physical intuition makes it evident that the skin effect remains conceptually the same for samples of any *shape*, besides possibly some quantitative details of the field distribution.

[17] The loop (vortex) character of the induced current lines, responsible for the term "eddy", is not very apparent in the 1D geometry explored above, with the near-surface currents (Fig. 2b) looping only implicitly, at $z \to \pm\infty$.

(where $\mathbf{n} = -\mathbf{n}_x$ is the outer normal to the surface – see Fig. 2b) or, in a different form,

$$\Delta\mathbf{H}(t) = \mathbf{n} \times \mathbf{J}(t), \tag{6.38b}$$

Coarse-grain boundary relation

where $\Delta\mathbf{H}$ is the full change of the field through the skin layer. This simple *coarse-grain relation* (independent of the choice of coordinate axes), is also independent of the used constituent relations (22), and is by no means occasional. Indeed, it may be readily obtained from the macroscopic Ampère law (5.116), by applying it to a contour drawn around a fragment of the surface, extending under it substantially deeper than the skin depth – see the contour C_2 in Fig. 2b. Hence, Eq. (38) is valid regardless of the exact law of the field penetration.

For the skin effect, this fundamental relationship between the linear current density and the external magnetic field implies that the skin effect's implementation does not necessarily require a dedicated ac magnetic field source. For example, the effect takes place in any wire that carries an ac current, leading to a current's concentration in a surface sheet of thickness $\sim\delta_s$. (Of course, the quantitative analysis of this problem in a wire with an arbitrary cross-section may be technically complicated, because it requires solving Eq. (23) for the corresponding 2D geometry; even for the round cross-section, the solution involves the Bessel functions.) In this case, the ac magnetic field outside the conductor, which still obeys Eq. (38), may be better interpreted as the effect, rather than the cause, of the ac current flow.

Finally, please mind the limited validity of all the above results. First, for the quasistatic approximation to be valid, the field frequency ω should not be too high, so the displacement current effects are negligible. (Again, this condition will be quantified in Sec. 7 below; it will show that for metals, the condition is violated only at extremely high frequencies above $\sim 10^{18}$ s^{-1}.) A more practical upper limit on ω is that the skin depth δ_s should stay much larger than the mean free path l of charge carriers,[18] because beyond this point, the constituent relation between the vectors $\mathbf{j}(\mathbf{r})$ and $\mathbf{E}(\mathbf{r})$ becomes essentially *non-local*. Both theory and experiment show that at δ_s below l, the skin effect persists, but acquires a frequency dependence slightly different from Eq. (33): $\delta_s \propto \omega^{-1/3}$ rather than $\omega^{-1/2}$. Historically, this *anomalous skin effect* has been very useful for the measurements of the Fermi surfaces of metals.[19]

6.4. Electrodynamics of superconductivity, and the gauge invariance

The effect of superconductivity[20] takes place (in certain materials only, mostly metals) when temperature T is reduced below a certain *critical temperature* T_c specific for each material. For most metallic superconductors, T_c is of the order of typically a few kelvins, though several compounds (the so-called *high-temperature superconductors*) with T_c above 100 K have been found since 1987. The most notable property of superconductors is the absence, at $T < T_c$, of measurable resistance to (not very high) dc currents. However, the electromagnetic properties of superconductors cannot be described by just taking $\sigma = \infty$ in our previous results. Indeed, for this case, Eq. (33) would give $\delta_s = 0$, i.e., no ac

[18] A discussion of the mean free path may be found, for example, in SM Chapter 6. In very clean metals at very low temperatures, δ_s may approach l at frequencies as low as ~1 GHz, but at room temperature, the crossover between the normal to the anomalous skin effect takes place only at ~ 100 GHz.

[19] See, e.g., A. Abrikosov, *Introduction to the Theory of Normal Metals*, Academic Press, 1972.

[20] Discovered experimentally in 1911 by Heike Kamerlingh Onnes.

magnetic field penetration at all. Experiment shows something substantially different: weak magnetic fields do penetrate into superconductors by a material-specific distance $\delta_L \sim 10^{-7}$-10^{-6} m, the so-called *London's penetration depth*,[21] which is virtually frequency-independent until the skin depth δ_s, of the same material in its "normal" state, i.e. the absence of superconductivity, becomes less than δ_L. (This crossover happens typically at frequencies $\omega \sim 10^{13}$-10^{14} s^{-1}.) The smallness of δ_L on the human scale means that the magnetic field is pushed out from macroscopic samples at their transition into the superconducting state.

This *Meissner-Ochsenfeld effect*, discovered experimentally in 1933,[22] may be partly understood using the following classical reasoning. Our discussion of the Ohm law in Sec. 4.2 implied that the current's (and hence the electric field's) frequency ω is either zero or sufficiently low. In the classical Drude reasoning, this is acceptable while $\omega\tau << 1$, where τ is the effective carrier scattering time participating in Eqs. (4.12)-(4.13). If this condition is not satisfied, we should take into account the charge carrier inertia; moreover, in the opposite limit $\omega\tau >> 1$, we may neglect the scattering at all. Classically, we can describe the charge carriers in such a "perfect conductor" as particles with a non-zero mass m, which are accelerated by the electric field following the 2nd Newton law (4.11),

$$m\dot{\mathbf{v}} = \mathbf{F} = q\mathbf{E}, \tag{6.39}$$

so the current density $\mathbf{j} = qn\mathbf{v}$ that they create, changes in time as

$$\dot{\mathbf{j}} = qn\dot{\mathbf{v}} = \frac{q^2 n}{m}\mathbf{E}. \tag{6.40}$$

In terms of the Fourier amplitudes of the functions $\mathbf{j}(t)$ and $\mathbf{E}(t)$, this means

$$-i\omega \mathbf{j}_\omega = \frac{q^2 n}{m}\mathbf{E}_\omega. \tag{6.41}$$

Comparing this formula with the relation $\mathbf{j}_\omega = \sigma\mathbf{E}_\omega$ implied in the last section, we see that we can use all its results with the following replacement:

$$\sigma \to i\frac{q^2 n}{m\omega}. \tag{6.42}$$

This change replaces the characteristic equation (29) with

$$-i\omega = \frac{\kappa^2 m\omega}{iq^2 n\mu}, \qquad \text{i.e. } \kappa^2 = \frac{\mu q^2 n}{m}, \tag{6.43}$$

i.e. replaces the skin effect with the field penetration by the following frequency-independent depth:

$$\delta \equiv \frac{1}{\kappa} = \left(\frac{m}{\mu q^2 n}\right)^{1/2}. \tag{6.44}$$

Superficially, this means that the field decay into the superconductor does not depend on frequency:

[21] Named so to acknowledge the pioneering theoretical work of brothers Fritz and Heinz London – see below.

[22] It is hardly fair to shorten this name to just the "Meissner effect" as it is frequently done, because of the reportedly crucial contribution by Robert Ochsenfeld, then a Walther Meissner's student, to the discovery.

$$H(x,t) = H(0,t)e^{-x/\delta}, \tag{6.45}$$

thus explaining the Meissner-Ochsenfeld effect.

However, there are two problems with this result. First, for the parameters typical for good metals ($q = -e$, $n \sim 10^{29}$ m^{-3}, $m \sim m_e$, $\mu \approx \mu_0$), Eq. (44) gives $\delta \sim 10^{-8}$ m, one or two orders of magnitude lower than the experimental values of δ_L. Experiment also shows that the penetration depth diverges at $T \to T_c$, which is not predicted by Eq. (44).

The second, much more fundamental problem with Eq. (44) is that it has been derived for $\omega\tau >> 1$. Even if we assume that somehow there is no scattering at all, i.e. $\tau = \infty$, at $\omega \to 0$ both parts of the characteristic equation (43) vanish, and we cannot make any conclusion about κ. This is not just a mathematical artifact we could ignore. For example, let us place a non-magnetic metal into a static external magnetic field at $T > T_c$. The field would completely penetrate the sample. Now let us cool it. As soon as the temperature is decreased below T_c, the above calculations would become valid, forbidding the penetration into the superconductor of any *change* of the field, so the initial field would be "frozen" inside the sample. The Meissner-Ochsenfeld experiments have shown something completely different: as T is lowered below T_c, the initial field is being expelled out of the sample.

The resolution of these contradictions is provided by quantum mechanics. As was explained in 1957 in a seminal work by J. Bardeen, L. Cooper, and J. Schrieffer (commonly referred to as the *BCS theory*), superconductivity is due to the correlated motion of electron pairs, with opposite spins and nearly opposite momenta. Such *Cooper pairs*, each with the electric charge $q = -2e$ and zero spin, may form only in a narrow energy layer near the Fermi surface, of a certain thickness $\Delta(T)$. This parameter $\Delta(T)$, which may be also interpreted as the binding energy of the pair, tends to zero at $T \to T_c$, while at $T << T_c$ it has a virtually constant value $\Delta(0) \approx 3.5\, k_B T_c$, of the order of a few meV for most superconductors. This fact readily explains the relatively low spatial density of the Cooper pairs: $n_p \sim n\Delta(T)/\varepsilon_F \sim 10^{26}$ m^{-3}. With the correction $n \to n_p$, Eq. (44) for the penetration depth becomes

$$\delta \to \delta_L = \left(\frac{m}{\mu q^2 n_p}\right)^{1/2}. \tag{6.46}$$

London's penetration depth

This result diverges at $T \to T_c$, and generally fits the experimental data reasonably well, at least for the so-called "clean" superconductors with the mean free path $l = v_F\tau$ (where $v_F \sim (2m\varepsilon_F)^{1/2}$ is the r.m.s. velocity of electrons on the Fermi surface) much longer than the Cooper pair size ξ – see below.

The smallness of the coupling energy $\Delta(T)$ is also a key factor in the explanation of the Meissner-Ochsenfeld effect. Because of Heisenberg's quantum uncertainty relation $\delta r \delta p \sim \hbar$, the spatial extension of the Cooper-pair's wavefunction (the so-called *coherence length* of the superconductor) is relatively large: $\xi \sim \delta r \sim \hbar/\delta p \sim \hbar v_F/\Delta(T) \sim 10^{-6}$ m. As a result, $n_p\xi^3 >> 1$, meaning that the wavefunctions of the pairs are strongly overlapped in space. Due to their integer spin, Cooper pairs behave like bosons, which means in particular that at low temperatures they exhibit the so-called *Bose-Einstein condensation* onto the same ground energy level ε_g.[23] This means that the quantum frequency ω

[23] A quantitative discussion of the Bose-Einstein condensation of bosons may be found in SM Sec. 3.4, though the full theory of superconductivity is more complicated because it has to describe the condensation taking place *simultaneously* with the formation of effective bosons (Cooper pairs) from fermions (single electrons). For a

$= \varepsilon_g/\hbar$ of the time evolution of each pair's wavefunction $\Psi = \psi\exp\{-i\omega t\}$ is exactly the same and that the phases φ of the wavefunctions, defined by the relation

$$\psi = |\psi| e^{i\varphi}, \tag{6.47}$$

coincide, so the electric current is carried not by individual Cooper pairs but rather by their *Bose-Einstein condensate* described by a single wavefunction (47). Due to this coherence, the quantum effects (which are, in the usual Fermi-gases of single electrons, masked by the statistical spread of their energies, and hence of their phases), become very explicit – "macroscopic".

To illustrate this, let us write the well-known quantum-mechanical formula for the probability current density of a free, non-relativistic particle,[24]

$$\mathbf{j}_w = \frac{i\hbar}{2m}\left(\psi\nabla\psi^* - \text{c.c.}\right) \equiv \frac{1}{2m}\left[\psi^*\left(-i\hbar\nabla\right)\psi - \text{c.c.}\right], \tag{6.48}$$

where c.c. means the complex conjugate of the previous expression. Now let me borrow one result that will be proved later in this course (in Sec. 9.7) when we discuss the analytical mechanics of a charged particle moving in an electromagnetic field. Namely, to account for the magnetic field effects, the particle's *kinetic momentum* $\mathbf{p} \equiv m\mathbf{v}$ (where $\mathbf{v} \equiv d\mathbf{r}/dt$ is the particle's velocity) has to be distinguished from its *canonical momentum*,[25]

$$\mathbf{P} \equiv \mathbf{p} + q\mathbf{A}. \tag{6.49}$$

where $\mathbf{A}$ is the field's vector potential defined by Eq. (5.27). In contrast with the Cartesian components $p_j = mv_j$ of the kinetic momentum $\mathbf{p}$, the canonical momentum's components are the generalized momenta corresponding to the Cartesian components r_j of the radius-vector $\mathbf{r}$, considered as generalized coordinates of the particle: $P_j = \partial\mathscr{L}/\partial v_j$, where $\mathscr{L}$ is the particle's Lagrangian function. According to the general rules of transfer from classical to quantum mechanics,[26] it is the vector $\mathbf{P}$ whose operator (in the coordinate representation) equals $-i\hbar\nabla$, so the operator of the kinetic momentum $\mathbf{p} = \mathbf{P} - q\mathbf{A}$ is $-i\hbar\nabla + q\mathbf{A}$. Hence, to account for the magnetic field[27] effects, we should make the following replacement,

$$-i\hbar\nabla \to -i\hbar\nabla - q\mathbf{A}, \tag{6.50}$$

in all quantum-mechanical relations. In particular, Eq. (48) has to be generalized as

$$\mathbf{j}_w = \frac{1}{2m}\left[\psi^*\left(-i\hbar\nabla - q\mathbf{A}\right)\psi - \text{c.c.}\right]. \tag{6.51}$$

This expression becomes more transparent if we take the wavefunction in form (47); then

---

detailed, but still very readable coverage of the physics of superconductors, I can recommend the reader the monograph by M. Tinkham, *Introduction to Superconductivity*, 2nd ed., McGraw-Hill, 1996.

[24] See, e.g., QM Sec. 1.4, in particular Eq. (1.47).

[25] I am sorry to use traditional notations $\mathbf{p}$ and $\mathbf{P}$ for the momenta – the same symbols which were used for the electric dipole moment and polarization in Chapter 3. I hope there will be no confusion because the latter notions are not used in this section.

[26] See, e.g., CM Sec. 10.1, in particular Eq. (10.26).

[27] The account of the electric field is easier, because the related energy $q\phi$ of the particle may be directly included in the potential energy operator.

$$\mathbf{j}_w = \frac{\hbar}{m}|\psi|^2\left(\nabla\varphi - \frac{q}{\hbar}\mathbf{A}\right). \tag{6.52}$$

This relation means, in particular, that in order to keep $\mathbf{j}_w$ gauge-invariant, the transformation (8)-(9) has to be accompanied by a simultaneous transformation of the wavefunction's phase:

$$\varphi \to \varphi + \frac{q}{\hbar}\chi\,. \tag{6.53}$$

It is fascinating that the quantum-mechanical wavefunction (or more exactly, its phase) is *not* gauge-invariant, meaning that you may change it in your mind – at your free will! Again, this does not change any observable (such as $\mathbf{j}_w$ or the probability density $\psi\psi^*$), i.e. any experimental results.

Now for the *electric* current density of the whole superconducting condensate, Eq. (52) yields the following constitutive relation:

$$\mathbf{j} \equiv \mathbf{j}_w q n_{\rm p} = \frac{\hbar q n_{\rm p}}{m}|\psi|^2\left(\nabla\varphi - \frac{q}{\hbar}\mathbf{A}\right), \tag{6.54}$$

Supercurrent density

The formula shows that this *supercurrent* may be induced by the dc magnetic field alone and does not require any electric field. Indeed, for the simple 1D geometry shown in Fig. 2a, $\mathbf{j}(\mathbf{r}) = j(x)\mathbf{n}_z$, $\mathbf{A}(\mathbf{r}) = A(x)\mathbf{n}_z$, and $\partial/\partial z = 0$, so the Coulomb gauge condition (5.48) is satisfied for any choice of the gauge function $\chi(x)$. For the sake of simplicity we can choose this function to provide $\varphi(\mathbf{r}) \equiv$ const,[28] so

$$\mathbf{j} = -\frac{q^2 n_{\rm p}}{m}\mathbf{A} \equiv -\frac{1}{\mu\delta_{\rm L}^2}\mathbf{A}\,. \tag{6.55}$$

where $\delta_{\rm L}$ is given by Eq. (46), and the field is assumed to be small and hence not affecting the probability $|\psi|^2$ (here normalized to 1 in the absence of the field). This is the so-called *London equation*, proposed (in a different form) by F. and H. London in 1935 for the Meissner-Ochsenfeld effect's explanation. Combining it with Eq. (5.44), generalized for a linear magnetic medium by the replacement $\mu_0 \to \mu$, we get

$$\nabla^2\mathbf{A} = \frac{1}{\delta_{\rm L}^2}\mathbf{A}\,, \tag{6.56}$$

For our 1D geometry, this simple differential equation, similar to Eq. (23), has an exponential solution similar to Eq. (32):

$$A(x) = A(0)\exp\left\{-\frac{x}{\delta_{\rm L}}\right\}, \qquad B(x) = B(0)\exp\left\{-\frac{x}{\delta_{\rm L}}\right\}, \qquad j(x) = j(0)\exp\left\{-\frac{x}{\delta_{\rm L}}\right\}, \tag{6.57}$$

which shows that the magnetic field and the supercurrent penetrate into a superconductor only by London's penetration depth $\delta_{\rm L}$, regardless of frequency.[29] By the way, integrating the last result through the penetration layer, and using the vector potential's definition, $\mathbf{B} = \nabla\times\mathbf{A}$ (for our geometry, giving

[28] This is the so-called *London gauge*; for our simple geometry, it is also the Coulomb gauge (5.48).

[29] Since at $T > 0$, not all electrons in a superconductor form Cooper pairs, at any frequency $\omega \neq 0$ the unpaired electrons provide energy-dissipating Ohmic currents, which are not described by Eq. (54). These losses become very substantial when the frequency ω becomes so high that the skin-effect length $\delta_{\rm s}$ of the material becomes less than $\delta_{\rm L}$. For typical metallic superconductors, this crossover takes place at frequencies of a few hundred GHz, so even for microwaves, Eq. (57) still gives a fairly accurate description of the field penetration.

$B(x) = dA(x)/dx = -\delta_L A(x)$) we may readily verify that the linear density **J** of the surface supercurrent still satisfies the universal coarse-grain relation (38).

This universality should bring to our attention the following common feature of the skin effect (in "normal" conductors) and the Meissner-Ochsenfeld effect (in superconductors): if the linear size of a bulk sample is much larger than, respectively, δ_s or δ_L, than **B** = 0 in the dominating part of its interior. According to Eq. (5.110), a formal description of such conductors (valid only on a coarse-grain scale much larger than either δ_s or δ_L), may be achieved by formally treating the sample as an *ideal diamagnet*, with $\mu = 0$. In particular, we can use this description and Eq. (5.124) to immediately obtain the magnetic field's distribution outside of a bulk sphere:

$$\mathbf{B} = \mu_0 \mathbf{H} = -\mu_0 \nabla \phi_m, \qquad \text{with } \phi_m = H_0 \left(-r - \frac{R^3}{2r^2} \right) \cos\theta, \qquad \text{for } r \geq R. \tag{6.58}$$

Figure 3 shows the corresponding surfaces of equal potential ϕ_m. It is evident that the magnetic field lines (which are normal to the equipotential surfaces) bend to become parallel to the surface near it.

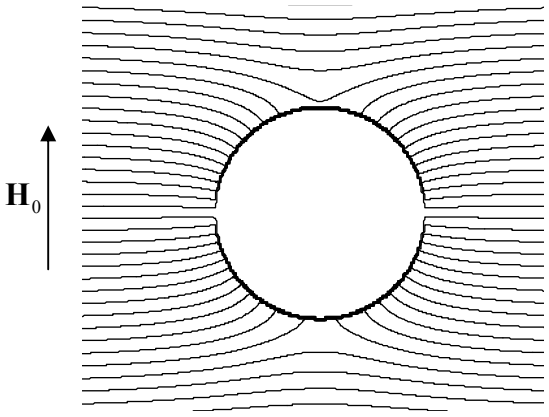


Fig. 6.3. Equipotential surfaces ϕ_m = const around a conducting sphere of radius $R >> \delta_s$ (or δ_L), placed into a uniform magnetic field, calculated within the coarse-grain (ideal-diamagnet) approximation $\mu = 0$.

This pattern also helps to answer the question that might arise at making the assumption (24): what happens to bulk conductors placed into a *normal* ac magnetic field – and to superconductors in a normal dc magnetic field as well? The answer is: the field is deformed outside of the conductor to sustain the following *coarse-grain boundary condition*:[30]

Coarse-grain boundary condition

$$B_n\big|_{\text{surface}} = 0, \tag{6.59}$$

which follows from Eq. (5.118) and the coarse-grain requirement $\mathbf{B}|_{\text{inside}} = 0$.

This answer should be taken with reservations. For normal conductors, it is only valid at sufficiently high frequencies where the skin depth (33) is relatively small: $\delta_s << a$, where a is the scale of the conductor's linear size – for a sphere, $a \sim R$. In superconductors, this simple picture requires not only that $\delta_s << a$, but also that magnetic field is relatively low because strong fields *do* penetrate

[30] Sometimes this boundary condition, as well as the (compatible) Eq. (38), are called "macroscopic". However, this term may lead to confusion with the genuine macroscopic boundary conditions (5.117)-(5.118), which also ignore the atomic-scale microstructure of the "effective currents" $\mathbf{j}_{ef} = \nabla \times \mathbf{M}$, but (as was shown earlier in this section) still allow explicit, detailed accounts of the skin-current (34) and supercurrent (55) distributions.

superconductors, destroying superconductivity (either completely or partly), and as a result violating the Meissner-Ochsenfeld effect – see the next section.

6.5. Electrodynamics of macroscopic quantum phenomena[31]

Despite the superficial similarity of the skin effect and the Meissner-Ochsenfeld effect, the electrodynamics of superconductors is much richer. For example, let us use Eq. (54) to describe the fascinating effect of *magnetic flux quantization*. Consider a closed ring/loop (not necessarily a round one) made of a superconducting "wire" with a cross-section much larger than δ_L^2 (Fig. 4a).

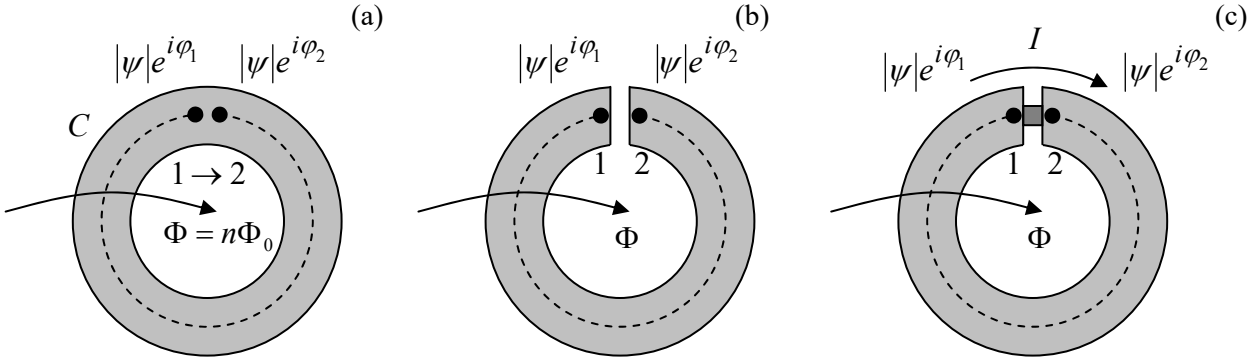


Fig. 6.4. (a) A closed, flux-quantizing superconducting ring, (b) a ring with a narrow slit, and (c) a Superconducting QUantum Interference Device (SQUID).

From the last section's discussion, we know that deep inside the wire the supercurrent is exponentially small. Integrating Eq. (54) along any closed contour C that does not approach the surface closer than a few δ_L at any point (see the dashed line in Fig. 4), so with $\mathbf{j} = 0$ at all its points, we get

$$\oint_C \nabla\varphi \cdot d\mathbf{r} - \frac{q}{\hbar}\oint_C \mathbf{A} \cdot d\mathbf{r} = 0 . \tag{6.60}$$

The first integral, i.e. the difference of φ in the initial and final points, has to be equal to either zero or an integer number of 2π because the change $\varphi \to \varphi + 2\pi n$ does not change the Cooper pair's condensate's wavefunction:

$$\psi' \equiv |\psi| e^{i(\varphi + 2\pi n)} = |\psi| e^{i\varphi} \equiv \psi . \tag{6.61}$$

On the other hand, according to Eq. (5.65), the second integral in Eq. (60) is just the magnetic flux Φ through the contour.[32] As a result, we get a wonderful result:

[31] The material of this section is not covered in most E&M textbooks, and will not be used in later sections of this course. Thus the "only" loss due to the reader's skipping this section would be the lack of familiarity with one of the most fascinating fields of physics. Note also that we already have virtually all formal tools necessary for its discussion, so reading this section should not require much effort.

[32] Due to the Meissner-Ochsenfeld effect, the exact path of the contour is not important, and we may discuss Φ just as the magnetic flux through the ring.

Magnetic flux quantization

$$\Phi = n\Phi_0, \quad \text{where } \Phi_0 \equiv \frac{2\pi\hbar}{|q|}, \quad \text{with } n = 0, \pm 1, \pm 2, \ldots, \tag{6.62}$$

saying that the magnetic flux inside any superconducting loop can only take values multiple of the *flux quantum* Φ_0. This effect, predicted in 1950 by the same Fritz London (who expected q to be equal to the electron charge $-e$), was observed experimentally in 1961,[33] but with $|q| = 2e$ – so $\Phi_0 \approx 2.07\times10^{-15}$ Wb. Historically, this observation gave decisive support to the BCS theory of superconductivity (implying Cooper pairs with charge $q = -2e$) that had been put forward just four years earlier.

Note the truly macroscopic character of this quantum effect: it has been repeatedly observed in human-scale superconducting loops, and from what is known about superconductors, there is no doubt that if we had made a giant superconducting wire loop extending, say, over the Earth's equator, the magnetic flux through it would still be quantized – though with a very large flux quanta number n. This means that the quantum coherence of Bose-Einstein condensates may extend over, using H. Casimir's famous expression, "miles of dirty lead wire". (Lead is a typical superconductor, with $T_c \approx 7.2$ K, and indeed retains its superconductivity even being highly contaminated by impurities.)

Moreover, hollow rings are not entirely necessary for flux quantization. In 1957, A. Abrikosov explained the counter-intuitive high-field behavior of superconductors with $\delta_L > \xi\sqrt{2}$, known experimentally as their *mixed* (or "Shubnikov") *phase* since the 1930s. He showed that a sufficiently high magnetic field may penetrate such superconductors in the form of self-formed magnetic field "threads" (or "tubes") surrounded by vortex-shaped supercurrents – the so-called *Abrikosov vortices.* In the simplest case, the core of such a vortex is a straight line, on which the superconductivity is completely suppressed ($|\psi| = 0$), surrounded by circular, axially-symmetric, persistent supercurrents $\mathbf{j}(\rho)$, where ρ is the distance from the vortex axis – see Fig. 5a. At the axis, the current vanishes, and with the growth of ρ, it first rises and then falls (with $\mathbf{j}(\infty) = 0$), reaching its maximum at $\rho \sim \xi$, while the magnetic field $\mathbf{B}(\rho)$, directed along the vortex axis, is largest at $\rho = 0$, and drops monotonically at distances of the order of δ_L (Fig. 5b).

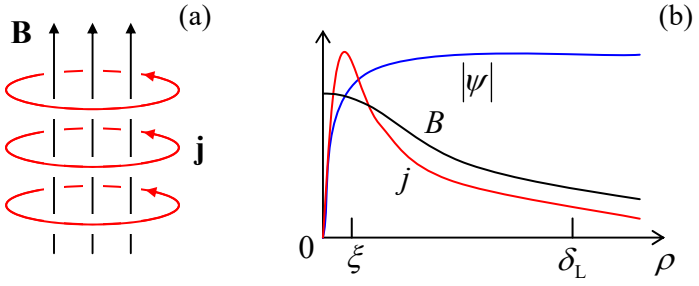


Fig. 6.5. The Abrikosov vortex: (a) a 3D structure's sketch, and (b) the main variables as functions of the distance ρ from the axis (schematically).

The total flux of the field equals exactly one flux quantum Φ_0, given by Eq. (62). Correspondingly, the wavefunction's phase φ performs just one $\pm2\pi$ revolution along *any* contour drawn around the vortex's axis, so $\nabla\varphi = \pm\mathbf{n}_\varphi/\rho$, where $\mathbf{n}_\varphi$ is the azimuthal unit vector.[34] This topological feature of the wavefunction's phase is sometimes called *fluxoid quantization* – to distinguish it from

[33] Independently and virtually simultaneously by two groups: B. Deaver and W. Fairbank, and R. Doll and M. Näbauer; their reports were published back-to-back in the same issue of the *Physical Review Letters*.

[34] The last (perhaps, evident) expression for $\nabla\varphi$ follows from MA Eq. (10.2) with $f = \pm\varphi + \text{const}$.

magnetic *flux quantization*, which is valid only for relatively large contours, not approaching the axis by distances $\sim\delta_L$.

A quantitative analysis of Abrikosov vortices requires, besides the equations we have discussed, one more constituent relation that would describe the suppression of the number of Cooper pairs (quantified by $|\psi|^2$) by the magnetic field – or rather by the field-induced supercurrent. In his original work, Abrikosov used for this purpose the famous *Ginzburg-Landau* equation,[35] which is quantitatively valid only at $T \approx T_c$. The equation may be conveniently represented in either of the following two forms:

$$\frac{1}{2m}\left(-i\hbar\nabla - q\mathbf{A}\right)^2\psi = a\psi - b\psi|\psi|^2, \qquad \xi^2\psi^*\left(\nabla - i\frac{q}{\hbar}\mathbf{A}\right)^2\psi = \left(1-|\psi|^2\right)|\psi|^2, \tag{6.63}$$

where a and b are certain temperature-dependent coefficients, with $a \to 0$ at $T \to T_c$. The first of these forms clearly shows that the Ginzburg-Landau equation (as well as the similar *Gross-Pitaevskii* equation describing electrically-neutral Bose-Einstein condensates) belongs to a broader class of *nonlinear Schrödinger equations*, differing from the usual Schrödinger equation, which is linear in ψ, only by the additional nonlinear terms. The equivalent, second form of Eq. (63) is more convenient for applications and shows more clearly that if the superconductor's condensate density, proportional to $|\psi|^2$, is suppressed only locally, it self-restores to its unperturbed value (with $|\psi|^2 = 1$) at the distances of the order of the coherence length $\xi \equiv \hbar/(2ma)^{1/2}$.

This fact enables a simple quantitative analysis of the Abrikosov vortex in the most important limit $\xi << \delta_L$. Indeed, as Fig. 5 shows, in this case. $|\psi|^2 = 1$ at most distances ($\rho \sim \delta_L$) where the field and current are distributed, so these distributions may be readily calculated without any further involvement of Eq. (63), just from Eq. (54) with $\nabla\varphi = \pm\mathbf{n}_\varphi/\rho$, and the Maxwell equations (21) for the magnetic field, giving $\nabla\times\mathbf{B} = \mu\mathbf{j}$, and $\nabla\cdot\mathbf{B} = 0$. Indeed, combining these equations just as this was done at the derivation of Eq. (23), for the only Cartesian component of the vector $\mathbf{B}(\mathbf{r}) = B(\rho)\mathbf{n}_z$ (where the z-axis is directed along the vortex' symmetry axis), we get a simple equation

$$\delta_L^2\nabla^2 B - B = -\frac{\hbar}{q}\nabla\times(\nabla\times\varphi) \equiv \mp\Phi_0\delta_2(\boldsymbol{\rho}), \qquad \text{at } \rho >> \xi, \tag{6.64}$$

which coincides with Eq. (56) at all regular points $\rho \neq 0$. Spelling out the Laplace operator for our current case of axial symmetry,[36] we get an ordinary differential equation,

$$\delta_L^2\frac{1}{\rho}\frac{d}{d\rho}\left(\rho\frac{dB}{d\rho}\right) - B = 0, \qquad \text{for } \rho \neq 0. \tag{6.65}$$

Comparing this equation with Eq. (2.155) with $\nu = 0$, and taking into account that we need the solution decreasing at $\rho \to \infty$, making any contribution proportional to the function I_0 unacceptable, we get

[35] This equation was derived by Vitaly Lazarevich Ginzburg and Lev Davidovich Landau from phenomenological arguments in 1950, i.e. before the advent of the "microscopic" BSC theory, and may be used for simple analyses of a broad range of nonlinear effects in superconductors. The Ginzburg-Landau and Gross-Pitaevskii equations will be further discussed in SM Sec. 4.3.

[36] See, e.g., MA Eq. (10.3) with $\partial/\partial\varphi = \partial/\partial z = 0$.

$$B = CK_0\left(\frac{\rho}{\delta_L}\right) \tag{6.66}$$

– see the plot of this Bessel function on the right panel of Fig. 2.22 (black line). The constant C should be calculated by fitting the 2D delta function on the right-hand side of Eq. (64), i.e. by requiring

$$\int_{\text{vortex}} B(\rho)d^2\rho \equiv 2\pi\int_0^\infty B(\rho)\rho d\rho \equiv 2\pi\delta_L^2 C\int_0^\infty K_0(\zeta)\zeta d\zeta = \mp\Phi_0 . \tag{6.67}$$

The last, dimensionless integral equals 1,[37] so finally

$$B(\rho) = \frac{\Phi_0}{2\pi\delta_L^2}K_0\left(\frac{\rho}{\delta_L}\right), \qquad \text{at } \rho >> \xi . \tag{6.68}$$

So the magnetic field of the vortex drops exponentially at distances ρ much larger than δ_L, and diverges at $\rho \to 0$ – see, e.g., the second of Eqs. (2.157). However, this divergence is very slow (logarithmic), and, as was repeatedly discussed in this series, is avoided by the account of virtually any other factor. In our current case, this factor is the decrease of $|\psi|^2$ to zero at $\rho \sim \xi$ (see Fig. 5), not taken into account in Eq. (68). As a result, we may estimate the field on the axis of the vortex as

$$B(0) \approx \frac{\Phi_0}{2\pi\delta_L^2}\ln\frac{\delta_L}{\xi}; \tag{6.69}$$

the exact (and much more involved) solution of the problem confirms this estimate with a minor correction: $\ln(\delta_L/\xi) \to \ln(\delta_L/\xi) - 0.28$, i.e. $\xi \to 1.3\xi$.

The current density distribution may be now calculated from the Maxwell equation $\nabla \times \mathbf{B} = \mu\mathbf{j}$, giving $\mathbf{j} = j(\rho)\mathbf{n}_\varphi$, with[38]

$$j(\rho) = -\frac{1}{\mu}\frac{\partial B}{\partial \rho} = -\frac{\Phi_0}{2\pi\mu\delta_L^2}\frac{\partial}{\partial\rho}K_0\left(\frac{\rho}{\delta_L}\right) \equiv \frac{\Phi_0}{2\pi\mu\delta_L^3}K_1\left(\frac{\rho}{\delta_L}\right), \qquad \text{at } \rho >> \xi , \tag{6.70}$$

where the same identity (2.158), with $J_n \to K_n$ and $n = 1$, was used. Now looking at Eqs. (2.157) and (2.158), with $n = 1$, we see that the supercurrent's density is exponentially low at $\rho >> \delta_L$ (thus outlining the vortex' periphery), and is proportional to $1/\rho$ within the broad range $\xi << \rho << \delta_L$. This rise of the current at $\rho \to 0$ (which could be readily predicted directly from Eq. (54) with $\nabla\varphi = \pm\mathbf{n}_\varphi/\rho$, and the $\mathbf{A}$-term negligible at $\rho << \delta_L$) is quenched at $\rho \sim \xi$ by a rapid drop of the factor $|\psi|^2$ in the same Eq. (54), i.e. by the suppression of the superconductivity near the axis (by the same supercurrent!) – see Fig. 5 again.

This structure of the Abrikosov vortex may be used to calculate, in a straightforward way, its energy per unit length (i.e. its linear tension)

[37] This fact follows, for example, from the integration of both sides of Eq. (2.143) (which is valid for any Bessel functions, including K_n) with $n = 1$, from 0 to ∞, and then using the asymptotic values given by Eqs. (2.157)-(2.158): $K_1(\infty) = 0$, and $K_1(\zeta) \to 1/\zeta$ at $\zeta \to 0$.

[38] See, e.g., MA Eq. (10.5), with $f_\rho = f_\varphi = 0$, and $f_z = B(\rho)$.

$$\mathcal{T} \equiv \frac{U}{l} \approx \frac{\Phi_0^2}{4\pi\mu\delta_L^2} \ln\frac{\delta_L}{\xi}, \tag{6.71}$$

and hence the so-called "first critical" value H_{c1} of the external magnetic field,[39] at which the vortex formation becomes possible (in a long cylindrical sample parallel to the field):

$$H_{c1} = \frac{\mathcal{T}}{\Phi_0} \approx \frac{\Phi_0}{4\pi\mu\delta_L^2} \ln\frac{\delta_L}{\xi}. \tag{6.72}$$

Let me leave the proof of these two formulas for the reader's exercise.

The flux quantization and the Abrikosov vortices discussed above are just two of several *macroscopic quantum effects* in superconductivity. Let me discuss just one more, but perhaps the most interesting of such effects. Let us consider a superconducting ring/loop interrupted with a very narrow slit (Fig. 4b). Integrating Eq. (54) along any current-free path from point 1 to point 2 (see, e.g., dashed line in Fig. 4b), we get

$$0 = \int_1^2 \left(\nabla\varphi - \frac{q}{\hbar}\mathbf{A}\right)\cdot d\mathbf{r} = \varphi_2 - \varphi_1 - \frac{q}{\hbar}\Phi. \tag{6.73}$$

Using the flux quantum definition (62), this result may be rewritten as

$$\varphi \equiv \varphi_1 - \varphi_2 = \frac{2\pi}{\Phi_0}\Phi, \tag{6.74}$$

Josephson phase difference

where φ is called the *Josephson phase difference*. Note that in contrast to each of the phases $\varphi_{1,2}$, their difference φ is gauge-invariant: Eq. (74) directly relates it to the gauge-invariant magnetic flux Φ.

Can this φ be measured? Yes, for example, using the *Josephson effect*.[40] Let us consider two (for the argument simplicity, similar) superconductors, connected with some sort of *weak link*, for example, a small tunnel junction, or a point contact, or a narrow thin-film bridge, through which a weak Cooper-pair supercurrent can flow. (Such a system of two weakly coupled superconductors is called a *Josephson junction*.) Let us think about what this supercurrent I may be a function of. For that, reverse thinking is helpful: let us imagine that we change the current; what parameter of the superconducting condensate can it affect? If the current is very weak, it cannot perturb the superconducting condensate's density, proportional to $|\psi|^2$; hence it may only change the Cooper condensate phases $\varphi_{1,2}$. However, according to Eq. (53), the phases are not gauge-invariant, while the current should be. Hence the current may affect (or, if you like, may be affected by) only the phase difference φ defined by Eq. (74). Moreover, just has already been argued during the flux quantization discussion, a change of any of $\varphi_{1,2}$ (and hence of φ) by 2π or any of its multiples should not change the current. Also, if the wavefunction is the same in both superconductors ($\varphi = 0$), the supercurrent should vanish due to the system's symmetry. Hence the function $I(\varphi)$ should satisfy the following conditions:

[39] This term is used to distinguish H_{c1} from the higher "second critical field" H_{c2}, at which the Abrikosov vortices are pressed to each other so tightly (to distances $d \sim \xi$) that they merge, and the remains of superconductivity vanish: $\psi \to 0$. Unfortunately, I do not have time/space to discuss these effects; the interested reader may be referred, for example, to Chapter 5 of M. Tinkham's monograph cited above.

[40] It was predicted in 1961 by Brian David Josephson (then a PhD student!) and observed experimentally by several groups soon after that.

$$I(0) = 0, \qquad I(\varphi + 2\pi) = I(\varphi). \tag{6.75}$$

With these conditions on hand, we should not be terribly surprised by the following Josephson's result that for the weak link provided by tunneling,[41]

Josephson (super)current

$$\boxed{I(\varphi) = I_c \sin\varphi}, \tag{6.76}$$

where constant I_c, which depends on the weak link's strength and temperature, is called the *critical current*. Actually, Eqs. (54) and (63) enable not only a straightforward calculation of this relation but even obtaining a simple expression of the critical current I_c via the link's normal-sate resistance – the task left for the (creative :-) reader's exercise.

Now let us see what happens if a Josephson junction is placed into the gap in a superconductor loop – see Fig. 4c. In this case, we may combine Eqs. (74) and (76), getting

Macroscopic quantum interference

$$\boxed{I = I_c \sin\left(2\pi \frac{\Phi}{\Phi_0}\right)}. \tag{6.77}$$

This effect of a periodic dependence of the current on the magnetic flux is called *macroscopic quantum interference*,[42] while the system shown in Fig. 4c, the *superconducting quantum interference device – SQUID* (with all letters capitalized, please :-). The low value of the magnetic flux quantum Φ_0, and hence the high sensitivity of φ to external magnetic fields, allows using such SQUIDs as ultrasensitive magnetometers. Indeed, for a superconducting ring of area ~1 cm^2, one period of the change of the supercurrent (77) is produced by a magnetic field change of the order of 10^{-11} T (10^{-7} Gs), while sensitive electronics allows measuring a tiny fraction of this period – limited by thermal noise at a level of the order of a few fT. Such sensitivity allows measurements, for example, of the miniscule magnetic fields induced outside of the body by the beating human heart, and even by brain activity.[43]

An important aspect of quantum interference is the so-called *Aharonov-Bohm* (AB) *effect* – which actually takes place for single quantum particles as well.[44] Let the magnetic field lines be limited to the central, hollow part of the SQUID loop so that no appreciable magnetic field ever touches the ring itself. (This may be done experimentally with very good accuracy, for example using high-μ magnetic cores – see their discussion in Sec. 5.6.) As predicted by Eq. (77), and confirmed by several careful experiments carried out in the mid-1960s,[45] this restriction does not matter – the interference is observed

[41] For some other types of weak links, the function $I(\varphi)$ may deviate from the sinusoidal form Eq. (76) rather considerably, while still satisfying the general conditions (75).

[42] The name is due to a deep analogy between this phenomenon and the interference between two coherent waves, to be discussed in detail in Sec. 8.4.

[43] Other practical uses of SQUIDs include MRI signal detectors, high-sensitive measurements of magnetic properties of materials, and weak field detection in a broad variety of physical experiments – see, e.g., J. Clarke and A. Braginski (eds.), *The SQUID Handbook*, vol. II, Wiley, 2006. For a comparison of these devices with other sensitive magnetometers see, e.g., the review collection by A. Grosz *et al.* (eds.), *High Sensitivity Magnetometers*, Springer, 2017.

[44] For a more detailed discussion of the AB effect see, e.g., QM Sec. 3.2.

[45] Similar experiments have been carried out with single (unpaired) electrons – moving either ballistically, in vacuum, or in "normal" (non-superconducting) conducting rings. In the last case, the effect is much harder to observe than in SQUIDs: the ring size has to be very small, and temperature very low, to avoid the so-called

anyway. This means that not only the magnetic field **B** but also the vector potential **A** represents physical reality, albeit in a quite peculiar way – remember the gauge transformation (5.46), which you may carry out in your head, without changing any physical reality? (Fortunately, this transformation does not change the contour integral participating in Eq. (5.65), and hence the magnetic flux Φ, and hence the interference pattern.)

Actually, the magnetic flux quantization (62) and the macroscopic quantum interference (77) are not completely different effects, but just two manifestations of the interrelated macroscopic quantum phenomena. To show that, one should note that if the critical current I_c (or rather its product by the loop's self-inductance L) is high enough, the flux Φ in the SQUID loop is due not only to the external magnetic field flux Φ_{ext} but also has a self-field component – cf. Eq. (5.68):[46]

$$\Phi = \Phi_{ext} - LI, \qquad \text{where } \Phi_{ext} \equiv \int_S (B_{ext})_n d^2 r . \tag{6.78}$$

Now the relation between Φ and Φ_{ext} may be readily found by solving this equation together with Eq. (77). Figure 6 shows this relation for several values of the dimensionless parameter $\lambda \equiv 2\pi LI_c/\Phi_0$.

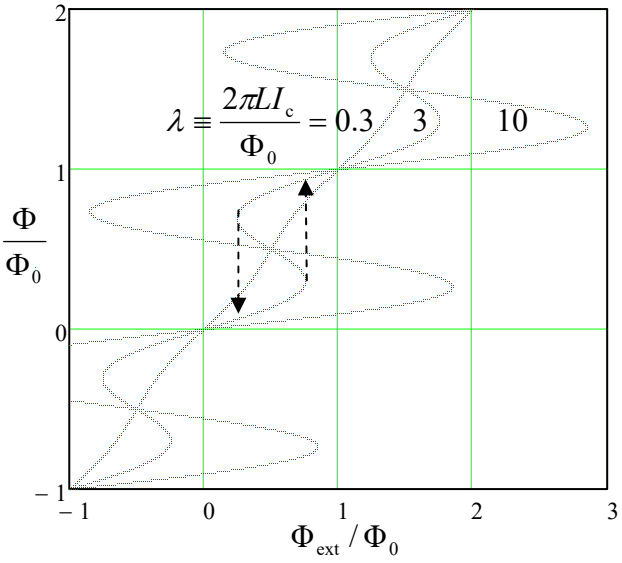


Fig. 6.6. The function $\Phi(\Phi_{ext})$ for SQUIDs with various values of the normalized LI_c product. Dashed arrows show the flux leaps as the external field is changed. (The branches with $d\Phi/d\Phi_{ext} < 0$ are unstable.)

These plots show that if the critical current (and/or the inductance) is low, $\lambda << 1$, the self-field effects are negligible, and the total flux follows the external field (i.e., Φ_{ext}) faithfully. However, at $\lambda > 1$, the function $\Phi(\Phi_{ext})$ becomes hysteretic, and at $\lambda >> 1$, its stable (positive-slope) branches are nearly flat, with the total flux values corresponding to Eq. (62). Thus, a superconducting ring closed with a high-I_c Josephson junction exhibits a nearly-perfect flux quantization.

The self-field effects described by Eq. (78) create certain technical problems for SQUID magnetometry, but they are the basis for one more useful application of these devices: ultrafast

dephasing effects due to unavoidable interactions of the electrons with their environment – see, e.g., QM Chapter 7.

[46] The sign before *LI* would be positive, as in Eq. (5.70), if *I* was the current flowing *into* the inductance. However, in order to keep the sign in Eq. (76) intact, *I* should mean the current flowing into the Josephson junction, i.e. *from* the inductance, thus changing the sign of the *LI* term in Eq. (78).

computing. Indeed, Fig. 6 shows that at the values of λ modestly above 1 (e.g., $\lambda \approx 3$), and within a certain range of applied field, the SQUID has two stable flux states, which differ by $\Delta\Phi \approx \Phi_0$ and may be used for coding binary 0 and 1. For practical superconductors (like Nb), the time of switching between these states (see dashed arrows in Fig. 4) is of the order of a picosecond, while the energy dissipated at such event may be as low as $\sim 10^{-19}$ J. (This bound is determined not by device's physics, by the fundamental requirement for the energy barrier between the two states to be much higher than the thermal fluctuation energy scale $k_B T$, ensuring a sufficiently long information retention time.) While the picosecond switching speed may be also achieved with some semiconductor devices, the power consumption of the SQUID-based digital devices may be 5 to 6 orders of magnitude lower, enabling large-scale digital integrated circuits with 100-GHz-scale clock frequencies. Unfortunately, the range of practical applications of these *Rapid Single-Flux-Quantum* (RSFQ) digital circuits is still very narrow, due to the inconvenience of their deep refrigeration to temperatures below T_c.[47]

Since we have already got the basic relations (74) and (76) describing the macroscopic quantum phenomena in superconductivity, let me mention in brief two other prominent members of this group, called the *dc* and *ac Josephson effects*. Differentiating Eq. (74) over time, and using the Faraday induction law (2), we get[48]

Josephson phase-to-voltage relation

$$\frac{d\varphi}{dt} = \frac{2e}{\hbar} V . \tag{6.79}$$

This famous *Josephson phase-to-voltage relation* should be valid regardless of the way how the voltage V has been created,[49] so let us apply Eqs. (76) and (79) to the simplest circuit with a non-superconducting source of dc voltage – see Fig. 7.

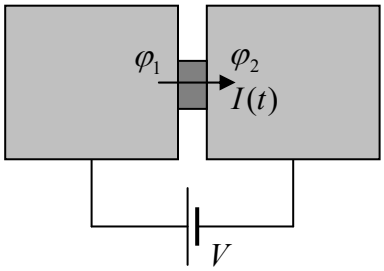


Fig. 6.7. DC-voltage-biased Josephson junction.

If the current's magnitude is below the critical value, Eq. (76) allows phase φ to have the time-independent value

$$\varphi = \sin^{-1}\frac{I}{I_c}, \qquad \text{if } -I_c < I < +I_c , \tag{6.80}$$

and hence, according to Eq. (79), a vanishing voltage drop across the junction: $V = 0$. This *dc Josephson effect* is not quite surprising – indeed, we have postulated from the very beginning that the Josephson junction may pass a certain supercurrent. Much more fascinating is the so-called *ac Josephson effect* that occurs if the voltage across the junction has a non-zero average (dc) component V_0. For simplicity, let us

[47] For more on that technology, see, e.g., the review paper by P. Bunyk *et al.*, *Int. J. High Speed Electron. Syst.* **11**, 257 (2001), and references therein.

[48] Since the induced e.m.f. $\mathcal{V}_{ind}$ cannot drop on the superconducting path between the Josephson junction electrodes 1 and 2 (see Fig. 4c), it should be equal to $(-V)$, where V is the voltage across the junction.

[49] Indeed, it may be also obtained from simple Schrödinger-equation-based arguments – see, e.g., QM Sec. 1.6.

assume that this is the *only* voltage component: $V(t) = V_0 = \text{const}$;[50] then Eq. (79) may be easily integrated to give $\varphi = \omega_J t + \varphi_0$, where

$$\omega_J \equiv \frac{2e}{\hbar} V_0 . \tag{6.81}$$

Josephson oscillation frequency

This result, plugged into Eq. (76), shows that the supercurrent oscillates,

$$I = I_c \sin\left(\omega_J t + \varphi_0\right), \tag{6.82}$$

with the so-called *Josephson frequency* ω_J (81) proportional to the applied dc voltage. For practicable voltages (above the typical noise level), the frequency $f_J = \omega_J/2\pi$ corresponds to the GHz or even THz ranges, because the proportionality coefficient in Eq. (81) is very high: $f_J/V_0 = e/\pi\hbar \approx 483$ MHz/μV.[51]

An important experimental fact is the universality of this coefficient. For example, in the mid-1980s, a Stony Brook group led by J. Lukens proved that this factor is material-independent with a relative accuracy of at least 10^{-15}. Very few experiments, especially in solid-state physics, have ever reached such precision. This fundamental nature of the Josephson voltage-to-frequency relation (81) allows an important application of the ac Josephson effect in metrology. Namely, phase-locking[52] the Josephson oscillations with an external microwave signal from an atomic frequency standard, one can get a more precise dc voltage than from any other source. In NIST and other metrological institutions around the globe, this effect is used for the calibration of simpler "secondary" voltage standards that can operate at room temperature.

6.6. Inductors, transformers, and ac Kirchhoff laws

Let a *wire coil* (meaning either a single loop illustrated in Fig. 5.4b or a series of such loops, such as one of the solenoids shown in Fig. 5.6) have a self-inductance L much larger than that of the wires connecting it to other components of our system: ac voltage sources, voltmeters, etc. (Since, according to Eq. (5.75), L scales as the square of the number N of wire turns, this condition is easier to satisfy at $N >> 1$.) Then in a quasistatic system consisting of such *lumped induction coils*, external wires, and other lumped circuit elements such as resistors, capacitances, etc., we may neglect the electromagnetic induction effects everywhere outside the coil, so the electric field in those external regions is potential. Then the voltage V between the coil's terminals may be defined, just as in electrostatics, as the difference of values of ϕ between the terminals, i.e. as the integral

$$V = \int \mathbf{E} \cdot d\mathbf{r} \tag{6.83}$$

between the coil terminals along any path outside the coil. This voltage has to be balanced by the induction e.m.f. (2) in the coil, so if the Ohmic resistance of the coil is negligible, we may write

[50] In experiment, this condition is hard to implement, due to the relatively high inductances of the current leads providing the dc voltage supply. However, this technical complication does not affect the main conclusion of the simple analysis described here.

[51] This 1962 prediction (by the same B. Josephson) was confirmed experimentally – in 1963 indirectly, by phase-locking of the oscillations (82) with an external microwave signal, and in 1967 explicitly, by the direct detection of the emitted microwave radiation.

[52] For a discussion of this very important (and general) effect, see, e.g., CM Sec. 5.4.

$$V = \frac{d\Phi}{dt}, \tag{6.84}$$

where Φ is the magnetic flux in the coil.[53] If the flux is due to the current I in the same coil only (i.e. if it is magnetically uncoupled from other coils), we may use Eq. (5.70) to get the well-known relation

Voltage drop on inductance coil

$$V = L\frac{dI}{dt}, \tag{6.85}$$

where compliance with the Lenz sign rule is achieved by selecting the relations between the assumed voltage polarity and the current direction as shown in Fig. 8a.

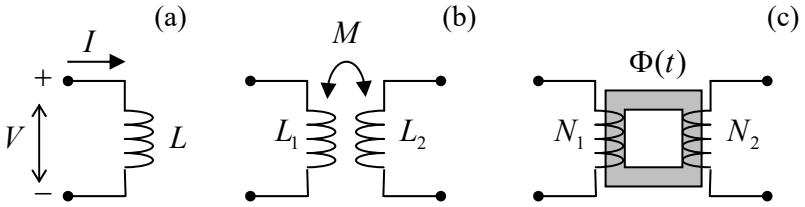


Fig. 6.8. Some lumped ac circuit elements: (a) an induction coil, (b) two inductively coupled coils, and (c) an ac transformer.

If similar conditions are satisfied for two magnetically coupled coils (Fig. 8b), then, in Eq. (84), we need to use Eqs. (5.69) instead, getting

$$V_1 = L_1\frac{dI_1}{dt} + M\frac{dI_2}{dt}, \qquad V_2 = L_2\frac{dI_2}{dt} + M\frac{dI_1}{dt}. \tag{6.86}$$

Such systems of inductively coupled coils have numerous applications in electrical engineering and physical experiment. Perhaps the most important of them is the *ac transformer*, in which the coils share a common soft-ferromagnetic core of the toroidal ("doughnut") topology – see Fig. 8c.[54] As we already know from the discussion in Sec. 5.6, such cores, with $\mu >> \mu_0$, "try" to absorb all magnetic field lines, so the magnetic flux $\Phi(t)$ in the core is nearly the same in each of its cross-sections. With this, Eq. (84) yields

$$V_1 \approx N_1\frac{d\Phi}{dt}, \quad V_2 \approx N_2\frac{d\Phi}{dt}, \tag{6.87}$$

so the voltage ratio is completely determined by the ratio N_1/N_2 of the number of wire turns.

Now we may generalize, to the ac current case, the Kirchhoff laws already discussed in Sec. 4.1 – see Fig. 4.3 reproduced in Fig. 9a below. Let not only inductances but also capacitances and resistances of the wires be negligible in comparison with those of the lumped (compact) circuit elements, whose list now would include not only resistors and current sources (as in the dc case), but also the induction coils (including magnetically coupled ones) and capacitors – see Fig. 9b. In the quasistatic approximation, the current flowing in each wire is conserved, so the "node rule", i.e. the 1st Kirchhoff law (4.7a),

[53] If the resistance is substantial, it may be represented by a separate lumped circuit element (resistor) connected in series with the coil.

[54] The first practically acceptable form of this device, called the *Stanley transformer*, was invented in 1886. In it, multi-turn windings could be easily mounted onto a toroidal ferromagnetic (at that time, silicon-steel-plate) core.

$$\sum_j I_j = 0 . \tag{6.88a}$$

remains valid. Also, if the electromagnetic induction effect is restricted to the interior of lumped induction coils as discussed above, the voltage drops V_k across each circuit element may be still represented, just as in dc circuits, with differences between the adjacent node potentials. As a result, the "loop rule", i.e. 2nd Kirchhoff law (4.7b),

$$\sum_k V_k = 0 , \tag{6.88b}$$

is also valid. Now, in contrast to the dc case, Eqs. (88) may be the (ordinary) differential equations. However, if all circuit elements are linear (as in the examples presented in Fig. 9b), these equations may be readily reduced to linear algebraic equations, using the Fourier expansion. (In the common case of sinusoidal ac sources, the final stage of the Fourier series summation is unnecessary.)

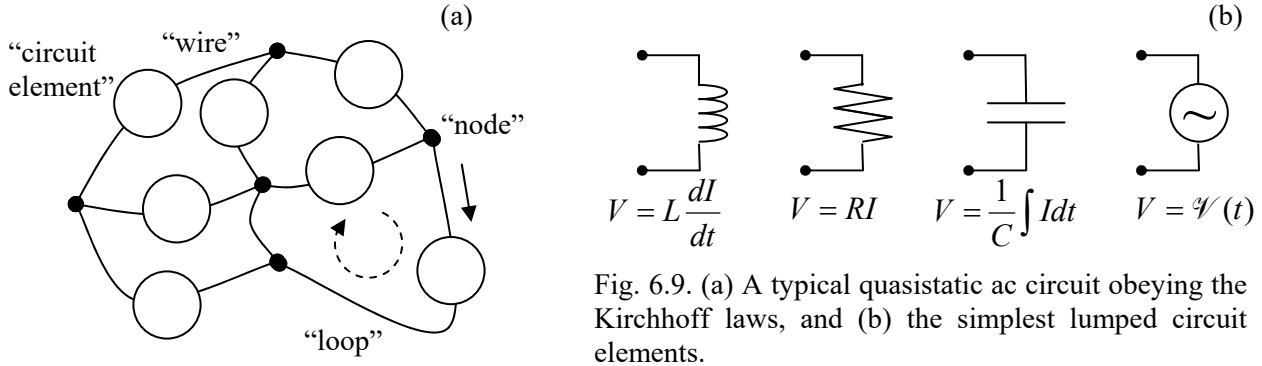


Fig. 6.9. (a) A typical quasistatic ac circuit obeying the Kirchhoff laws, and (b) the simplest lumped circuit elements.

My teaching experience shows that the potential readers of these notes are well familiar with the application of Eqs. (88) to such problems from their undergraduate studies, so I will save time/space by skipping discussions of even the simplest examples of such circuits, such as *LC*, *LR*, *RC*, and *LRC* loops and periodic structures.[55] However, since such problems are very important for practice, my sincere advice to the reader is to carry out a self-test by solving a few problems of this type, provided in Sec. 9 below, and if they cause any difficulty, pursue some remedial reading.

6.7. Displacement currents

Electromagnetic induction is not the only new effect arising in non-stationary electrodynamics. Indeed, though Eqs. (21) are adequate for the description of quasistatic phenomena, a deeper analysis shows that one of these equations, namely $\nabla \times \mathbf{H} = \mathbf{j}$, cannot be exact. To see that, let us take the divergence of both sides:

$$\nabla \cdot (\nabla \times \mathbf{H}) = \nabla \cdot \mathbf{j} . \tag{6.89}$$

But, as the divergence of any curl,[56] the left-hand side should equal zero. Hence we get

[55] Curiously enough, these effects include wave propagation in periodic *LC* circuits, even within the quasistatic approximation! However, the speed $1/(LC)^{1/2}$ of these waves in lumped circuits is much lower than the speed $1/(\varepsilon\mu)^{1/2}$ of electromagnetic waves in the surrounding medium – see Sec. 8 below.

[56] Again, see MA Eq. (11.2) – if you need it.

$$\nabla \cdot \mathbf{j} = 0 . \tag{6.90}$$

This is fine in statics, but in dynamics, this equation forbids any charge accumulation, because according to the continuity relation (4.5),

$$\nabla \cdot \mathbf{j} = -\frac{\partial \rho}{\partial t} . \tag{6.91}$$

This discrepancy had been recognized by James Clerk Maxwell who suggested, in the 1860s, a way out of this contradiction. If we generalize the equation for $\nabla \times \mathbf{H}$ by adding to the term $\mathbf{j}$ (that describes the density of real electric currents) the so-called *displacement current* density term,

Displacement current density

$$\boxed{\mathbf{j}_{\mathrm{d}} \equiv \frac{\partial \mathbf{D}}{\partial t}}, \tag{6.92}$$

(which of course vanishes in statics), then the equation takes the form

$$\nabla \times \mathbf{H} = \mathbf{j} + \mathbf{j}_{\mathrm{d}} \equiv \mathbf{j} + \frac{\partial \mathbf{D}}{\partial t} . \tag{6.93}$$

In this case, due to the equation (3.22), $\nabla \cdot \mathbf{D} = \rho$, the divergence of the right-hand side equals zero due to the continuity equation (92), and the discrepancy is removed. This incredible theoretical feat,[57] confirmed by the 1886 experiments carried out by Heinrich Hertz (see below) was perhaps the main triumph of theoretical physics of the 19th century.

Maxwell's displacement current concept, expressed by Eq. (93), is so important that it is worthwhile to have one more look at its derivation using a particular model shown in Fig. 10.[58]

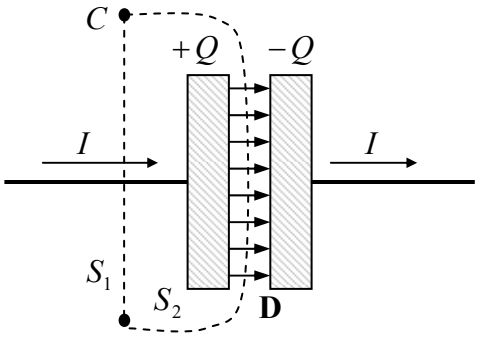


Fig. 6.10. The Ampère law applied to capacitor recharging.

Neglecting the fringe field effects, we may use Eq. (4.1) to describe the relationship between the current I flowing through the wires and the electric charge Q of the capacitor:[59]

$$\frac{dQ}{dt} = I . \tag{6.94}$$

[57] It looks deceivingly simple now – after the fact, and with the current mathematical tools (especially the del operator), which are much superior to those that were available to J. Maxwell.

[58] No physicist should be ashamed of doing this. For example, J. Maxwell's main book, *A Treatise of Electricity and Magnetism*, is full of drawings of plane capacitors, inductance coils, and voltmeters. More generally, the whole history of science teaches us that snobbery regarding particular examples and practical systems is a virtually certain path toward producing nothing of either practical value or fundamental importance.

[59] This is of course just the integral form of the continuity equation (91).

Now let us consider a closed contour C drawn around the wire. (Solid points in Fig. 10 show the places where the contour intercepts the plane of the drawing.) This contour may be seen as the line limiting either surface S_1 (crossed by the wire) or surface S_2 (avoiding such crossing by passing through the capacitor's gap). Applying the macroscopic Ampère law (5.116) to the former surface, we get

$$\oint_C \mathbf{H}\cdot d\mathbf{r} = \int_{S_1} j_n d^2r = I, \tag{6.95}$$

while for the latter surface the same law gives a different result,

$$\oint_C \mathbf{H}\cdot d\mathbf{r} = \int_{S_2} j_n d^2r = 0, \qquad \text{[WRONG!]} \tag{6.96}$$

for the same integral. This is just an integral-form manifestation of the discrepancy outlined above, but it shows clearly how serious the problem is (or rather it was – before Maxwell).

Now let us see how the introduction of the displacement currents saves the day, considering for the sake of simplicity a plane capacitor of area A, with a small and constant electrode spacing. In this case, as we already know, the field inside it is uniform, with $D = \sigma$, so the total capacitor's charge $Q = A\sigma = AD$, and the current (94) may be represented as

$$I = \frac{dQ}{dt} = A\frac{dD}{dt}. \tag{6.97}$$

So, instead of the wrong Eq. (96), the Ampère law modified following Eq. (93), gives

$$\oint_C \mathbf{H}\cdot d\mathbf{r} = \int_{S_2} (j_{\text{d}})_n d^2r = \int_{S_2} \frac{\partial D_n}{\partial t} d^2r = \frac{dD}{dt} A = I, \tag{6.98}$$

i.e. the Ampère integral becomes independent of the choice of the surface limited by the contour C – as it has to be.

6.8. Finally, the full Maxwell equation system

. This is a very special moment in this course: with the displacement currents in, i.e. with the replacement of Eq. (5.107) with Eq. (93), we have finally arrived at the full set of macroscopic Maxwell equations for time-dependent fields,[60]

$$\nabla\times\mathbf{E} + \frac{\partial\mathbf{B}}{\partial t} = 0, \qquad \nabla\times\mathbf{H} - \frac{\partial\mathbf{D}}{\partial t} = \mathbf{j}, \tag{6.99a}$$

$$\nabla\cdot\mathbf{D} = \rho, \qquad \nabla\cdot\mathbf{B} = 0, \tag{6.99b}$$

Macroscopic Maxwell equations

whose validity has been confirmed by an enormous body of experimental data. Indeed, despite numerous efforts, no other corrections (e.g., additional terms) to the Maxwell equations have been ever found, and these equations are still considered exact within the range of their validity, i.e. while the electric and magnetic fields may be considered classically. Moreover, even in quantum theory, these

[60] This vector form of the Maxwell equations, magnificent in its symmetry and simplicity, was developed in 1884-85 by Oliver Heaviside, with substantial contributions by H. Lorentz. (The original Maxwell's result circa 1864 looked like a system of 20 equations for Cartesian components of the vector and scalar potentials.)

equations are believed to be *strictly* valid as relations between the Heisenberg operators of the electric and magnetic fields.[61] (Note that the *microscopic* Maxwell equations for the genuine fields **E** and **B** may be formally obtained from Eqs. (99) by the substitutions $\mathbf{D} = \varepsilon_0\mathbf{E}$ and $\mathbf{H} = \mathbf{B}/\mu_0$, and the simultaneous replacement of the stand-alone charge and current densities on their right-hand sides with the full ones.)

Perhaps the most striking feature of these equations is that, even in the absence of stand-alone charges and currents inside the region of our interest, when the equations become fully homogeneous,

$$\nabla\times\mathbf{E} = -\frac{\partial\mathbf{B}}{\partial t}, \qquad \nabla\times\mathbf{H} = \frac{\partial\mathbf{D}}{\partial t}, \tag{6.100a}$$

$$\nabla\cdot\mathbf{D} = 0, \qquad \nabla\cdot\mathbf{B} = 0, \tag{6.100b}$$

they still describe something very non-trivial: *electromagnetic waves*, including light. The physics of the waves may be clearly seen from Eqs. (100a): according to the first of them, the change of the magnetic field in time creates a vortex-like (divergence-free) electric field. On the other hand, the second of Eqs. (100a) describes how the changing electric field, in turn, creates a vortex-like magnetic field. So-coupled electric and magnetic fields may propagate as waves – even very far from their sources.

We will carry out a detailed quantitative analysis of the waves in the next chapter, and here I will only use this notion to make good on the promise given in Sec. 3, namely to establish the condition of validity of the quasistatic approximation (21). For simplicity, let us consider an electromagnetic wave with a time period $\mathcal{T}$, velocity v, and hence the wavelength[62] $\lambda = v\mathcal{T}$ in a linear medium with $\mathbf{D} = \varepsilon\mathbf{E}$, $\mathbf{B} = \mu\mathbf{H}$. Then the magnitude of the left-hand side of the first of Eqs. (100a) is of the order of $E/\lambda = E/v\mathcal{T}$, while that of its right-hand side may be estimated as $B/\mathcal{T} \sim \mu H/\mathcal{T}$. Using similar estimates for the second of Eqs. (100a), we arrive at the following two requirements:[63]

$$\frac{E}{H} \sim \mu v \sim \frac{1}{\varepsilon v}. \tag{6.101}$$

To ensure the compatibility of these two relations, the wave's speed should satisfy the estimate

$$v \sim \frac{1}{(\varepsilon\mu)^{1/2}}, \tag{6.102}$$

reduced to $v \sim 1/(\varepsilon_0\mu_0)^{1/2} \equiv c$ in free space, while the ratio of the electric and magnetic field amplitudes should be of the following order:

$$\frac{E}{H} \sim \mu v \sim \mu\frac{1}{(\varepsilon\mu)^{1/2}} \equiv \left(\frac{\mu}{\varepsilon}\right)^{1/2}. \tag{6.103}$$

(In the next chapter we will see that for plane electromagnetic waves, these results are exact.)

Now, let a system of a linear size $\sim a$ carry currents producing a certain magnetic field H. Then, according to Eqs. (100a), their magnetic field Faraday-induces the electric field of magnitude $E \sim \mu Ha/\mathcal{T}$, whose displacement currents, in turn, produce an additional magnetic field with magnitude

[61] See, e.g., QM Chapter 9.

[62] Let me hope the reader knows that the relation $\lambda = v\mathcal{T}$ is universal and valid for waves of any nature – see, e.g., CM Chapter 6. (In the case of substantial dispersion, v means the phase velocity.)

[63] The fact that $\mathcal{T}$ has canceled, shows that these estimates are valid for waves of any frequency.

$$H' \sim \frac{a\varepsilon}{\mathcal{T}} E \sim \frac{a\varepsilon}{\mathcal{T}} \frac{\mu a}{\mathcal{T}} H \equiv \left(\frac{a\lambda}{v\mathcal{T}\lambda} \right)^2 H \equiv \left(\frac{a}{\lambda} \right)^2 H . \tag{6.104}$$

Hence, the displacement current effects are negligible for a system of size $a << \lambda$.[64]

In particular, the quasistatic picture of the skin effect, discussed in Sec. 3, is valid while the skin depth (33) remains much *smaller* than the corresponding wavelength,

$$\lambda = v\mathcal{T} = \frac{2\pi v}{\omega} = \left(\frac{4\pi^2}{\varepsilon\mu\omega^2} \right)^{1/2} . \tag{6.105}$$

The wavelength decreases with the frequency as $1/\omega$, i.e. faster than $\delta_s \propto 1/\omega^{1/2}$, so they become comparable at the crossover frequency

$$\omega_r = \frac{\sigma}{\varepsilon} \equiv \frac{\sigma}{\kappa\varepsilon_0}, \tag{6.106}$$

which is nothing else than the reciprocal charge relaxation time (4.10). As was discussed in Sec. 4.2, for good metals this frequency is extremely high (about 10^{18} s^{-1}), so the validity of Eq. (33) is typically limited by the anomalous skin effect (which was briefly discussed in Sec. 3), rather than the wave effects.

Before going after the analysis of the full Maxwell equations for particular situations (that will be the main goal of the next chapters of this course), let us have a look at the energy balance they yield for a certain volume V, which may include both some charged particles and the electromagnetic field. Since, according to Eq. (5.10), the magnetic field performs no work on charged particles even if they move, the total power $\mathscr{P}$ being transferred from the field to the particles inside the volume is due to the electric field alone – see Eq. (4.38):

$$\mathscr{P} = \int_V \mathscr{p}\, d^3r, \qquad \text{with } \mathscr{p} = \mathbf{j}\cdot\mathbf{E}, \tag{6.107}$$

Expressing **j** from the corresponding Maxwell equation of the system (99), we get

$$\mathscr{P} = \int_V \left[\mathbf{E}\cdot(\nabla\times\mathbf{H}) - \mathbf{E}\cdot\frac{\partial\mathbf{D}}{\partial t} \right] d^3r. \tag{6.108}$$

Let us pause here for a second, and transform the divergence of **E**×**H**, using the well-known vector algebra identity:[65]

$$\nabla\cdot(\mathbf{E}\times\mathbf{H}) = \mathbf{H}\cdot(\nabla\times\mathbf{E}) - \mathbf{E}\cdot(\nabla\times\mathbf{H}). \tag{6.109}$$

The last term on the right-hand side of this equality is exactly the first term in the square brackets of Eq. (108), so we may rewrite that formula as

$$\mathscr{P} = \int_V \left[-\nabla\cdot(\mathbf{E}\times\mathbf{H}) + \mathbf{H}\cdot(\nabla\times\mathbf{E}) - \mathbf{E}\cdot\frac{\partial\mathbf{D}}{\partial t} \right] d^3r. \tag{6.110}$$

[64] Let me emphasize that if this condition is *not* fulfilled, the lumped-circuit representation of the system (see Fig. 9 and its discussion) is typically inadequate – besides some special cases, to be discussed in the next chapter.

[65] See, e.g., MA Eq. (11.7) with **f** = **E** and **g** = **H**.

However, according to the Maxwell equation for $\nabla \times \mathbf{E}$, this curl is equal to $-\partial\mathbf{B}/\partial t$, so the second term in the square brackets of Eq. (110) equals $-\mathbf{H}\cdot\partial\mathbf{B}/\partial t$ and, according to Eq. (14), is just the (minus) time derivative of the magnetic energy per unit volume. Similarly, according to Eq. (3.76), the third term under the integral is the (minus) time derivative of the electric energy per unit volume. Finally, we can use the divergence theorem to transform the integral of the first term in the square brackets to a 2D integral over the surface S limiting the volume V. As a result, we get the so-called *Poynting theorem*[66] for the power balance in the system:

Poynting theorem

$$\int_V \left(\mathcal{P} + \frac{\partial u}{\partial t} \right) d^3 r + \oint_S S_n d^2 r = 0 . \tag{6.111}$$

Here u is the density of the total (electric plus magnetic) energy of the electromagnetic field, with

Field's energy variation

$$\delta u \equiv \mathbf{E} \cdot \delta\mathbf{D} + \mathbf{H} \cdot \delta\mathbf{B} \tag{6.112}$$

– just the sum of the expressions given by Eqs. (3.76) and (14). For the particular case of an isotropic, linear, and dispersion-free medium, with $\mathbf{D}(t) = \varepsilon\mathbf{E}(t)$, $\mathbf{B}(t) = \mu\mathbf{H}(t)$, Eq. (112) yields

Field's energy

$$u = \frac{\mathbf{E}\cdot\mathbf{D}}{2} + \frac{\mathbf{H}\cdot\mathbf{B}}{2} \equiv \frac{\varepsilon E^2}{2} + \frac{B^2}{2\mu} . \tag{6.113}$$

Another key notion participating in Eq. (111) is the *Poynting vector*, defined as[67]

Poynting vector

$$\mathbf{S} \equiv \mathbf{E} \times \mathbf{H} . \tag{6.114}$$

The first integral in Eq. (111) is evidently the net change of the energy of the system (particles + field) per unit time, so the second (surface) integral has to be the power flowing out from the system through the surface. As a result, it is tempting to interpret the Poynting vector **S** locally, as the power flow density at the given point. In many cases, such a local interpretation of vector **S** is legitimate; however, in other cases, it may lead to wrong conclusions. Indeed, let us consider the simple system shown in Fig. 11: a charged plane capacitor placed into a static and uniform external magnetic field, so that the electric and magnetic fields are mutually perpendicular.

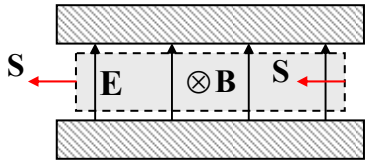


Fig. 6.11. The Poynting vector paradox.

In this static situation, with no charges moving, both $\mathcal{P}$ and $\partial/\partial t$ are equal to zero, and there should be no power flow in the system. However, Eq. (114) shows that the Poynting vector is not equal

[66] It is named after John Henry Poynting for his work published in 1884, though this fact was independently discovered by O. Heaviside in 1885 in a simpler form, while a similar result for the intensity of mechanical elastic waves had been obtained earlier (in 1874) by Nikolay Alekseevich Umov – see, e.g., CM Sec. 7.7.

[67] Actually, an addition to **S** of the curl of an arbitrary vector function $\mathbf{f}(\mathbf{r}, t)$ does not change Eq. (111). Indeed, we may use the divergence theorem to transform the corresponding change of the surface integral in Eq. (111) to a volume integral of scalar function $\nabla\cdot(\nabla\times\mathbf{f})$ that equals zero at any point – see, e.g., MA Eq. (11.2).

to zero inside the capacitor, being directed as the red arrows in Fig. 11 show. From the point of view of the only unambiguous corollary of the Maxwell equations, Eq. (111), there is no contradiction here, because the fluxes of the vector **S** through the side boundaries of the volume shaded in Fig. 11 are equal and opposite (and they are zero for other faces of this rectilinear volume), so the total flux of the Poynting vector through the volume boundary equals zero, as it should. It is, however, useful to recall this example each time before giving a local interpretation of the vector **S**.

The paradox illustrated in Fig. 11 is closely related to the *radiation recoil effects*, due to the electromagnetic field's momentum – more exactly, it *linear momentum*. Indeed, acting as at the Poynting theorem derivation, it is straightforward to use the *microscopic* Maxwell equations[68] to prove that, neglecting the boundary effects, the vector sum of the mechanical linear momentum of the particles in an arbitrary volume, and the integral of the following vector,

$$\mathbf{g} \equiv \frac{\mathbf{S}}{c^2}, \tag{6.115}$$

Electro-magnetic field's momentum

over the same volume, is conserved, enabling an interpretation of **g** as the density of the linear momentum of the electromagnetic field. (It will be more convenient for me to prove this relation, and discuss the related issues, in Sec. 9.8, using the 4-vector formalism of special relativity.) Due to this conservation, if some static fields coupled to mechanical bodies are suddenly decoupled from them and are allowed to propagate in space, i.e. to change their local integral of **g**, they give the bodies an equal and opposite impulse of force.

Finally, to complete our initial discussion of the Maxwell equations,[69] let us rewrite them in terms of potentials **A** and ϕ, because this is more convenient for the solution of some (though not all!) problems. Even when dealing with the system (99) of the more general Maxwell equations than discussed before, Eqs. (7) are still used for the definition of the potentials. It is straightforward to verify that with these definitions, the two homogeneous Maxwell equations (99b) are satisfied automatically. Plugging Eqs. (7) into the inhomogeneous equations (99a), and considering, for simplicity, a linear, uniform medium with frequency-independent ε and μ, we get

$$\nabla^2\phi + \frac{\partial}{\partial t}(\nabla\cdot\mathbf{A}) = -\frac{\rho}{\varepsilon}, \qquad \nabla^2\mathbf{A} - \varepsilon\mu\frac{\partial^2\mathbf{A}}{\partial t^2} - \nabla\left(\nabla\cdot\mathbf{A} + \varepsilon\mu\frac{\partial\phi}{\partial t}\right) = -\mu\mathbf{j}. \tag{6.116}$$

This is a more complex result than what we would like to get. However, let us select a special gauge, which is frequently called (especially for the free space case, when $v = c$) the *Lorenz gauge condition*[70]

$$\nabla\cdot\mathbf{A} + \varepsilon\mu\frac{\partial\phi}{\partial t} = 0, \tag{6.117}$$

Lorenz gauge condition

[68] The situation with the *macroscopic* Maxwell equations is more complex, and is still a subject of some lingering discussions (usually called the *Abraham-Minkowski controversy*, despite contributions by many other scientists including A. Einstein), because of the ambiguity of the momentum's division between its field and particle components – see, e.g., the review paper by R. Pfeiffer *et al.*, *Rev. Mod. Phys.* **79**, 1197 (2007).

[69] We will return to their general discussion (in particular, to the analytical mechanics of the electromagnetic field, and its stress tensor) in Sec. 9.8, after we get equipped with the special relativity theory.

[70] This condition, named after *Ludwig Lorenz*, should not be confused with the so-called *Lorentz invariance condition* of relativity, due to *Hendrik Lorentz*, to be discussed in Sec. 9.4. (Note the last names' spelling.)

which is a natural generalization of the Coulomb gauge (5.48) to time-dependent phenomena. With this condition, Eqs. (107) are reduced to a simpler, beautifully symmetric form:

Potentials' dynamics

$$\nabla^2\phi - \frac{1}{v^2}\frac{\partial^2\phi}{\partial t^2} = -\frac{\rho}{\varepsilon}, \qquad \nabla^2\mathbf{A} - \frac{1}{v^2}\frac{\partial^2\mathbf{A}}{\partial t^2} = -\mu\mathbf{j}, \tag{6.118}$$

where $v^2 \equiv 1/\varepsilon\mu$. Note that these equations are essentially a set of 4 similar equations for 4 scalar functions (namely, ϕ and three Cartesian components of $\mathbf{A}$) and thus clearly invite the 4-component vector formalism of the relativity theory; it will be discussed in Chapter 9.[71]

If ϕ and $\mathbf{A}$ depend on just one spatial coordinate, say z, then in a region without field sources: $\rho = 0$, $\mathbf{j} = 0$, Eqs. (118) are reduced to the following 1D *wave equations*

$$\frac{\partial^2\phi}{\partial z^2} - \frac{1}{v^2}\frac{\partial^2\phi}{\partial t^2} = 0, \qquad \frac{\partial^2\mathbf{A}}{\partial z^2} - \frac{1}{v^2}\frac{\partial^2\mathbf{A}}{\partial t^2} = 0. \tag{6.119}$$

It is well known[72] that these equations describe waves, with arbitrary waveforms (including sinusoidal waves of any frequency), propagating with the same speed v in either of the z-axis directions. According to the definitions of the constants ε_0 and μ_0, in free space, v is just the speed of light:

$$v = \frac{1}{(\varepsilon_0\mu_0)^{1/2}} \equiv c\,. \tag{6.120}$$

Historically, the experimental observation of relatively low-frequency (GHz-scale) electromagnetic waves, with their speed equal to that of light, was the decisive proof (actually, a real triumph!) of the Maxwell theory and his prediction of such waves.[73] This was first accomplished in 1886 by Heinrich Rudolf Hertz, using the electronic circuits and antennas he had invented for this purpose.

Before proceeding to the detailed analysis of these waves in the following chapters, let me mention that the invariance of Eqs. (119) with respect to the wave propagation direction is not occasional; it is just a manifestation of one more general property of the Maxwell equations (99), called the *Lorentz reciprocity*. We have already met its simplest example, for time-independent electrostatic fields, in one of the problems of Chapter 1. In a much more general case when two monochromatic electromagnetic fields of the same frequency, with complex amplitudes, say, $\{\mathbf{E}_1(\mathbf{r}), \mathbf{H}_1(\mathbf{r})\}$ and $\{\mathbf{E}_2(\mathbf{r}),$

---

[71] Here I have to mention in passing the so-called *Hertz vector potentials* $\mathbf{\Pi}_e$ and $\mathbf{\Pi}_m$ (whose introduction may be traced back at least to the 1904 work by E. Whittaker). They may be defined by the following relations:

$$\mathbf{A} = \mu\frac{\partial\mathbf{\Pi}_e}{\partial t} + \mu\nabla\times\mathbf{\Pi}_m, \qquad \phi = -\frac{1}{\varepsilon}\nabla\cdot\mathbf{\Pi}_e,$$

which make the Lorentz gauge condition (117) automatically satisfied. These potentials are especially convenient for the solution of problems in which the electromagnetic field is induced by sources characterized by field-independent electric and magnetic polarizations $\mathbf{P}$ and $\mathbf{M}$ – rather than by field-independent charge and current densities ρ and $\mathbf{j}$. Indeed, it is straightforward to check that both $\mathbf{\Pi}_e$ and $\mathbf{\Pi}_m$ satisfy the equations similar to Eqs. (118), but with their right-hand sides equal to, respectively, $-\mathbf{P}$ and $-\mathbf{M}$. Unfortunately, I would not have time/space to discuss such problems and have to refer interested readers elsewhere – for example, to a classical text by J. Stratton, *Electromagnetic Theory*, Adams Press, 2008.

[72] See, e.g., CM Secs. 6.3-6.4 and 7.7-7.8.

[73] By that time, the speed of light (estimated very reasonably by Ole Rømer as early as 1676) has been experimentally measured, by Hippolyte Fizeau and then Léon Foucault, with an accuracy better than 1%.

$\mathbf{H}_2(\mathbf{r})\}$ are induced, separately, by stand-alone currents with complex amplitudes $\mathbf{j}_1(\mathbf{r})$ and $\mathbf{j}_2(\mathbf{r})$ of their densities. Then it may be proved[74] that if the medium is linear and either isotropic or even anisotropic but with symmetric tensors $\varepsilon_{jj'}$ and $\mu_{jj'}$, then for any volume V limited by a closed surface S,

$$\int_V \left(\mathbf{j}_1 \cdot \mathbf{E}_2 - \mathbf{j}_2 \cdot \mathbf{E}_1\right) d^3r = \oint_S \left(\mathbf{E}_1 \times \mathbf{H}_2 - \mathbf{E}_2 \times \mathbf{H}_1\right)_n d^2r \,. \tag{6.121}$$

This property implies, in particular, that the waves propagate similarly in two reciprocal directions even in situations much more general than the 1D case described by Eqs. (119). For some important practical applications (e.g., for low-noise amplifiers and detectors) such reciprocity is rather inconvenient. Fortunately, Eq. (121) may be violated in anisotropic media with asymmetric tensors $\varepsilon_{jj'}$ and/or $\mu_{jj'}$. The simplest case of such an anisotropy, the *Faraday rotation* of the wave polarization in plasma, will be discussed in the next chapter.

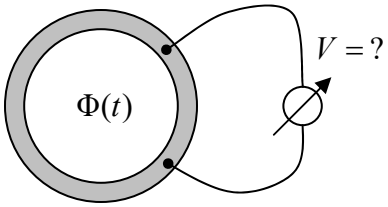
6.9. Exercise problems

6.1. Prove that the electromagnetic induction e.m.f. $\mathscr{V}_{\text{ind}}$ in a conducting loop may be measured as shown on two panels of Fig. 1:

(i) by measuring the current $I = \mathscr{V}_{\text{ind}}/R$ induced in the loop closed with an Ohmic resistor R, or

(ii) using a voltmeter inserted into the loop.

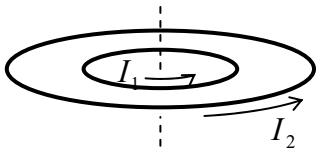
6.2. The flux Φ of the magnetic field that pierces a resistive ring is being changed in time, while the field outside of the ring is negligibly low. A voltmeter is connected to a part of the ring, as shown in the figure on the right. What would the voltmeter show?



6.3. A weak constant magnetic field **B** is applied to an axially-symmetric permanent magnet with the dipole magnetic moment **m** directed along its axis, rapidly rotating about the same axis, with an angular momentum **L**. Calculate the electric field resulting from the magnetic field's application, and formulate the conditions of your result's validity.

6.4. The similarity of Eq. (5.53) obtained in Sec. 5.3 without any use of the Faraday induction law, and Eq. (5.54) proved in Sec. 2 of this chapter using it, implies that the law may be derived from magnetostatics. Prove that this is indeed true for a particular case of a current loop being slowly deformed in a fixed magnetic field **B**(**r**).

6.5. Could Problem 5.2 (i.e. the semi-quantitative analysis of the mechanical stability of the system shown in the figure on the right) be solved using potential energy arguments?



[74] It will be more convenient for me to give this proof (or rather offer it for the reader's exercise :-) in the next chapter, after we have discussed the Fourier expansion of the fields in linear media.

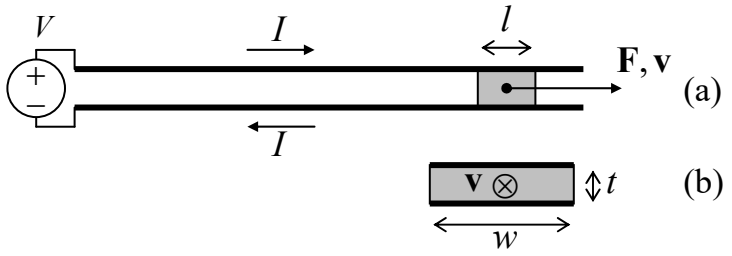
6.6. Use energy arguments to calculate the pressure exerted by the magnetic field **B** inside a long uniform solenoid of length l, and a cross-section of area $A << l^2$, with $N >> l/A^{1/2} >> 1$ turns, on its "walls" (windings), and the forces exerted by the field on the solenoid's ends, for two cases:

(i) the current through the solenoid is fixed by an external source, and

(ii) after the initial current setting, the ends of the solenoid's wire, with negligible resistance, are connected, so that it continues to carry a non-zero current.

Compare the results, and give a physical interpretation of the direction of these forces.

6.7. The *electromagnetic railgun* is a projectile launch system consisting of two long parallel conducting rails and a sliding conducting projectile shorting the current I fed into the system by a powerful source – see panel (a) in the figure on the right. Calculate the force exerted on the projectile, using two approaches:

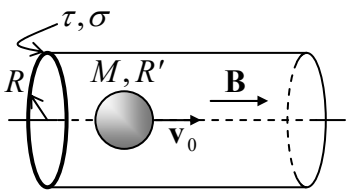


(i) by a direct calculation, assuming that the cross-section of the system has the simple shape shown on panel (b) of the figure above, with $t << w$, l, and

(ii) by using the energy balance (for simplicity, neglecting the Ohmic resistances in the system),

and compare the results.

6.8. A uniform, static magnetic field **B** is applied along the axis of a long thin pipe of a radius R and wall thickness $\tau << R$, made of a material with Ohmic conductivity σ. A sphere of mass M and radius $R' << R$, made of a linear magnetic material with permeability $\mu >> \mu_0$, is launched, with an initial velocity v_0, to fly ballistically along the pipe's axis – see the figure on the right. Use the quasistatic approximation to calculate the distance the sphere would pass before it stops. Formulate the conditions of validity of your result.



6.9. A planar thin-wire loop with inductance L, resistance R, and area A is launched to fly ballistically from field-free space into a region where the magnetic field **B** is constant. Calculate the final change of the kinetic energy of the loop, assuming that the time of its entry into the field region is much shorter than the relaxation time constant L/R and that the loop cannot rotate.

6.10. AC current of frequency ω is being passed through a long uniform wire with a round cross-section of a radius R comparable with the skin depth δ_s. In the quasistatic approximation, find the current's distribution across the cross-section, and analyze it in the limits $R << \delta_s$ and $\delta_s << R$. Calculate the effective ac resistance of the wire (per unit length) in these two limits.

6.11. A long round cylinder of radius R, made of a uniform conductor with an Ohmic conductivity σ and magnetic permeability μ, is placed into a uniform ac magnetic field $\mathbf{H}_{ext}(t) = \mathbf{H}_0 \cos\omega t$ directed along its symmetry axis. Calculate the spatial distribution of the magnetic field's amplitude and, in particular, its value on the cylinder's axis. Spell out the last result in the limits of relatively small and large R.

6.12.* Define and calculate an appropriate spatial-temporal Green's function for Eq. (25), and then use this function to analyze the dynamics of propagation of the external magnetic field that is suddenly turned on at $t = 0$ and then kept constant:

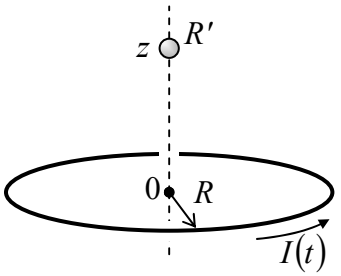
$$H(x<0,t)=\begin{cases}0, & \text{at } t<0,\\ H_0, & \text{at } t>0,\end{cases}$$

into an Ohmic conductor occupying the semi-space $x > 0$ – see Fig. 2.

Hint: Try to use a function proportional to $\exp\{-(x-x')^2/2(\delta x)^2\}$, with a suitable time dependence of the parameter δx and a properly selected pre-exponential factor.

6.13. Solve the previous problem using the variable separation method, and compare the results.

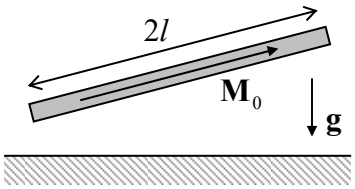
6.14. Calculate the average force exerted by ac current $I(t)$ of amplitude I_0, flowing in a planar round coil of radius R, on a conducting sphere with a much smaller radius R' (which is still much larger than the skin depth δ_s at the ac current's frequency), located on the loop's axis, at distance z from its center – see the figure on the right.



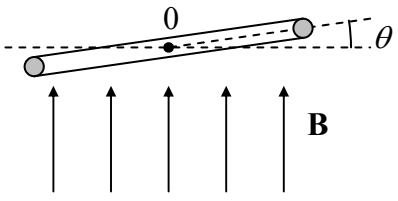
6.15. A small planar wire loop carrying current I is located relatively far from a planar surface of a superconductor. Within the coarse-grain (ideal-diamagnetic) description of the Meissner-Ochsenfeld effect, calculate:

(i) the energy of the loop-superconductor interaction,
(ii) the force and torque acting on the loop, and
(iii) the distribution of supercurrents on the superconductor surface.

6.16. A straight uniform magnet of length $2l$, cross-section area $A << l^2$, and mass m, with a permanent longitudinal magnetization M_0, is placed over a horizontal surface of a superconductor – see the figure on the right. Within the ideal-diamagnet description of superconductivity, find the stable equilibrium position of the magnet.



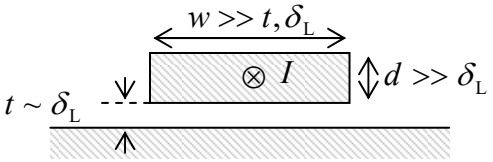
6.17. A plane superconducting wire loop of area A and inductance L may turn, without static friction, about a horizontal axis 0 (in the figure on the right, normal to the plane of the drawing) passing through its center of mass. Initially, the loop had been horizontal (with $\theta = 0$) and carried supercurrent I_0 in such a direction that its magnetic dipole vector had been directed down. Then a uniform magnetic field **B**, directed vertically up, was applied. Using the ideal-diamagnet description of the Meissner-Ochsenfeld effect, find all possible equilibrium positions of the loop, analyze their stability, and give a physical interpretation of the results.



6.18. Use the London equation to analyze the penetration of a uniform external magnetic field into a thin ($t \sim \delta_L$) planar superconducting film whose plane is parallel to the field.

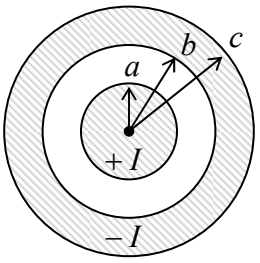
6.19. Use the London equation to calculate the distribution of supercurrent density **j** inside a long straight superconducting wire with a circular cross-section of radius $R \sim \delta_L$, carrying current I.

6.20. Use the London equation to calculate the inductance (per unit length) of a long uniform superconducting strip placed close to the surface of a similar superconductor – see the figure on the right, which shows the structure's cross-section.



6.21. Calculate the inductance (per unit length) of a superconducting cable with the round cross-section shown in the figure on the right, in the following limits:

(i) $\delta_L << a, b, c - b$, and
(ii) $a << \delta_L << b, c - b$.



6.22. Use the London equation to analyze the magnetic field shielding by a superconducting thin film of thickness $t << \delta_L$, by calculating the penetration of the field induced by current I in a thin wire that runs parallel to a wide planar thin film, at a distance $d >> t$ from it, into the space behind the film.

6.23. Assuming that the magnetic monopole does exist and has a magnetic charge q_m, calculate the change ΔI of current in a superconducting loop due to a passage of a single monopole through its area. Evaluate ΔI for a monopole with the charge conjectured by P. Dirac, $q_m = nq_0 \equiv n(2\pi\hbar/e)$ with an integer n, and compare the result with the magnetic flux quantum Φ_0 (62). Review your result for a similar passage of a single quasi-monopole magnetic charge formed at one of the ends of a permanent-magnet needle – see, e.g., Fig. 19 and the accompanying discussion.

Hint: To simplify calculations, you may consider the monopole's passage along the symmetry axis of a round ring of radius R, made of a superconducting wire with a cross-section's area A satisfying the conditions $\delta_L^2 << A << R^2$.

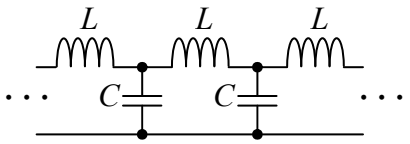
6.24. Use the Ginzburg-Landau equations (54) and (63) to calculate the largest ("critical") value of supercurrent in a uniform superconducting wire with a cross-section area much smaller than δ_L^2.

6.25. Use the discussion of a long straight Abrikosov vortex, in the limit $\xi << \delta_L$, in Sec. 5 to prove Eqs. (71)-(72) for its energy per unit length and the first critical field.

6.26.* Use the Ginzburg-Landau equations (54) and (63) to prove the Josephson relation (76) for a small superconducting weak link, and express its critical current I_c via the Ohmic resistance R_n of the same weak link in its normal state.

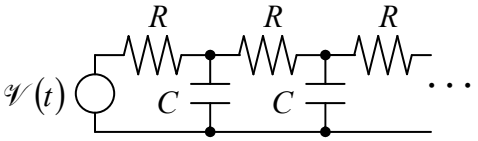
6.27. Use Eqs. (76) and (79) to calculate the coupling energy of a Josephson junction and the full potential energy of the SQUID shown in Fig. 4c.

6.28. Analyze the possibility of wave propagation in a long uniform chain of lumped inductances and capacitances – see the figure on the right.



Hint: Readers without prior experience in electromagnetic wave analysis may like to use a substantial analogy between this effect and mechanical waves in a 1D chain of elastically coupled particles.[75]

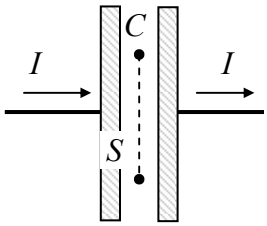
6.29. A sinusoidal e.m.f. of amplitude V_0 and frequency ω is applied to an end of a long chain of similar lumped resistors and capacitors, shown in the figure on the right. Calculate the law of decay of the ac voltage amplitude along the chain.



6.30. As was discussed in Sec. 7, the displacement current concept allows one to extend the Ampère law to time-dependent processes as

$$\oint_C \mathbf{H}\cdot d\mathbf{r} = I_S + \frac{\partial}{\partial t}\int_S D_n d^2 r\,.$$

We also have seen that this generalization makes the integral $\int \mathbf{H}\cdot d\mathbf{r}$ over an external contour, such as the one shown in Fig. 10, independent of the choice of the surface S limited by the contour. However, it may look like the situation is different for a contour drawn inside a capacitor – see the figure on the right. Indeed, if the contour's size is much larger than the capacitor's thickness, the magnetic field $\mathbf{H}$ created by the linear current I on the contour's line is virtually the same as that of a continuous wire, and hence the integral $\int \mathbf{H}\cdot d\mathbf{r}$ along the contour apparently does not depend on its area, while the magnetic flux $\int D_n d^2 r$ does, so the equation displayed above seems invalid. (The current I_S piercing this contour evidently equals zero.) Resolve this paradox, for simplicity considering an axially-symmetric system.



6.31. A straight, uniform, long wire with a circular cross-section of radius R, made of an Ohmic conductor with conductivity σ, carries dc current I. Calculate the flux of the Poynting vector through its surface, and compare it with the Joule rate of energy dissipation.

[75] See, e.g., CM Sec. 6.3.

Chapter 7. Electromagnetic Wave Propagation

This (rather extensive) chapter focuses on the most important effect that follows from the time-dependent Maxwell equations, namely the electromagnetic waves, at this stage avoiding the issue of their origin, i.e. of the wave radiation process – which will be the subject of Chapters 8 and 10. We will start from the simplest, plane waves in uniform and isotropic media, and then proceed to a discussion of nonuniform systems, bringing up such effects as reflection and refraction. Then we will discuss the so-called guided waves, propagating along various transmission lines – such as cables, waveguides, and optical fibers. Finally, the end of the chapter is devoted to final-length fragments of such lines, serving as resonant cavities, and to the effects of energy dissipation in transmission lines and cavities.

7.1. Plane waves

Let us start by considering a spatial region that does not contain field sources ($\rho = 0$, $\mathbf{j} = 0$), and is filled with a uniform, isotropic, linear medium, which therefore obeys Eqs. (3.46) and (5.110):

$$\mathbf{D} = \varepsilon\mathbf{E}, \qquad \mathbf{B} = \mu\mathbf{H}. \tag{7.1}$$

Moreover, let us assume for a while that these constitutive equations hold for all frequencies of interest. (Of course, these relations are *exactly* valid for the very important particular case of free space, where we may formally use the macroscopic Maxwell equations (6.100), but with $\varepsilon = \varepsilon_0$ and $\mu = \mu_0$.) As was already shown in Sec. 6.8, in this case, the Lorenz gauge condition (6.117) allows the Maxwell equations to be recast into the wave equations (6.118) for the scalar and vector potentials. However, for most purposes, it is more convenient to use the homogeneous Maxwell equations (6.100) for the electric and magnetic fields – which are independent of the gauge choice. After an elementary elimination of **D** and **B** using Eqs. (1),[1] these equations take a simple, very symmetric form:

Maxwell equations for uniform linear media

$$\nabla\times\mathbf{E} + \mu\frac{\partial\mathbf{H}}{\partial t} = 0, \qquad \nabla\times\mathbf{H} - \varepsilon\frac{\partial\mathbf{E}}{\partial t} = 0, \tag{7.2a}$$

$$\nabla\cdot\mathbf{E} = 0, \qquad \nabla\cdot\mathbf{H} = 0. \tag{7.2b}$$

Now, let us act by the operator $\nabla\times$ on each of Eqs. (2a), i.e. take their curl, and then use the vector algebra identity (5.31). The appearing terms $\nabla\cdot\mathbf{E}$ and $\nabla\cdot\mathbf{H}$ vanish due to Eqs. (2b), so the first terms of Eqs. (2a) turn into the Laplace operators of these vectors (with the minus sign). Now swapping, in the second terms, the operators $\partial/\partial t$ and $\nabla\times$, and using Eqs. (2a) again, we get fully similar wave equations for the electric and magnetic fields:[2]

EM wave equations

$$\left(\nabla^2 - \frac{1}{v^2}\frac{\partial^2}{\partial t^2}\right)\mathbf{E} = 0, \qquad \left(\nabla^2 - \frac{1}{v^2}\frac{\partial^2}{\partial t^2}\right)\mathbf{H} = 0, \tag{7.3}$$

[1] Though in a medium, **B** rather than **H** is the actual macroscopic magnetic field, mathematically it is a bit more convenient (just as it was in Sec. 6.2) to use the vector pair {**E**, **H**} in the following discussion, because at sharp media boundaries, it is **H** that obeys the boundary condition (5.117) similar to that for **E** – cf. Eq. (3.37).

[2] The two vector equations (3) are of course just a shorthand for six similar equations for three Cartesian components of **E** and **H**.

where the parameter v is defined as

$$v^2 \equiv \frac{1}{\varepsilon\mu}. \tag{7.4}$$

Wave velocity

with $v^2 = 1/\varepsilon_0\mu_0 \equiv c^2$ in free space – see Eq. (6.120) again.

These equations allow, in particular, solutions of the following type;

$$E \propto H \propto f(z - vt), \tag{7.5}$$

Plane wave

where z is the Cartesian coordinate along a certain (*arbitrary*) direction **n**, and f is an *arbitrary* function of one argument. Note that this solution, first of all, describes a *traveling wave* – meaning a certain field pattern moving, without deformation, along the z-axis, with the constant velocity v. Second, according to Eq. (5), both **E** and **H** have the same values at all points of each plane perpendicular to the direction **n** ≡ $\mathbf{n}_z$ of the wave propagation; hence the second name – *plane wave*.

According to Eqs. (2), the independence of the wave *equations* (3) for vectors **E** and **H** does not mean that their plane-wave *solutions* are independent. Indeed, plugging any solution of the type (5) into Eqs. (2a), we get

$$\mathbf{H} = \frac{\mathbf{n}\times\mathbf{E}}{Z}, \qquad \text{i.e. } \mathbf{E} = Z\,\mathbf{H}\times\mathbf{n}, \tag{7.6}$$

Field vector relation

where

$$Z \equiv \frac{E}{H} = \left(\frac{\mu}{\varepsilon}\right)^{1/2}. \tag{7.7}$$

Wave impedance

The vector relationship (6) means, first of all, that at any point of space and at any time instant, the vectors **E** and **H** are perpendicular not only to the propagation vector **n** (such waves are called *transverse*) but also to each other — see Fig. 1.

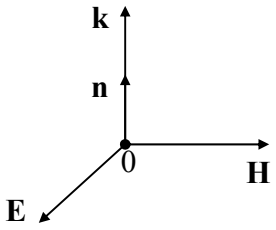


Fig. 7.1. Field vectors in a plane electromagnetic wave propagating along direction **n**.

Second, this equality does not depend on the function f, meaning that the electric and magnetic fields increase and decrease *simultaneously*. Finally, the field magnitudes are related by the constant Z called the *wave impedance* of the medium. Very soon we will see that this impedance plays a pivotal role in many problems, in particular at the wave reflection from the interface between two media. Since the dimensionality of E, in SI units, is V/m, and that of H is A/m, Eq. (7) shows that Z has the dimensionality of V/A, i.e. ohms (Ω).[3] In particular, in free space,

[3] In the Gaussian units, E and H have a similar dimensionality (in particular, in a free-space wave, $E = H$), making the (very useful) notion of the wave impedance less manifestly exposed – so in some older physics textbooks it is not mentioned at all!

Wave impedance of free space

$$Z = Z_0 \equiv \left(\frac{\mu_0}{\varepsilon_0}\right)^{1/2} = 4\pi \times 10^{-7} c \approx 377\ \Omega\,. \tag{7.8}$$

Next, plugging Eq. (6) into Eqs. (6.113) and (6.114), we get:

Wave's energy

$$u = \varepsilon E^2 = \mu H^2\,, \tag{7.9a}$$

Wave's power

$$\mathbf{S} \equiv \mathbf{E} \times \mathbf{H} = \mathbf{n}\frac{E^2}{Z} = \mathbf{n} Z H^2\,, \tag{7.9b}$$

so, according to Eqs. (4) and (7), the wave's energy and power densities are universally related as

$$\mathbf{S} = \mathbf{n} u v\,. \tag{7.9c}$$

In view of the Poynting vector paradox discussed in Sec. 6.8 (see Fig. 6.11), one may wonder whether the last equality may be interpreted as the actual density of power flow. In contrast to the static situation shown in Fig. 6.11, which limits the electric and magnetic fields to the vicinity of their sources, waves may travel far from them. As a result, they can form *wave packets* of a finite length in free space – see Fig. 2.

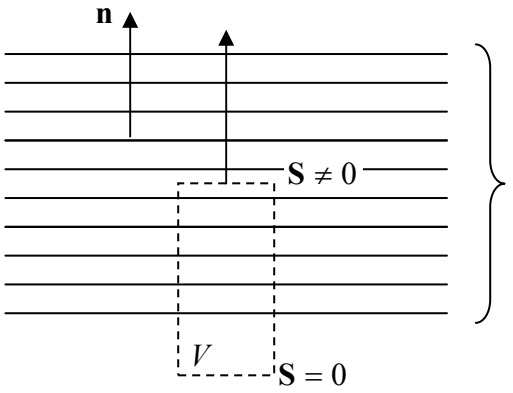


Fig. 7.2. Interpreting the Poynting vector in a plane electromagnetic wave. (Horizontal lines show equal-field planes.)

Let us apply the Poynting theorem (6.111) to the cylinder shown with dashed lines in Fig. 2, with one lid inside the wave packet, and another lid in the region already passed by the wave. Then, according to Eq. (6.111), the rate of change of the full field energy $\mathscr{E}$ inside the volume is $d\mathscr{E}/dt = -SA$ (where A is the lid area), so S may be indeed interpreted as the power flow (per unit area) from the volume. Making a reasonable assumption that the finite length of a sufficiently long wave packet does not affect the physics inside it, we may indeed interpret the **S** given by Eqs. (9b-c) as the power flow density inside a plane electromagnetic wave.

As we will see later in this chapter, the free-space value Z_0 of the wave impedance, given by Eq. (8), establishes the scale of Z of virtually all wave transmission lines, so we may use it, together with Eq. (9), to get a better feeling of how much different are the electric and magnetic field amplitudes in the waves – on the scale of typical electrostatics and magnetostatics experiments. For example, according to Eqs. (9), a wave of a modest intensity $S = 1$ W/m^2 (this is what we get from a usual electric bulb a few meters away from it) has $E \sim (SZ_0)^{1/2} \sim 20$ V/m, quite comparable with the dc field created by a standard AA battery right outside it. On the other hand, the wave's magnetic field $H = (S/Z_0)^{1/2} \approx 0.05$ A/m. For this particular case, the relation following from Eqs. (1), (4), and (7),

$$B = \mu H = \mu \frac{E}{Z} = \mu \frac{E}{(\mu/\varepsilon)^{1/2}} \equiv (\varepsilon\mu)^{1/2} E = \frac{E}{v}, \tag{7.10}$$

gives $B = \mu_0 H = E/c \sim 7\times10^{-8}$T, i.e. a magnetic field a thousand times lower than the Earth's field, and about 7 orders of magnitude lower than the field of a typical permanent magnet. This huge difference may be interpreted as follows: the scale $B \sim E/c$ of magnetic fields in the waves is "normal" for electromagnetism, while the permanent magnet fields are abnormally high because they are due to the ferromagnetic alignment of electron spins, essentially relativistic objects – see the discussion in Sec. 5.5.

The fact that Eq. (5) is valid for an arbitrary function f means, in the standard terminology, that a medium with frequency-independent ε and μ supports the propagation of plane waves without either decay (*attenuation*) or waveform deformation (*dispersion*). However, for any real medium but pure vacuum, this approximation is valid only within limited frequency intervals. We will discuss the effects of attenuation and dispersion in the next section and will see that all our prior formulas remain valid even for an arbitrary linear media, provided that we limit them to single-frequency (i.e. sinusoidal, frequently called *monochromatic*) waves. Such waves may be most conveniently represented as[4]

$$f = \mathrm{Re}\left[f_\omega e^{i(kz - \omega t)} \right], \tag{7.11}$$

Mono-chromatic wave

where f_ω is the *complex amplitude* of the wave, and k is its *wave number* (the magnitude of the *wave vector* $\mathbf{k} \equiv \mathbf{n}k$), sometimes called the *spatial frequency*. The last term is justified by the fact, evident from Eq. (11), that k is related to the wavelength λ exactly as the usual ("temporal") frequency ω is related to the time period $\mathcal{T}$:

$$k = \frac{2\pi}{\lambda}, \qquad \omega = \frac{2\pi}{\mathcal{T}}. \tag{7.12}$$

Spatial and temporal frequencies

In the dispersion-free case (5), the compatibility of that relation with Eq. (11) requires the argument $(kz - \omega t) \equiv k\,[z - (\omega/k)t]$ to be proportional to $(z - vt)$, so $\omega/k = v$, i.e.

$$k = \frac{\omega}{v} \equiv (\varepsilon\mu)^{1/2}\omega, \tag{7.13}$$

Dispersion relation

so in that particular case, the *dispersion relation* $\omega(k)$ is linear.

Now note that Eq. (6) does not mean that the vectors **E** and **H** retain their direction in space. (The wave in which they do, is called *linearly polarized*.[5]) Indeed, nothing in the Maxwell equations prevents, for example, a joint rotation of this vector pair around the fixed vector **n**, while still keeping all these three vectors perpendicular to each other at any instant – see Fig. 1. However, an arbitrary rotation law or even an arbitrary constant frequency of such rotation would violate the single-frequency (monochromatic) character of the elementary sinusoidal wave (11). To understand what is the most general type of polarization the wave may have without violating that condition, let us represent two

[4] As we have already seen in the previous chapter (see also CM Sec. 5.1), such complex-exponential representation of sinusoidally changing variables is more convenient for mathematical manipulation than by using sine and cosine functions, especially because in all linear relations, the operator Re may be omitted (implied) until the very end of the calculation. Note, however, that this is *not* valid for the quadratic forms such as Eqs. (9).

[5] The possibility of different polarizations of electromagnetic waves was discovered (for light) in 1699 by Rasmus Bartholin, a.k.a. Erasmus Bartholinus.

Cartesian components of one of these vectors (say, **E**) along any two fixed axes *x* and *y*, perpendicular to each other and the *z*-axis (i.e. to the vector **n)**, in the same form as used in Eq. (11):

$$E_x = \mathrm{Re}\left[E_{\omega x} e^{i(kz - \omega t)} \right], \qquad E_y = \mathrm{Re}\left[E_{\omega y} e^{i(kz - \omega t)} \right]. \tag{7.14}$$

To keep the wave monochromatic, the complex amplitudes $E_{\omega x}$ and $E_{\omega y}$ have to be constant in time; however, they may have different magnitudes and an arbitrary phase shift between them.

In the simplest case when the arguments of these complex amplitudes are equal,

$$E_{\omega x,y} = \left| E_{\omega x,y} \right| e^{i\varphi}. \tag{7.15}$$

the real field components have the same phase:

$$E_x = \left| E_{\omega x} \right| \cos(kz - \omega t + \varphi), \qquad E_y = \left| E_{\omega y} \right| \cos(kz - \omega t + \varphi), \tag{7.16}$$

so their ratio is constant in time – see Fig. 3a. This means that the wave is linearly polarized, with the polarization plane defined by the relation

$$\tan\theta = \left| E_{\omega y} \right| / \left| E_{\omega x} \right|. \tag{7.17}$$

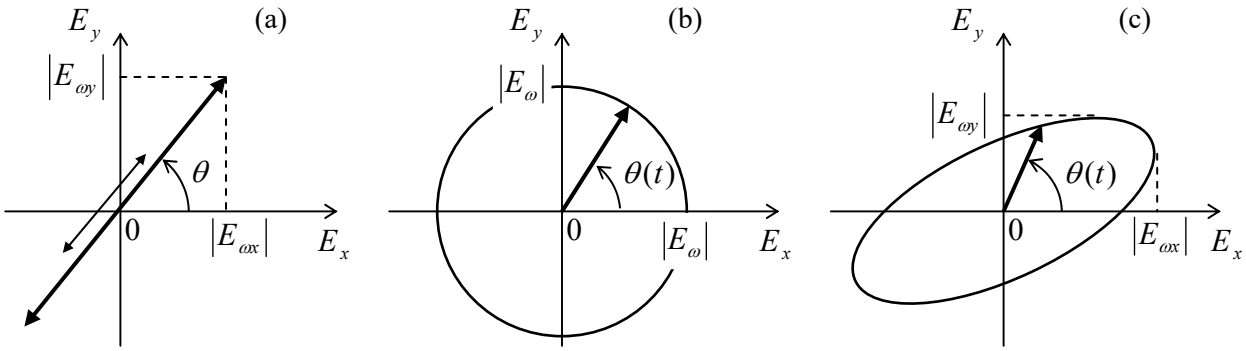


Fig. 7.3. Time evolution of the instantaneous electric field vector in monochromatic waves with: (a) a linear polarization, (b) a circular polarization, and (c) an elliptical polarization.

Another simple case is when the moduli of the complex amplitudes $E_{\omega x}$ and $E_{\omega y}$ are equal, but their phases are shifted by $+\pi/2$ or $-\pi/2$:

$$E_{\omega x} = \left| E_\omega \right| e^{i\varphi}, \quad E_{\omega y} = \left| E_\omega \right| e^{i(\varphi \pm \pi/2)}. \tag{7.18}$$

In this case

$$E_x = \left| E_\omega \right| \cos(kz - \omega t + \varphi), \quad E_y = \left| E_\omega \right| \cos\left(kz - \omega t + \varphi \pm \frac{\pi}{2} \right) \equiv \mp \left| E_\omega \right| \sin(kz - \omega t + \varphi). \tag{7.19}$$

This means that on the wave's plane (normal to **n**), the end of the vector **E** moves, with the wave's frequency ω, either clockwise or counterclockwise around a circle – see Fig. 3b:

$$\theta(t) = \mp(\omega t - \varphi). \tag{7.20}$$

Such waves are called *circularly polarized*. In the dominant convention, the wave is called *right-polarized* (RP) if it is described by the lower sign in Eqs. (18)-(20), i.e. if the vector **ω** of the angular frequency of the field vector's rotation coincides with the wave propagation's direction **n**, and *left-polarized* (LP) in the opposite case. These particular solutions of the Maxwell equations are very convenient for quantum electrodynamics, because single electromagnetic field quanta with a certain (positive or negative) spin direction may be considered elementary excitations of the corresponding circularly polarized wave.[6] (This fact does not exclude, from the quantization scheme, waves of other polarizations, because any monochromatic wave may be presented as a linear combination of two opposite circularly polarized waves – just as Eqs. (14) represent it as a linear combination of two linearly polarized waves.)

Finally, in the general case of arbitrary complex amplitudes $E_{\omega x}$ and $E_{\omega y}$, the field vector's end moves along an ellipse (Fig. 3c); such a wave is called *elliptically polarized*. The elongation ("eccentricity") and orientation of the ellipse are completely described by one complex number, the ratio $E_{\omega x}/E_{\omega y}$, i.e. by two real numbers – for example, $|E_{\omega x}/E_{\omega y}|$ and $\varphi = \arg(E_{\omega x}/E_{\omega y})$.[7]

7.2. Attenuation and dispersion

Let me start the discussion of the dispersion and attenuation effects by considering a particular case of the time evolution of the electric polarization $\mathbf{P}(t)$ of a dilute, non-polar medium, with negligible interaction between its elementary dipoles $\mathbf{p}(t)$. As was discussed in Sec. 3.3, in this case, the local electric field acting on each elementary dipole, may be taken equal to the macroscopic field $\mathbf{E}(t)$. Then, the dipole moment $\mathbf{p}(t)$ may be caused not only by the values of the field $\mathbf{E}$ at the same moment of time (t), but also by those at the earlier moments $t' < t$. Due to the linear superposition principle, the macroscopic polarization $\mathbf{P}(t) = n\mathbf{p}(t)$ should be a sum (practically, an integral) of the values of $\mathbf{E}(t')$ at all moments $t' \le t$, weighed by some function of t and t':[8]

$$P(t) = \int_{-\infty}^{t} E(t')G(t,t')dt'. \qquad (7.21)$$

Temporal Green's function

[6] This issue is closely related to that of the wave's angular momentum; it will be more convenient for me to discuss it later in this chapter (in Sec. 7).

[7] Note that the same information may be expressed via four so-called *Stokes parameters* s_0, s_1, s_2, and s_3, which are popular in practical optics, because they may be used for the description of not only completely coherent waves that were discussed here but also of party coherent or even fully incoherent waves – including the *natural light* emitted by thermal sources such as our Sun. (In contrast to the coherent waves (14), whose complex amplitudes are deterministic numbers, the amplitudes of incoherent waves should be treated as random variables.) For more on the Stokes parameters, as well as many other optics topics I will not have time to cover, I can recommend the classical text by M. Born *et al.*, *Principles of Optics*, 7th ed., Cambridge U. Press, 1999.

[8] In an isotropic media, the vectors **E**, **P**, and hence $\mathbf{D} = \varepsilon_0\mathbf{E} + \mathbf{P}$, are all parallel, and for notation simplicity, I will drop the vector sign in the following formulas of this section. I am also assuming that **P** at any point **r** is only dependent on the electric field at the same point, and hence drop the factor $\exp\{ikz\}$, the same for all variables. This last assumption is valid if the wavelength λ is much larger than the elementary dipole's size a. In most systems of interest, the scale of a is atomic ($\sim 10^{-10}$m), so this approximation is valid up to extremely high frequencies, $\omega \sim c/a \sim 10^{18}$ s^{-1}, corresponding to hard X-rays.

The condition $t' \le t$, which is implied by this relation, expresses a keystone principle of all science, the *causal relation* between a cause (in our case, the electric field $E(t')$ applied to each dipole) and its effect (the polarization $P(t)$ it creates). The function $G(t, t')$ is called the *temporal Green's function* for the electric polarization.[9] To reveal its physical sense, let us consider the case when the applied field $E(t)$ is a very short pulse at the moment $t_0 < t$, which may be well approximated with Dirac's delta function:

$$E(t) = \delta(t - t'') . \tag{7.22}$$

Then Eq. (21) yields just $P(t) = G(t, t'')$, so the Green's function $G(t, t')$ is just the polarization at moment t, created by a unit δ-functional pulse of the applied field at moment t' (Fig. 4).

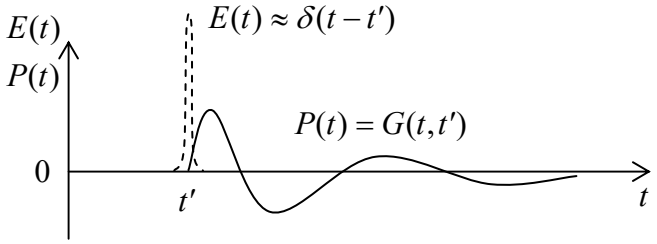


Fig. 7.4. An example of the temporal Green's function for the electric polarization (schematically).

What are the general properties of the temporal Green's function? First, for systems without infinite internal "memory", G should tend to zero at $t - t' \to \infty$, although the type of this approach (e.g., whether the function G oscillates approaching zero, as in Fig. 4, or not) depends on the medium's properties. Second, if the parameters of the medium do not change in time, the polarization response to an electric field pulse should be dependent not on its absolute timing, but only on the time difference $\theta \equiv t - t'$ between the pulse and observation instants, when Eq. (21) is reduced to

$$P(t) = \int_{-\infty}^{t} E(t')G(t - t')dt' \equiv \int_{0}^{\infty} E(t - \theta)G(\theta)d\theta . \tag{7.23}$$

For a sinusoidal waveform, $E(t) = \mathrm{Re}\,[E_\omega e^{-i\omega t}]$, this equation yields

$$P(t) = \mathrm{Re}\int_{0}^{\infty} E_\omega e^{-i\omega(t-\theta)}G(\theta)d\theta \equiv \mathrm{Re}\left[\left(E_\omega \int_{0}^{\infty} G(\theta)e^{i\omega\theta}d\theta\right)e^{-i\omega t}\right]. \tag{7.24}$$

The expression in the last parentheses is of course nothing else than the complex amplitude P_ω of the polarization. This means that though even if the static linear relation (3.43), $P = \chi_e\varepsilon_0 E$, is invalid for an arbitrary time-dependent process, we may still keep its Fourier analog,

$$P_\omega = \chi_e(\omega)\varepsilon_0 E_\omega, \qquad \text{with } \chi_e(\omega) \equiv \frac{1}{\varepsilon_0}\int_{0}^{\infty} G(\theta)e^{i\omega\theta}d\theta , \tag{7.25}$$

for each sinusoidal component of the process, using it as the definition of the frequency-dependent electric susceptibility $\chi_e(\omega)$. Similarly, the frequency-dependent electric permittivity may be defined using the Fourier analog of Eq. (3.46):

[9] The idea of these functions is very similar to that of the spatial Green's functions (see Sec. 2.10), but with a new twist, due to the causality principle. A discussion of the temporal Green's functions in application to classical mechanics (which to some extent overlaps with our current discussion) may be found in CM Sec. 5.1.

$$D_\omega \equiv \varepsilon(\omega) E_\omega . \tag{7.26a}$$

Complex electric permittivity

Then, according to Eq. (3.47), the complex permittivity is related to the temporal Green's function by the usual Fourier transform:

$$\varepsilon(\omega) \equiv \varepsilon_0 + \frac{P_\omega}{E_\omega} = \varepsilon_0 + \int_0^\infty G(\theta) e^{i\omega\theta} d\theta . \tag{7.26b}$$

This relation shows that the function $\varepsilon(\omega)$ may be complex,

$$\varepsilon(\omega) = \varepsilon'(\omega) + i\varepsilon''(\omega), \quad \text{with } \varepsilon'(\omega) = \varepsilon_0 + \int_0^\infty G(\theta)\cos\omega\theta\, d\theta, \quad \varepsilon''(\omega) = \int_0^\infty G(\theta)\sin\omega\theta\, d\theta , \tag{7.27}$$

and that its real part $\varepsilon'(\omega)$ is always an even function of frequency, while the imaginary part $\varepsilon''(\omega)$ is an odd function of ω. Note that though the particular causal relationship (21) between $P(t)$ and $E(t)$ is conditioned by the elementary dipole independence, the frequency-dependent complex electric permittivity $\varepsilon(\omega)$ may be introduced, in a similar way, if *any* two linear combinations of these variables are related by a similar formula.

Absolutely similar arguments show that magnetic properties of a linear, isotropic medium may be characterized by a frequency-dependent, complex permeability $\mu(\omega)$. Now rewriting Eqs. (1) for the complex amplitudes of the fields at a particular frequency, we may readily repeat all calculations of Sec. 1, and verify that all its results are valid for monochromatic waves even for a dispersive (but necessarily linear!) medium. In particular, Eqs. (7) and (13) now become

$$Z(\omega) = \left(\frac{\mu(\omega)}{\varepsilon(\omega)}\right)^{1/2}, \qquad k(\omega) = \omega\left[\varepsilon(\omega)\mu(\omega)\right]^{1/2}, \tag{7.28}$$

Complex Z and k

so the wave impedance and the wave number may be both complex functions of frequency.[10]

This fact has important consequences for electromagnetic wave propagation. First, plugging the representation of the complex wave number as the sum of its real and imaginary parts, $k(\omega) \equiv k'(\omega) + ik''(\omega)$, into Eq. (11):

$$f = \mathrm{Re}\left\{f_\omega e^{i[k(\omega)z - \omega t]}\right\} = e^{-k''(\omega)z}\,\mathrm{Re}\left\{f_\omega e^{i[k'(\omega)z - \omega t]}\right\}, \tag{7.29}$$

we see that $k''(\omega)$ describes the rate of wave *attenuation* in the medium at frequency ω.[11] Second, if the waveform is not sinusoidal (and hence should be represented as a sum of several/many sinusoidal components), the frequency dependence of $k'(\omega)$ provides for wave *dispersion*, i.e. the waveform deformation at the propagation, because the propagation velocity (4) of component waves is now different.[12]

---

[10] The first unambiguous observations of dispersion (for the case of light refraction) were described by Sir Isaac Newton in his *Optics* (1704) – even though this genius has never recognized the wave nature of light!

[11] It may be tempting to attribute this effect to wave *absorption*, i.e. the dissipation of the wave's energy, but we will see very soon that wave attenuation may be due to different effects as well.

[12] The reader is probably familiar with the most noticeable effect of the dispersion: the difference between the *group velocity* $v_{gr} \equiv d\omega/dk'$ giving the speed of the envelope of a wave packet with a narrow frequency spectrum, and the *phase velocity* $v_{ph} \equiv \omega/k'$ of the component waves. The second-order dispersion effect, proportional to $d^2\omega/d^2k'$, leads to the deformation (gradual broadening) of the envelope itself. Following tradition, these effects

As an example of such a dispersive medium, let us consider a simple but very representative *Lorentz oscillator model*.[13] In dilute atomic or molecular systems (e.g., gases), electrons respond to the external electric field especially strongly when its frequency ω is close to certain frequencies ω_j corresponding to the spectrum of quantum interstate transitions of a single atom/molecule. A phenomenological description of this behavior may be obtained from a classical model of several externally driven harmonic oscillators, generally with non-zero damping. For a single oscillator, driven by the electric field's force $F(t) = qE(t)$, we can write the 2nd Newton law as

$$m\left(\ddot{x} + 2\delta_0\dot{x} + \omega_0^2 x\right) = qE(t), \tag{7.30}$$

where ω_0 is the own frequency of the oscillator, and δ_0 is its damping coefficient. For the electric field of a monochromatic wave, $E(t) = \mathrm{Re}\,[E_\omega \exp\{-i\omega t\}]$, we may look for a particular, *forced-oscillation* solution of this equation in a similar form $x(t) = \mathrm{Re}\,[x_\omega \exp\{-i\omega t\}]$.[14] Plugging this solution into Eq. (30), we readily find the complex amplitude of these oscillations:

$$x_\omega = \frac{q}{m}\frac{E_\omega}{\left(\omega_0^2 - \omega^2\right) - 2i\omega\delta_0}. \tag{7.31}$$

Using this result to calculate the complex amplitude of the dipole moment as $p_\omega = qx_\omega$, and then the electric polarization $P_\omega = np_\omega$ of a dilute medium with n independent oscillators for unit volume, for its frequency-dependent permittivity (26) we get

Lorentz oscillator model

$$\varepsilon(\omega) = \varepsilon_0 + n\frac{q^2}{m}\frac{1}{\left(\omega_j^2 - \omega^2\right) - 2i\omega\delta_0}. \tag{7.32}$$

This result may be readily generalized to the case when the system has several types of oscillators with different masses and frequencies:

$$\varepsilon(\omega) = \varepsilon_0 + nq^2\sum_j \frac{f_j}{m_j\left[\left(\omega_j^2 - \omega^2\right) - 2i\omega\delta_j\right]}, \tag{7.33}$$

where $f_j \equiv n_j/n$ is the fraction of oscillators with frequency ω_j, so the sum of all f_j equals 1. Figure 5 shows a typical behavior of the real and imaginary parts of the complex dielectric constant, described by Eq. (33), as functions of frequency. The oscillator resonances' effect is clearly visible, and dominates the media response at $\omega \approx \omega_j$, especially in the case of low damping, $\delta_j << \omega_j$. Note that in the low-damping limit, the imaginary part of the dielectric constant ε'', and hence the wave attenuation k'', are negligibly small at all frequencies besides small vicinities of ω_j, where the derivative $d\varepsilon'(\omega)/d\omega$ is

---

are discussed in more detail in the quantum-mechanics part of this series (QM Sec. 2.2), because they are a crucial factor of Schrödinger's wave mechanics. (See also a brief discussion in CM Sec. 6.3.)

[13] This example is focused on the frequency dependence of ε rather than μ, because electromagnetic waves interact with "usual" media via their electric field much more than via the magnetic field. Indeed, according to Eq. (7), the magnetic field of the wave is of the order of E/c, so the magnetic component of the Lorentz force (5.10), acting on a non-relativistic particle, $F_m \sim quB \sim (u/c)qE$, is much smaller than that of its electric component, $F_e = qE$, and may be neglected. However, as will be discussed in Sec. 6, forgetting about the possible dispersion of $\mu(\omega)$ may result in missing some remarkable opportunities for manipulating the waves.

[14] If this point and Eq. (30) are not absolutely clear, please see CM Sec. 5.1 for a more detailed discussion.

negative.[15] Thus, for a system of weakly-damped oscillators, Eq. (33) may be well approximated by a sum of singularities ("poles"):

$$\varepsilon(\omega) \approx \varepsilon_0 + n\frac{q^2}{2}\sum_j \frac{f_j}{m_j\omega_j(\omega_j - \omega)}, \qquad \text{for } \delta_j << \left|\omega - \omega_j\right| << \left|\omega_j - \omega_{j'}\right|. \tag{7.34}$$

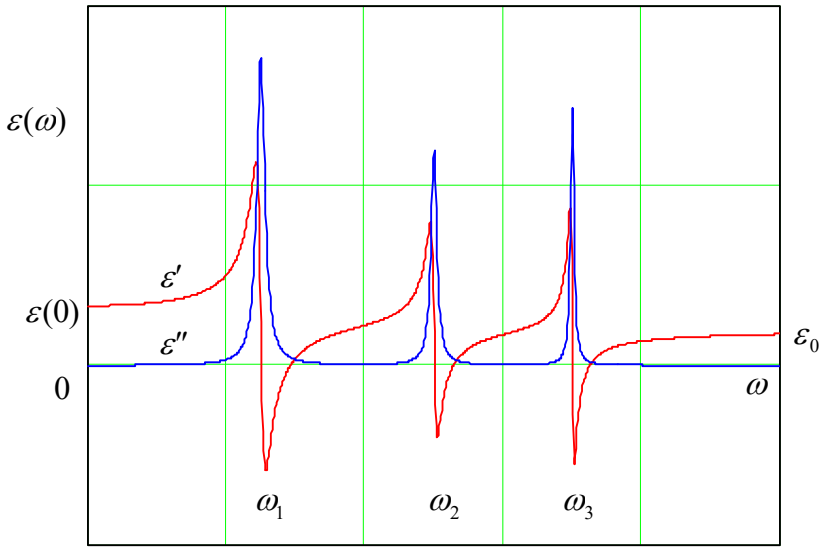


Fig. 7.5. Typical frequency dependence of the real and imaginary parts of the complex electric permittivity, according to the generalized Lorentz oscillator model.

This result is especially important because according to quantum mechanics,[16] Eq. (34) (with all m_j equal) is also valid for a set of non-interacting, similar quantum systems (whose dynamics may be completely different from that of a harmonic oscillator!), provided that ω_j are replaced with frequencies of possible quantum interstate transitions, and coefficients f_j are replaced with the so-called *oscillator strengths* of the transitions – which obey the same *sum rule*, $\Sigma_j f_j = 1$.

At $\omega \to 0$, the imaginary part of the complex permittivity (33) also vanishes (for any δ_j), while its real part approaches its electrostatic ("dc") value

$$\varepsilon(0) = \varepsilon_0 + q^2\sum_j \frac{n_j}{m_j\omega_j^2}. \tag{7.35}$$

Note that according to Eq. (30), the denominator of the fraction in Eq. (35) is just the effective spring constant $\kappa_j = m_j\omega_j^2$ of the j^{th} oscillator, so the oscillator masses m_j as such are actually (and quite naturally) not involved in the static dielectric response.

In the opposite limit of very high frequencies, $\omega >> \omega_j, \delta_j$, the permittivity also becomes real and may be represented as

$$\varepsilon(\omega) = \varepsilon_0\left(1 - \frac{\omega_p^2}{\omega^2}\right), \qquad \text{where } \omega_p^2 \equiv \frac{q^2}{\varepsilon_0}\sum_j \frac{n_j}{m_j}. \tag{7.36}$$

$\varepsilon(\omega)$ in plasma

[15] In optics, such behavior is called *anomalous dispersion*.

[16] See, e.g., QM Chapters 5-6.

This result is very important because it is also valid at *all* frequencies if all ω_j and δ_j vanish, for example for gases of free charged particles, in particular for *plasmas* – ionized atomic gases, provided that the ion collision effects are negligible. (This is why the parameter ω_p defined by Eq. (36) is called the *plasma frequency*.) Typically, the plasma as a whole is neutral, i.e. the density n of positive atomic ions is equal to that of the free electrons. Since the ratio n_j/m_j for electrons is much higher than that for ions, the general formula (36) for the plasma frequency is usually well approximated by the following simple expression:

$$\omega_p^2 = \frac{ne^2}{\varepsilon_0 m_e}. \tag{7.37}$$

This expression has a simple physical sense: the effective spring constant $\kappa_{ef} \equiv m_e\omega_p^2 = ne^2/\varepsilon_0$ describes the Coulomb force that appears when the electron subsystem of the plasma is shifted, as a whole, from its positive-ion subsystem, thus violating the electroneutrality.[17] Hence, there is no surprise that the function $\varepsilon(\omega)$ given by Eq. (36) vanishes at $\omega = \omega_p$: at this resonance frequency, the polarization electric field **E** may oscillate, i.e. have a non-zero amplitude $E_\omega = D_\omega/\varepsilon(\omega)$, even in the absence of external forces induced by external (stand-alone) charges, i.e. in the absence of the field **D** these charges induce – see Eq. (3.32).

The behavior of electromagnetic waves in a medium that obeys Eq. (36), is very remarkable. If the wave frequency ω is above ω_p, the dielectric constant $\varepsilon(\omega)$ and hence the wave number (28) are positive and real, and waves propagate without attenuation, following the dispersion relation,

Plasma dispersion relation

$$k(\omega) = \omega\left[\varepsilon(\omega)\mu_0\right]^{1/2} = \frac{1}{c}\left(\omega^2 - \omega_p^2\right)^{1/2}, \tag{7.38}$$

whose plot is shown in Fig. 6.

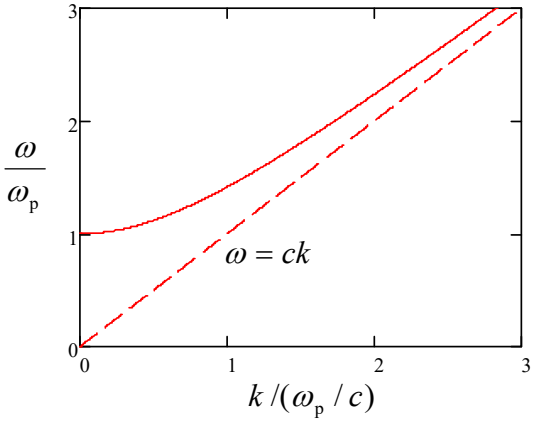


Fig. 7.6. The plasma dispersion law (solid line) in comparison with the linear dispersion in the free space (dashed line).

At $\omega \to \omega_p$ the wave number k tends to zero. Beyond that point (i.e. at $\omega < \omega_p$), we still can use Eq. (38), but it is instrumental to rewrite it in the mathematically equivalent form

[17] Indeed, let us consider such a small shift Δx, perpendicular to the plane surface of a broad, plane slab filled with plasma. The uncompensated ion charges, with equal and opposite surface densities $\sigma = \pm en\Delta x$, that appear at the slab surfaces, create inside it, according to Eq. (2.3), a uniform electric field with $E_x = en\Delta x/\varepsilon_0$. This field exerts the force $-eE_x = -(ne^2/\varepsilon_0)\Delta x \equiv -\kappa_{ef}\Delta x$ on each electron, pulling it back to its equilibrium position.

$$k(\omega) = \frac{i}{c}\left(\omega_{\rm p}^2 - \omega^2\right)^{1/2} = \frac{i}{\delta}, \qquad \text{where } \delta \equiv \frac{c}{\left(\omega_{\rm p}^2 - \omega^2\right)^{1/2}}. \tag{7.39}$$

At $\omega < \omega_{\rm p}$, the so-defined parameter δ is real, and Eq. (29) shows that the electromagnetic field exponentially decreases with distance:

$$f = \mathrm{Re}\, f_\omega e^{i(kz-\omega t)} \equiv \exp\left\{-\frac{z}{\delta}\right\} \mathrm{Re}\, f_\omega e^{-i\omega t}. \tag{7.40}$$

Does this mean that the wave is being absorbed in the plasma? Answering this question is a good pretext to calculate the time average of the Poynting vector $\mathbf{S} = \mathbf{E}\times\mathbf{H}$ of a monochromatic electromagnetic wave in an *arbitrary* dispersive (but still linear and isotropic) medium. First, let us spell out the real fields' time dependences:

$$E(t) = \mathrm{Re}\left[E_\omega e^{-i\omega t}\right] \equiv \frac{1}{2}\left[E_\omega e^{-i\omega t} + \text{c.c.}\right], \qquad H(t) = \mathrm{Re}\left[H_\omega e^{-i\omega t}\right] \equiv \frac{1}{2}\left[\frac{E_\omega}{Z(\omega)} e^{-i\omega t} + \text{c.c.}\right]. \tag{7.41}$$

Now, a straightforward calculation yields[18]

$$\overline{S} = \overline{E(t)H(t)} = \frac{E_\omega E_\omega^*}{4}\left[\frac{1}{Z(\omega)} + \frac{1}{Z^*(\omega)}\right] \equiv \frac{E_\omega E_\omega^*}{2}\mathrm{Re}\frac{1}{Z(\omega)} \equiv \frac{\left|E_\omega\right|^2}{2}\mathrm{Re}\left[\frac{\varepsilon(\omega)}{\mu(\omega)}\right]^{1/2}. \tag{7.42}$$

Let us apply this important general formula to our simple model of plasma at $\omega < \omega_{\rm p}$. In this case, the magnetic permeability equals μ_0, i.e. $\mu(\omega) = \mu_0$ is positive and real, while $\varepsilon(\omega)$ is real and negative, so $1/Z(\omega) = [\varepsilon(\omega)/\mu(\omega)]^{1/2}$ is purely imaginary, and the average Poynting vector (42) vanishes. This means that the energy, on average, does not flow along the z-axis. So, the waves with $\omega < \omega_{\rm p}$ are not absorbed in plasma. (Indeed, the Lorentz model with $\delta_j = 0$ does not describe any energy dissipation mechanism.) Instead, as we will see in the next section, the waves are rather *reflected* from the plasma's boundary, more exactly from its surface layer of a thickness $\sim\delta$.

Note also that in the limit $\omega << \omega_{\rm p}$, Eq. (39) yields

$$\delta \to \frac{c}{\omega_{\rm p}} = \left(\frac{c^2\varepsilon_0 m_{\rm e}}{ne^2}\right)^{1/2} = \left(\frac{m_{\rm e}}{\mu_0 n e^2}\right)^{1/2}. \tag{7.43}$$

But this is just a particular case (for $q = e$, $m = m_{\rm e}$, and $\mu = \mu_0$) of Eq. (6.44) that was derived in Sec. 6.4 for the depth of the magnetic field's penetration into a lossless (collision-free) conductor in the quasistatic approximation. This fact shows again that, as was already discussed in Sec. 6.7, this

[18] For an arbitrary plane wave, the total average power flow may be calculated as an integral of Eq. (42) over all frequencies. By the way, combining this integral and the Poynting theorem (6.111), is it straightforward to prove the following interesting expression for the average electromagnetic energy density of a narrow ($\Delta\omega << \omega$) wave packet propagating in an arbitrary dispersive (but linear and isotropic) medium:

$$\overline{u} = \frac{1}{2}\int_{\text{packet}} \left\{\frac{d\left[\omega\varepsilon'(\omega)\right]}{d\omega} E_\omega E_\omega^* + \frac{d\left[\omega\mu'(\omega)\right]}{d\omega} H_\omega H_\omega^*\right\} d\omega.$$

approximation (in which the displacement currents are neglected) gives an adequate description of the time-dependent phenomena at $\omega << \omega_p$, i.e. at $\delta << c/\omega = 1/k = \lambda/2\pi$.[19]

There are two most important examples of natural plasmas. For the Earth's ionosphere, i.e. the upper part of its atmosphere, which is almost completely ionized by the ultra-violet and X-ray components of the Sun's radiation, the maximum value of n, reached about 300 km over the Earth's surface, is between 10^{10} and 10^{12} m^{-3} (depending on the time of the day and the Sun's activity phase), so the maximum plasma frequency (37) is between 1 and 10 MHz. This is much higher than the particles' typical reciprocal collision time τ^{-1}, so Eq. (38) gives a good description of wave dispersion in this plasma. The effect of reflection of electromagnetic waves with $\omega < \omega_p$ from the ionosphere enables long-range (over-the-globe) radio communications and broadcasting at the so-called *short waves*, with cyclic frequencies of the order of 10 MHz:[20] they may propagate in the flat channel formed by the Earth's surface and the ionosphere, being reflected repeatedly by these parallel "walls". Unfortunately, due to the random variations of the Sun's activity, and hence of ω_p, this natural radio communication channel is not too reliable, and in our age of transworld optical-fiber cables (see Sec. 7 below), its practical importance has diminished.

Another important example of plasmas is free electrons in metals and other conductors. For a typical metal, n is of the order of 10^{23} cm^{-3} $\equiv 10^{29}$ m^{-3}, so Eq. (37) yields $\omega_p \sim 10^{16}$ s^{-1}. This value of ω_p is somewhat higher than the mid-optical frequencies ($\omega \sim 3\times10^{15}$ s^{-1}), explaining why planar, clean metallic surfaces, such as the aluminum and silver films used in mirrors, are so shiny: at these frequencies, their complex permittivity $\varepsilon(\omega)$ is almost exactly real and negative, leading to light reflection, with very little absorption.

The simple model (36), which neglects electron scattering, becomes inadequate at lower frequencies, $\omega\tau \sim 1$. A good phenomenological way of extending the model to the account of scattering is to take, in Eq. (33), the lowest frequency ω_j equal zero (to describe the free electrons), while keeping the damping coefficient δ_0 of this mode larger than zero, to account for the energy dissipation due to their scattering. Then Eq. (33) is reduced to

$$\varepsilon_{ef}(\omega) = \varepsilon_{opt}(\omega) + \frac{n_0 q^2}{m}\frac{1}{-\omega^2 - 2i\omega\delta_0} \equiv \varepsilon_{opt}(\omega) + i\frac{n_0 q^2}{2\delta_0 m\omega}\frac{1}{1 - i\omega/2\delta_0}, \tag{7.44}$$

where the response $\varepsilon_{opt}(\omega)$ at high (in practice, optical) frequencies is still given by Eq. (33), but now with $j > 0$. The result (44) allows for a simple interpretation. To show that, let us incorporate into our calculations the Ohmic conduction of the medium, generalizing Eq. (4.7) as $\mathbf{j}_\omega = \sigma(\omega)\mathbf{E}_\omega$ to account for the possible frequency dependence of the Ohmic conductivity. Plugging this relation into the Fourier image of the relevant macroscopic Maxwell equation, $\nabla\times\mathbf{H}_\omega = \mathbf{j}_\omega - i\omega\mathbf{D}_\omega \equiv \mathbf{j}_\omega - i\omega\varepsilon(\omega)\mathbf{E}_\omega$, we get

$$\nabla\times\mathbf{H}_\omega = \left[\sigma(\omega) - i\omega\varepsilon(\omega)\right]\mathbf{E}_\omega. \tag{7.45}$$

[19] One more convenience of the simple model of a collision-free plasma, which has led us to Eq. (36), is that it may be readily generalized to the case of an additional strong dc magnetic field $\mathbf{B}_0$ (much higher than that of the wave) applied in the direction $\mathbf{n}$ of wave propagation. It is straightforward (and hence left for the reader) to show that such plasma exhibits the *Faraday effect* of the polarization plane's rotation, and hence gives an example of an anisotropic media that violates the Lorentz reciprocity relation (6.121).

[20] These frequencies are an order of magnitude lower than those used for TV and FM-radio broadcasting.

This relation shows that for a monochromatic wave, the addition of the Ohmic current density $\mathbf{j}_\omega$ to the displacement current density is equivalent to the addition of $\sigma(\omega)$ to $-i\omega\varepsilon(\omega)$, i.e. to the following change of the ac electric permittivity:[21]

$$\varepsilon(\omega) \to \varepsilon_{\text{ef}}(\omega) \equiv \varepsilon_{\text{opt}}(\omega) + i\frac{\sigma(\omega)}{\omega}. \tag{7.46}$$

Now the comparison of Eqs. (44) and (46) shows that they coincide if we take

$$\sigma(\omega) = \frac{n_0 q^2 \tau}{m_0}\frac{1}{1-i\omega\tau} \equiv \sigma(0)\frac{1}{1-i\omega\tau}, \tag{7.47}$$

Generalized Drude formula

where the dc conductivity $\sigma(0)$ is described by the Drude formula (4.13), and the phenomenologically introduced coefficient δ_0 is associated with $1/2\tau$. Eq. (47), which is frequently called the *generalized* (or "ac", or "rf") *Drude formula*,[22] gives a very reasonable (semi-quantitative) description of the ac conductivity of many metals almost up to optical frequencies.

Now returning to our discussion of the generalized Lorentz model (33), we see that the frequency dependences of the real (ε') and imaginary (ε'') parts of the complex permittivity it yields are not quite independent. For example, let us have one more look at the resonance peaks in Fig. 5. Each time the real part drops with frequency, $d\varepsilon'/d\omega < 0$, its imaginary part ε'' has a positive peak. Ralph Kronig (in 1926) and Hendrik ("Hans") Kramers (in 1927) independently showed that this is not an occasional coincidence pertinent only to this particular model. Moreover, the full knowledge of the function $\varepsilon'(\omega)$ enables the *calculation* of the function $\varepsilon''(\omega)$, and vice versa. The mathematical reason for this fact is that both these functions are always related to a single real function $G(\theta)$ – see Eqs. (27).

To derive these relations, let us consider Eq. (26b) on the complex frequency plane, $\omega \to \boldsymbol{\omega} \equiv \omega' + i\omega''$:

$$f(\boldsymbol{\omega}) \equiv \varepsilon(\boldsymbol{\omega}) - \varepsilon_0 = \int_0^\infty G(\theta) e^{i\boldsymbol{\omega}\theta} d\theta \equiv \int_0^\infty G(\theta) e^{i\omega'\theta} e^{-\omega''\theta} d\theta. \tag{7.48}$$

For all stable physical systems, $G(\theta)$ has to be finite for all important values of the real integration variable ($\theta > 0$), and tend to zero at $\theta \to 0$ and $\theta \to \infty$. (Indeed, according to Eq. (23), a non-zero $G(0)$ would mean an instantaneous response of the medium to the external force, while $G(\infty) \neq 0$ would mean that it has an infinitely long memory.) Because of that, and thanks to the factor $e^{-\omega''\theta}$, the expression under the integral in Eq. (48) tends to zero at $|\boldsymbol{\omega}| \to \infty$ in all upper half-plane ($\omega'' \geq 0$). As a result, we may claim that the complex function $f(\boldsymbol{\omega})$ given by this relation, is analytical in that half-plane. This fact allows us to apply to it the general *Cauchy integral* formula[23]

$$f(\boldsymbol{\omega}) = \frac{1}{2\pi i}\oint_C f(\boldsymbol{\Omega})\frac{d\boldsymbol{\Omega}}{\boldsymbol{\Omega} - \boldsymbol{\omega}}, \tag{7.49}$$

[21] Alternatively, according to Eq. (45), it is possible (and in the field of infrared spectroscopy, conventional) to attribute the ac response of a medium, at *all* frequencies, to its effective complex conductivity: $\sigma_{\text{ef}}(\omega) \equiv \sigma(\omega) - i\omega\varepsilon(\omega) \equiv -i\omega\varepsilon_{\text{ef}}(\omega)$.

[22] It may be also derived from the Boltzmann kinetic equation in the so-called relaxation-time approximation (RTA) – see, e.g., SM Sec. 6.2.

[23] See, e.g., MA Eq. (15.2).

where $\mathbf{\Omega} \equiv \Omega' + i\Omega''$ is also a complex variable. Let us take the integration contour C of the form shown in Fig. 7, with the radius R of the larger semicircle tending to infinity, and the radius r of the smaller semicircle (around the singular point $\mathbf{\Omega} = \omega$) tending to zero. Due to the exponential decay of $|f(\mathbf{\Omega})|$ at $|\mathbf{\Omega}| \to \infty$, the contribution to the right-hand side of Eq. (49) from the larger semicircle vanishes,[24] while the contribution from the small semicircle, where $\mathbf{\Omega} = \omega + r\exp\{i\varphi\}$, with $-\pi \le \varphi \le 0$, is

$$\lim_{r\to 0} \frac{1}{2\pi i} \int_{\mathbf{\Omega}=\omega+r\exp\{i\varphi\}} f(\mathbf{\Omega}) \frac{d\mathbf{\Omega}}{\mathbf{\Omega}-\omega} = \frac{f(\omega)}{2\pi i} \int_{-\pi}^{0} \frac{ir\exp\{i\varphi\}d\varphi}{r\exp\{i\varphi\}} \equiv \frac{f(\omega)}{2\pi} \int_{-\pi}^{0} d\varphi = \frac{1}{2} f(\omega). \quad (7.50)$$

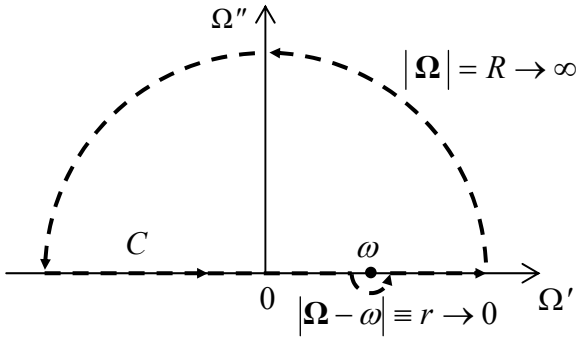


Fig. 7.7. Deriving the Kramers-Kronig dispersion relations.

As a result, for our contour C, Eq. (49) yields

$$f(\omega) = \lim_{r\to 0} \frac{1}{2\pi i} \left(\int_{-\infty}^{\omega-r} + \int_{\omega+r}^{+\infty} \right) f(\Omega) \frac{d\Omega}{\Omega-\omega} + \frac{1}{2} f(\omega)., \quad (7.51)$$

where $\Omega \equiv \Omega'$ on the real axis (where $\Omega'' = 0$). Such an integral, excluding a symmetric infinitesimal vicinity of a pole singularity, is called the *principal value* of the (formally, diverging) integral from $-\infty$ to $+\infty$, and is denoted by the letter P before it.[25] Using this notation, subtracting $f(\omega)/2$ from both parts of Eq. (51), and multiplying them by 2, we get

$$f(\omega) = \frac{1}{\pi i} \mathrm{P} \int_{-\infty}^{+\infty} f(\Omega) \frac{d\Omega}{\Omega-\omega}. \quad (7.52)$$

Now plugging into this complex equality the polarization-related difference $f(\omega) \equiv \varepsilon(\omega) - \varepsilon_0$ in the form $[\varepsilon'(\omega) - \varepsilon_0] + i[\varepsilon''(\omega)]$, and requiring both real and imaginary components of the two sides of Eq. (52) to be equal separately, we get the famous *Kramers-Kronig dispersion relations*

Kramers-Kronig dispersion relations

$$\varepsilon'(\omega) = \varepsilon_0 + \frac{1}{\pi} \mathrm{P} \int_{-\infty}^{+\infty} \varepsilon''(\Omega) \frac{d\Omega}{\Omega-\omega}, \qquad \varepsilon''(\omega) = -\frac{1}{\pi} \mathrm{P} \int_{-\infty}^{+\infty} [\varepsilon'(\Omega) - \varepsilon_0] \frac{d\Omega}{\Omega-\omega}. \quad (7.53)$$

We may use the already mentioned fact that $\varepsilon'(\omega)$ is always an even function, while $\varepsilon''(\omega)$ an odd function of frequency, to rewrite these relations in the following equivalent form,

[24] Strictly speaking, this also requires $|f(\Omega)|$ to decrease faster than Ω^{-1} at the real axis (at $\Omega'' = 0$), but due to the inertia of charged particles, this requirement is fulfilled for all realistic models of dispersion – see, e.g., Eq. (36).

[25] I am typesetting this symbol in a Roman (upright) font, to avoid any possibility of confusion with the medium's polarization.

$$\varepsilon'(\omega) = \varepsilon_0 + \frac{2}{\pi} \mathrm{P} \int_0^{+\infty} \varepsilon''(\Omega) \frac{\Omega d\Omega}{\Omega^2 - \omega^2}, \qquad \varepsilon''(\omega) = -\frac{2\omega}{\pi} \mathrm{P} \int_0^{+\infty} [\varepsilon'(\Omega) - \varepsilon_0] \frac{d\Omega}{\Omega^2 - \omega^2}, \tag{7.54}$$

which is more convenient for most applications, because it involves only physical (positive) frequencies.

Though the Kramers-Kronig relations are "global" in frequency, in certain cases they allow an approximate calculation of dispersion from experimental data for absorption, collected even within a limited ("local") frequency range. Most importantly, if a medium has a sharp absorption peak at some frequency ω_j, we may describe it as

$$\varepsilon''(\omega) \approx c\delta(\omega - \omega_j) + \text{a more smooth function of } \omega, \tag{7.55}$$

and the first of Eqs. (54) immediately gives

$$\varepsilon'(\omega) \approx \varepsilon_0 + \frac{2c}{\pi} \frac{\omega_j}{\omega_j^2 - \omega^2} + \text{another smooth function of } \omega, \tag{7.56}$$

Dispersion near an absorption line

thus predicting the anomalous dispersion near such a point. This calculation shows that such behavior observed in the Lorentz oscillator model (see Fig. 5) is by no means occasional or model-specific.

Let me emphasize again that the Kramers-Kronig relations (53)-(54) are much more general than the Lorentz model (33), and require only a causal linear relation (21) between the polarization $P(t)$ with the electric field $E(t')$.[26] Hence, these relations are also valid for the complex functions relating Fourier images of any cause/effect-related pair of variables. In particular, at a measurement of *any* linear response $r(t)$ of *any* experimental sample to *any* external field $f(t')$, whatever the nature of this response and physics behind it, we may be confident that there is a causal relationship between the variables r and f, so the corresponding complex function $\chi(\omega) \equiv r_\omega/f_\omega$ does obey the Kramers-Kronig relations. However, it is still important to remember that a linear relationship between the Fourier amplitudes of two variables does *not* necessarily imply a causal relationship between them.[27]

7.3. Reflection

The most important new effect arising in nonuniform media is wave *reflection*. Let us start its discussion from the simplest case of a plane electromagnetic wave that is normally incident on a sharp interface between two uniform, linear, isotropic media.

Moreover, let us first consider an even simpler sub-case when one of the two media (say, that located at $z > 0$, see Fig. 8) cannot sustain any electric field at all – as implied, in particular, by the macroscopic model of a perfect conductor – see Eq. (2.1):

[26] Actually, in mathematics, the relations even somewhat more general than Eqs. (53) and valid for an arbitrary analytic function of complex argument (the *Sokhotski-Plemelj theorem*) have been known at least from 1868.

[27] For example, the function $\varphi(\omega) \equiv E_\omega / P_\omega$ in the Lorentz oscillator model, does not obey the Kramers-Kronig relations. This is evident not only physically, from the fact that $E(t)$ is *not* a causal function of $P(t)$, but even mathematically. Indeed, Green's function describing a causal relationship has to tend to zero at small time delays $\theta \equiv t - t'$, so its Fourier image has to tend to zero at $\omega \to \pm\infty$. This is certainly true for the function $f(\omega)$ given by Eq. (32), but not for the reciprocal function $\varphi(\omega) \equiv 1/f(\omega) \propto (\omega^2 - \omega_0^2) - 2i\delta\omega$, which diverges at large frequencies.

$$E\big|_{z\geq 0} = 0. \tag{7.57}$$

This condition is evidently incompatible with the single traveling wave (5). However, this solution may be readily generalized using the fact that the dispersion-free 1D wave equation,

$$\left(\frac{\partial^2}{\partial z^2} - \frac{1}{v^2}\frac{\partial^2}{\partial t^2}\right)E = 0, \tag{7.58}$$

supports waves propagating, with the same speed, in any of two opposite directions. As a result, the following linear superposition of two such waves,

$$E\big|_{z\leq 0} = f(z - vt) - f(-z - vt), \tag{7.59}$$

satisfies both the equation and the boundary condition (57), for an arbitrary function f. The second term on the right-hand side of Eq. (59) may be interpreted as a result of *total reflection* of the incident wave (described by its first term) – in this particular case, with the change of the electric field's sign. This means, in particular, that within the macroscopic model, a conductor acts as a perfect mirror. By the way, since the vector **n** of the reflected wave is opposite to that of the incident one (see the arrows in Fig. 8), Eq. (6) shows that the magnetic field of the wave does *not* change its sign at the reflection:

$$H\big|_{z\leq 0} = \frac{1}{Z}\left[f(z - vt) + f(-z - vt)\right]. \tag{7.60}$$

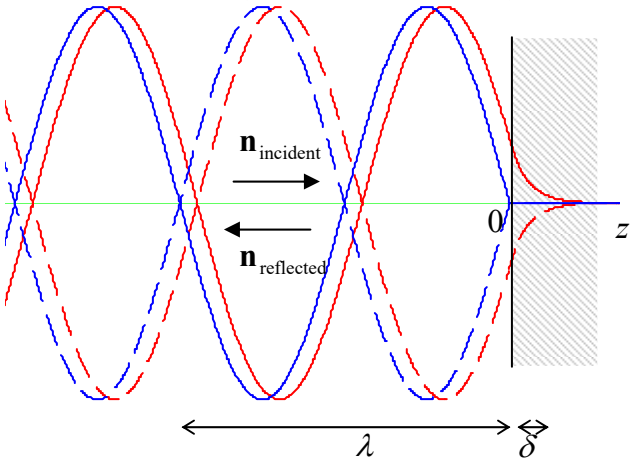


Fig. 7.8. A snapshot of the electric field at the reflection of a sinusoidal wave from a perfect conductor: a realistic pattern (red lines) and its macroscopic, ideal-mirror approximation (blue lines). Dashed lines show the snapshots after a half-period time delay ($\omega\Delta t = \pi$).

The blue lines in Fig. 8 show the resulting pattern (59) for the simplest, monochromatic wave:

Wave's total reflection

$$E\big|_{z\leq 0} = \mathrm{Re}\left[E_\omega e^{i(kz-\omega t)} - E_\omega e^{i(-kz-\omega t)}\right]. \tag{7.61a}$$

Depending on convenience in a particular context, this pattern may be legitimately represented and interpreted either as the linear superposition (61a) of two *traveling* waves or as a single *standing wave*:

$$E\big|_{z\leq 0} = -2\,\mathrm{Im}\left(E_\omega e^{-i\omega t}\right)\sin kz \equiv 2\,\mathrm{Re}\left(iE_\omega e^{-i\omega t}\right)\sin kz \equiv 2\,\mathrm{Re}\left[E_\omega e^{-i\left(\omega t - \pi/2\right)}\right]\sin kz, \tag{7.61b}$$

in which the electric and magnetic field oscillate with the phase shifts by $\pi/2$ both in time and space:

$$H\big|_{z\le 0} = \mathrm{Re}\left[\frac{E_\omega}{Z}e^{i(kz-\omega t)} + \frac{E_\omega}{Z}e^{i(-kz-\omega t)}\right] \equiv 2\,\mathrm{Re}\left(\frac{E_\omega}{Z}e^{-i\omega t}\right)\cos kz\,. \tag{7.62}$$

As a result of this shift, the time average of the Poynting vector's magnitude,

$$S(z,t) = EH = \frac{1}{Z}\mathrm{Re}\left[E_\omega^2 e^{-2i\omega t}\right]\sin 2kz\,, \tag{7.63}$$

equals zero, showing that at the total reflection, there is no *average* power flow. (This is natural because the perfect mirror can neither transmit the wave nor absorb it.) However, Eq. (63) shows that the standing wave features local oscillations of energy, transferring it periodically between the concentrations of the electric and magnetic fields, separated by the distance $\Delta z = \pi/2k = \lambda/4$.

In the case of the sinusoidal waves, the reflection effects may be readily explored even for the more general case of dispersive and/or lossy (but still linear) media in which $\varepsilon(\omega)$ and $\mu(\omega)$, and hence the wave vector $k(\omega)$ and the wave impedance $Z(\omega)$, defined by Eqs. (28), are certain complex functions of frequency. The "only" new factor we have to account for is that in this case, the reflection may not be total, so inside the second medium we have to use the traveling-wave solution as well. This factor may be taken care of by looking for the solution to our boundary problem in the form

$$E\big|_{z\le 0} = \mathrm{Re}\left[E_\omega\left(e^{ik_- z} + R\,e^{-ik_- z}\right)e^{-i\omega t}\right], \qquad E\big|_{z\ge 0} = \mathrm{Re}\left[E_\omega T e^{ik_+ z}e^{-i\omega t}\right], \tag{7.64}$$

Wave's partial reflection

and hence, according to Eq. (6),

$$H\big|_{z\le 0} = \mathrm{Re}\left[\frac{E_\omega}{Z_-(\omega)}\left(e^{ik_- z} - R\,e^{-ik_- z}\right)e^{-i\omega t}\right], \qquad H\big|_{z\ge 0} = \mathrm{Re}\left[\frac{E_\omega}{Z_+(\omega)}T e^{ik_+ z}e^{-i\omega t}\right]. \tag{7.65}$$

(The indices + and – correspond to the media located at $z > 0$ and $z < 0$, respectively.) Please note the following important features of Eqs. (64)-(65):

(i) They satisfy the Maxwell equations in both media. (Historically, the fact that at $z > 0$, these solutions do not include any components proportional to $\exp\{ik_- z\}$, looked surprising and was called the *wave extinction paradox*.)

(ii) Due to the problem's linearity, we could (and did :-) take the complex amplitudes of the reflected and transmitted wave proportional to that (E_ω) of the incident wave, while scaling them with dimensionless, generally complex coefficients R and T. As a comparison of Eqs. (64)-(65) with Eqs. (61)-(62) shows, the total reflection from an ideal mirror corresponds to $R = -1$ and $T = 0$.

(iii) Since in our current problem, the incident wave arrives from one side only (from $z = -\infty$), there is no need to include a term proportional to $\exp\{-ik_+ z\}$ into Eqs. (64)-(65) – even though this term is also a legitimate solution of our wave equation. However, we would need to add such a term if the medium at $z > 0$ had been nonuniform (e.g., had at least one more interface or any other inhomogeneity), because the wave reflected from that additional inhomogeneity would be incident on our interface (located at $z = 0$) from the right.

(iv) Eqs. (64)-(65) may be used even for the description of the cases when waves cannot propagate to $z \ge 0$, for example, a conductor or a plasma with $\omega_p > \omega$. Indeed, the exponential drop of the field amplitude at $z > 0$ in such cases is automatically described by the imaginary part of the wave number k_+ – see Eq. (29).

In order to calculate the coefficients R and T, we need to use boundary conditions at $z = 0$. Since in our current case of the normal incidence, the reflection does not change the transverse character of the partial waves, both vectors **E** and **H** remain tangential to the interface plane (in our notation, $z = 0$). Reviewing the arguments that have led us, in statics, to the boundary conditions (3.37) and (5.117) for these components, we see that they remain valid for the time-dependent situation as well,[28] so for our current case of normal incidence, we may write:

$$E\big|_{z=-0} = E\big|_{z=+0}, \qquad H\big|_{z=-0} = H\big|_{z=+0}. \tag{7.66}$$

Plugging Eqs. (64)-(65) into these conditions, we readily get two equations for the coefficients R and T:

$$1 + R = T, \qquad \frac{1}{Z_-}(1 - R) = \frac{1}{Z_+}T. \tag{7.67}$$

Solving this simple system of linear equations, we get[29]

Reflection and transmission: sharp interface

$$R = \frac{Z_+ - Z_-}{Z_+ + Z_-}, \qquad T = \frac{2Z_+}{Z_+ + Z_-}. \tag{7.68}$$

These formulas are very important, and much more general than one might think because they are applicable for virtually any 1D waves – electromagnetic or not, provided that the impedance Z is defined properly.[30] Since in the general case the wave impedances $Z_\pm$ defined by Eq. (28) with the corresponding indices, are complex functions of frequency, Eqs. (68) show that R and T may have imaginary parts as well. This fact has important consequences at $z < 0$, where the reflected wave, proportional to R, combines ("interferes") with the incident wave. Indeed, with $R = |R|e^{i\varphi}$ (where $\varphi \equiv \arg R$ is a real phase shift), the expression in the parentheses in the first of Eqs. (64) becomes

$$\begin{aligned} e^{ik_- z} + Re^{-ik_- z} &= \left(1 - |R| + |R|\right)e^{ik_- z} + |R|e^{i\varphi}e^{-ik_- z} \\ &\equiv \left(1 - |R|\right)e^{ik_- z} + 2|R|e^{i\varphi/2}\sin\left[k_-(z - \delta_-)\right], \qquad \text{where } \delta_- \equiv \frac{\varphi - \pi}{2k_-}. \end{aligned} \tag{7.69}$$

This means that the field may be represented as a sum of a traveling wave and a standing wave, with an amplitude proportional to $|R|$, shifted by the distance δ_- toward the interface, relative to the ideal-mirror pattern (61b) – see Fig. 8. This effect is frequently used for the experimental measurements of an unknown impedance Z_+ of some medium, provided than Z_- is known – most often, it is the free space, where $Z_- = Z_0$. For that, a small antenna (the *probe*), not disturbing the fields' distribution too much, is placed into the wave field, and the amplitude of the ac voltage induced in it by the wave is measured with a detector (e.g., a semiconductor diode with a nearly-quadratic *I-V* curve), as a function of z (Fig.

[28] For example, the first of Eqs. (66) may be obtained by integrating the full (time-dependent) Maxwell equation $\nabla\times\mathbf{E} + \partial\mathbf{B}/\partial t = 0$ over a narrow and long rectangular contour with dimensions l and d ($d << l$) stretched along the interface. At the application of the Stokes theorem to this integral, the first term gives $\Delta E_\tau l$, while the contribution of the second term is proportional to the product ld, so its contribution at $d/l \to 0$ is negligible. The proof of the second boundary condition is similar – as was already discussed in Sec. 6.2.

[29] Please note that only the media impedances (rather than their wave velocities) are important for reflection in this case! Unfortunately, this fact is not clearly emphasized in some textbooks that discuss only the case $\mu_\pm = \mu_0$, when $Z = (\mu_0/\varepsilon)^{1/2}$ and $v = 1/(\mu_0\varepsilon)^{1/2}$ are proportional to each other.

[30] See, e.g., the discussion of elastic waves of mechanical deformations in CM Secs. 6.3, 6.4, 7.7, and 7.8.

9). From the results of such a measurement, it is straightforward to find both $|R|$ and δ_-, and hence restore the complex R, and then use Eq. (68) to calculate both the modulus and the argument of Z_+. (Before computers became ubiquitous, a specially lined paper called the *Smith chart*, had been frequently used for performing this recalculation graphically; even nowadays, it is still used for result presentation.)

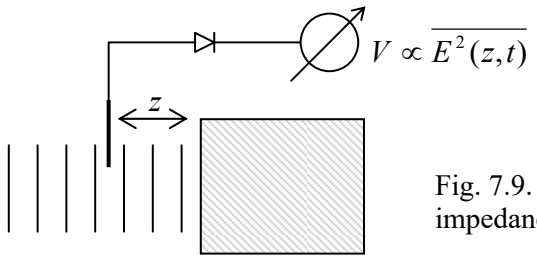


Fig. 7.9. Measurement of the complex impedance of a medium (schematically).

Now let us discuss what these results give for waves incident from the free space ($Z_-(\omega) = Z_0 =$ const, $k_- = k_0 = \omega/c$) onto the surfaces of two particular important media.

(i) For a collision-free plasma (with negligible magnetization) we may use Eq. (36) with $\mu(\omega) = \mu_0$, to represent the impedance (28) in either of two equivalent forms:

$$Z_+ = Z_0 \frac{\omega}{\left(\omega^2 - \omega_p^2\right)^{1/2}} \equiv -iZ_0 \frac{\omega}{\left(\omega_p^2 - \omega^2\right)^{1/2}}. \tag{7.70}$$

The first of these forms is more convenient in the case $\omega > \omega_p$, when the wave vector k_+ and the wave impedance Z_+ of the plasma are real, so a part of the incident wave does propagate into it. Plugging this expression into the latter of Eqs. (68), we see that T is real as well:

$$T = \frac{2\omega}{\omega + \left(\omega^2 - \omega_p^2\right)^{1/2}}. \tag{7.71}$$

Note that according to this formula, and somewhat counter-intuitively, $T > 1$ for any frequency (above ω_p), inviting the question: how can the transmitted wave be more intensive than the incident one that has induced it? To answer this question, we need to compare the powers (rather than the electric field amplitudes) of these two waves, i.e. their average Poynting vectors (42):

$$\overline{S_{\text{incident}}} = \frac{|E_\omega|^2}{2Z_0}, \qquad \overline{S_+} = \frac{|TE_\omega|^2}{2Z_+} = \frac{|E_\omega|^2}{2Z_0} \frac{4\omega\left(\omega^2 - \omega_p^2\right)^{1/2}}{\left[\omega + \left(\omega^2 - \omega_p^2\right)^{1/2}\right]^2}. \tag{7.72}$$

The ratio of these two values[31] is always below 1 (and tends to zero at $\omega \to \omega_p$), so only a fraction of the incident wave power may be transmitted. Hence the result $T > 1$ may be interpreted as follows: an interface between two media may be an *impedance transformer*: it can never transmit more *power* than the incident wave provides, i.e. can only decrease the product $S = EH$, but since the ratio $Z = E/H$ changes at the interface, the amplitude of *one of the fields* may increase at the transmission.

Now let us proceed to case $\omega < \omega_p$ when the waves cannot propagate in the plasma. In this case, the second of the expressions (70) is more convenient, because it immediately shows that Z_+ is purely

[31] This ratio is sometimes also called the "wave transmission coefficient", but to avoid its confusion with the T defined by Eq. (64), it is better to call it the *power transmission coefficient*.

imaginary, while $Z_- = Z_0$ is purely real. This means that $(Z_+ - Z_-) = (Z_+ + Z_-)^*$, i.e. according to the first of Eqs. (68), $|R| = 1$, so the reflection is total, i.e. no incident power (on average) is transferred into the plasma – as was already discussed in Sec. 2. However, the complex R has a finite argument,

$$\varphi \equiv \arg R = 2\arg(Z_+ - Z_0) = -2\tan^{-1}\frac{\omega}{\left(\omega_p^2 - \omega^2\right)^{1/2}}, \tag{7.73}$$

and hence provides a finite spatial shift (69) of the standing wave toward the plasma surface:

$$\delta_- = \frac{\varphi - \pi}{2k_0} = \frac{c}{\omega}\tan^{-1}\frac{\omega}{\left(\omega_p^2 - \omega^2\right)^{1/2}}. \tag{7.74}$$

On the other hand, we already know from Eq. (40) that the solution at $z > 0$ is exponential, with the decay length δ described by Eq. (39). Calculating, from the coefficient T, the exact coefficient before this exponent, it is straightforward to verify that the electric and magnetic fields are indeed continuous at the interface, completing the pattern shown with red lines in Fig. 8. This wave penetration into a fully reflecting material may be experimentally observed, for example, by thinning its sample. Even without solving this problem exactly, it is evident that if the sample's thickness d becomes comparable to δ, a part of the exponential "tail" of the field reaches its second interface, and induces a propagating wave. This is a classical-electromagnetic analog of the quantum-mechanical tunneling through a potential barrier.[32]

Note that at low frequencies, both δ_- and δ tend to the same frequency-independent value,

$$\delta, \delta_- \to \frac{c}{\omega_p} = \left(\frac{c^2\varepsilon_0 m_e}{ne^2}\right)^{1/2} = \left(\frac{m_e}{\mu_0 ne^2}\right)^{1/2}, \qquad \text{at } \frac{\omega}{\omega_p} \to 0, \tag{7.75}$$

which is just the field penetration depth (6.44) calculated for a perfect conductor model (assuming $m = m_e$ and $\mu = \mu_0$) in the quasistatic limit. This is natural, because the condition $\omega << \omega_p$ may be recast as $\lambda_0 \equiv 2\pi c/\omega >> 2\pi c/\omega_p \equiv 2\pi\delta$, i.e. as the quasistatic approximation's validity condition.

(ii) Now let us consider electromagnetic wave's reflection from an Ohmic, non-magnetic conductor. In the simplest low-frequency limit, when $\omega\tau$ is much less than 1, the conductor may be described by a frequency-independent conductivity σ.[33] According to Eq. (46), in this case we can take

$$Z_+ = \left(\frac{\mu_0}{\varepsilon_{opt}(\omega) + i\sigma/\omega}\right)^{1/2}. \tag{7.76}$$

With this substitution, Eqs. (68) immediately give us all the results of interest. In particular, in the most important quasistatic limit when $\delta_s \equiv (2/\mu_0\sigma\omega)^{1/2} << \lambda_0 \equiv 2\pi c/\omega$, i.e. $\sigma/\omega >> \varepsilon_0 \sim \varepsilon_{opt}$, the conductor' impedance is low:

$$Z_+ \approx \left(\frac{\mu_0\omega}{i\sigma}\right)^{1/2} \equiv \pi\left(\frac{2}{i}\right)^{1/2}\frac{\delta_s}{\lambda_0}Z_0, \qquad \text{i.e. } \left|\frac{Z_+}{Z_0}\right| << 1. \tag{7.77}$$

[32] See, e.g., QM Sec. 2.3.

[33] In a typical metal, $\tau \sim 10^{-13}$s, so this approximation works well up to $\omega \sim 10^{13}$ s^{-1}, i.e. up to the far-infrared frequencies.

This impedance is complex, and hence some fraction $\mathscr{f}$ of the incident wave is absorbed by the conductor. The fraction may be found as the ratio of the dissipated power (either calculated, as was done above, from Eqs. (68), or just taken from Eq. (6.36), with the magnetic field amplitude $|H_\omega| = 2|E_\omega|/Z_0$ – see Eq. (62)) to the incident wave's power given by the first of Eqs. (72). The result,

$$\mathscr{f} = \frac{2\omega\delta_s}{c} \equiv 4\pi\frac{\delta_s}{\lambda_0} << 1. \tag{7.78}$$

is used for crude estimates of the energy dissipation in metallic-wall waveguides and resonators. It shows that to keep the energy losses low, the characteristic size of such systems (which gives a scale of the free-space wavelengths λ_0 at which they are used) should be much larger than δ_s. A more detailed theory of these structures, and of the energy loss in them, will be discussed later in this chapter.

7.4. Refraction

Now let us consider the effects arising at a plane interface between two uniform media when the wave's incidence angle θ (Fig. 10) is arbitrary rather than equal to zero as in our previous analysis, for the simplest case of fully transparent media, with real $\varepsilon_\pm(\omega)$ and $\mu_\pm(\omega)$. (For the sake of notation simplicity, in most formulas below, the argument of these functions will be dropped, i.e. just implied.)

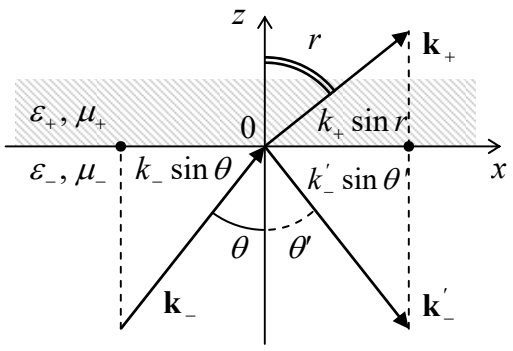


Fig. 7.10. Plane wave's reflection, transmission, and refraction at a plane interface. The plane of the drawing is selected to contain all three wave vectors: $\mathbf{k}_+$, $\mathbf{k}_-$, and $\mathbf{k}'_-$.

In contrast with the case of normal incidence, here the wave vectors $\mathbf{k}_-$, $\mathbf{k}'_-$, and $\mathbf{k}_+$ of the three components (incident, reflected, and transmitted) waves may have different directions. (Such change of the transmitted wave's direction is called *refraction*.) Hence let us start our analysis by writing a general expression for a single plane, monochromatic wave for the case when its wave vector $\mathbf{k}$ has all three Cartesian components, rather than one. An evident generalization of Eq. (11) for this case is

$$f(\mathbf{r},t) = \mathrm{Re}\left[f_\omega e^{i\left(k_x x + k_y y + k_z z\right) - \omega t)}\right] \equiv \mathrm{Re}\left[f_\omega e^{i(\mathbf{k}\cdot\mathbf{r} - \omega t)}\right]. \tag{7.79}$$

This expression enables a ready analysis of "kinematic" relations, which are independent of the media impedances. Indeed, it is sufficient to notice that to satisfy *any* linear, homogeneous boundary conditions at the interface ($z = 0$), all partial plane waves must have the same temporal and spatial dependence on this plane. Hence if we select the x-z plane so that the vector $\mathbf{k}_-$ lies in it, then $(k_-)_y = 0$, and $\mathbf{k}_+$ and $\mathbf{k}'_-$ cannot have any y-component either, i.e. all three wave vectors lie in the same plane –

that is selected as the plane of the drawing in Fig. 10. Moreover, because of the same reason, their x-components should be equal:

$$k_- \sin\theta = k_-' \sin\theta' = k_+ \sin r\,. \tag{7.80}$$

From here we immediately get two well-known laws: of reflection

Reflection angle

$$\theta' = \theta\,, \tag{7.81}$$

and of refraction:[34]

Snell law

$$\frac{\sin r}{\sin\theta} = \frac{k_-}{k_+}\,. \tag{7.82}$$

In this form, the laws are valid for plane waves of any nature. In optics, the Snell law (82) is frequently represented in the form

$$\frac{\sin r}{\sin\theta} = \frac{n_-}{n_+}\,, \tag{7.83}$$

where $n_\pm$ is the *index of refraction* (also called the "refractive index") of the corresponding medium, defined as its wave number normalized to that of the free space (at the particular wave's frequency):

Index of refraction

$$n_\pm \equiv \frac{k_\pm}{k_0} \equiv \left(\frac{\varepsilon_\pm \mu_\pm}{\varepsilon_0 \mu_0}\right)^{1/2}\,. \tag{7.84}$$

Perhaps the most famous corollary of the Snell law is that if a wave propagates from a medium with a higher index of refraction to that with a lower one (i.e. if $n_- > n_+$ in Fig. 10), for example from water to air, there is always a certain *critical* value θ_c of the incidence angle,

Critical angle

$$\theta_c = \sin^{-1}\frac{n_+}{n_-} \equiv \sin^{-1}\left(\frac{\varepsilon_+ \mu_+}{\varepsilon_- \mu_-}\right)^{1/2}\,, \tag{7.85}$$

at which the refraction angle r (see Fig. 10 again) reaches $\pi/2$. At a larger θ, i.e. within the range $\theta_c < \theta < \pi/2$, the boundary conditions (80) cannot be satisfied by a refracted wave with a real wave vector, so the wave experiences the so-called *total internal reflection*. This effect is very important for practice because it means that dielectric surfaces may be used as optical mirrors, in particular in optical fibers – to be discussed in more detail in Sec. 7 below. This is very fortunate for telecommunication technology because light's reflection from metals is rather imperfect. Indeed, according to Eq. (78), in the optical range ($\lambda_0 \sim 0.5$ μm, i.e. $\omega \sim 10^{15}$ s^{-1}), even the best conductors (with $\sigma \sim 6\times10^8$ S/m and hence the normal skin depth $\delta_s \sim 1.5$ nm) provide power loss of at least a few percent at each reflection.

Note, however, that even within the range $\theta_c < \theta < \pi/2$, the field at $z > 0$ is not identically equal to zero: it penetrates into the lower-n media by a distance of the order of λ_0, exponentially decaying inside it, just as it does at the normal incidence – see Fig. 8. However, at $\theta \neq 0$ the penetrating field still propagates, with the wave number (80), along the interface. Such a field, exponentially dropping in one direction but still propagating as a wave in another direction, is commonly called the *evanescent wave*.

[34] The latter relation is traditionally called the *Snell law*, after a 17th-century astronomer Willebrord Snellius, but it has been traced all the way back to a circa 984 work by Abu Saad al-Ala ibn Sahl. (Claudius Ptolemy who performed pioneering experiments on light refraction in the 2nd century AD, was just one step from this result.)

One more remark: just as at the normal incidence, the field's penetration into another medium causes a phase shift of the reflected wave – see, e.g., Eq. (69) and its discussion. A new feature of this phase shift, arising at $\theta \neq 0$, is that it also has a component parallel to the interface – the so-called *Goos-Hänchen effect*. In geometric optics, this effect leads to an image shift (relative to its position in a perfect mirror) with components both normal and parallel to the interface.

Now let us carry out an analysis of "dynamic" relations that determine amplitudes of the refracted and reflected waves. For this, we need to write explicitly the boundary conditions at the interface (i.e. the plane $z = 0$). Since now the electric and/or magnetic fields may have components normal to the plane, in addition to the continuity of their tangential components, which were repeatedly discussed above,

$$E_{x,y}\big|_{z=-0} = E_{x,y}\big|_{z=+0}, \qquad H_{x,y}\big|_{z=-0} = H_{x,y}\big|_{z=+0}, \tag{7.86}$$

we also need relations for the normal components. As it follows from the homogeneous macroscopic Maxwell equations (6.99b), they are also the same as in statics, i.e. $D_n = \text{const}$, and $B_n = \text{const}$, for our coordinate choice (Fig. 10) giving

$$\varepsilon_- E_z\big|_{z=-0} = \varepsilon_+ E_z\big|_{z=+0}, \qquad \mu_- H_z\big|_{z=-0} = \mu_+ H_z\big|_{z=+0}. \tag{7.87}$$

The expressions of these components via the amplitudes E_ω, RE_ω, and TE_ω of the incident, reflected, and transmitted waves depend on the incident wave's polarization. For example, for a linearly-polarized wave with the electric field vector *normal* to the plane of incidence, i.e. *parallel* to the interface plane, the reflected and refracted waves are similarly polarized – see Fig. 11a.

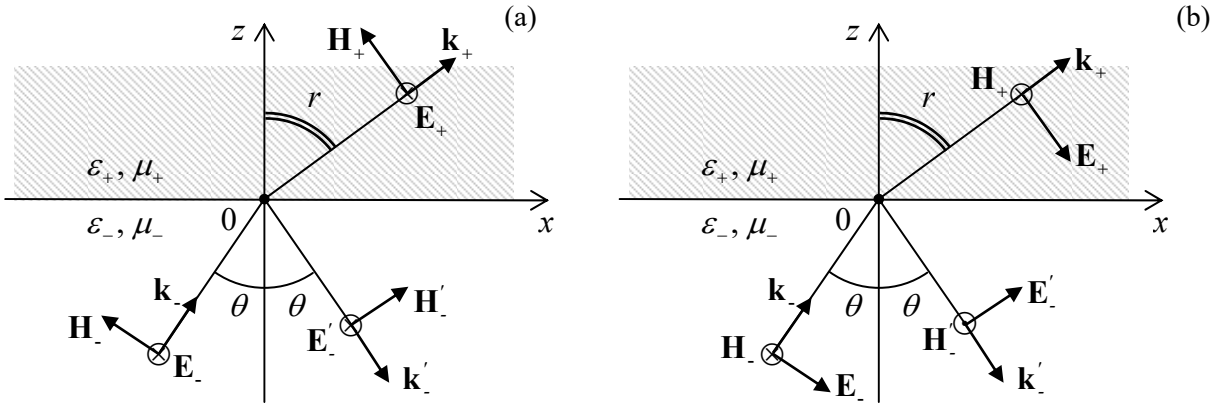


Fig. 7.11. Reflection and refraction at two different linear polarizations of the incident wave.

As a result, all E_z are equal to zero (so the first of Eqs. (87) is inconsequential), while the tangential components of the electric field are equal to their full amplitudes, just as at the normal incidence, so we still can use Eqs. (64) expressing these components via the coefficients R and T. However, at $\theta \neq 0$ the magnetic fields have not only tangential components

$$H_x\big|_{z=-0} = \mathrm{Re}\left[\frac{E_\omega}{Z_-}(1-R)\cos\theta\, e^{-i\omega t}\right], \qquad H_x\big|_{z=+0} = \mathrm{Re}\left[\frac{E_\omega}{Z_+}T\cos r\; e^{-i\omega t}\right], \tag{7.88}$$

but also normal components (see Fig. 11a):

$$H_z\big|_{z=-0} = \mathrm{Re}\left[\frac{E_\omega}{Z_-}(1+R)\sin\theta\ e^{-i\omega t}\right], \qquad H_z\big|_{z=+0} = \mathrm{Re}\left[\frac{E_\omega}{Z_+}T\sin r\ e^{-i\omega t}\right]. \tag{7.89}$$

Plugging these expressions into the boundary conditions expressed by Eqs. (86) (in this case, for the *y*-components only) and the second of Eqs. (87), we get *three* equations for *two* unknown coefficients *R* and *T*. However, two of these equations duplicate each other because of the Snell law, and we get just two independent equations,

$$1+R=T, \qquad \frac{1}{Z_-}(1-R)\cos\theta = \frac{1}{Z_+}T\cos r, \tag{7.90}$$

which are a very natural generalization of Eqs. (67), with the replacements $Z_- \to Z_-\cos r$, $Z_+ \to Z_+\cos\theta$. As a result, we can immediately use Eq. (68) to write the solution of the system (90):[35]

$$R = \frac{Z_+\cos\theta - Z_-\cos r}{Z_+\cos\theta + Z_-\cos r}, \qquad T = \frac{2Z_+\cos\theta}{Z_+\cos\theta + Z_-\cos r}. \tag{7.91a}$$

If we want to express these coefficients via the angle of incidence alone, we should use the Snell law (82) to eliminate the angle *r*, getting frequently quoted bulkier expressions:

$$R = \frac{Z_+\cos\theta - Z_-\left[1-(k_-/k_+)^2\sin^2\theta\right]^{1/2}}{Z_+\cos\theta + Z_-\left[1-(k_-/k_+)^2\sin^2\theta\right]^{1/2}}, \qquad T = \frac{2Z_+\cos\theta}{Z_+\cos\theta + Z_-\left[1-(k_-/k_+)^2\sin^2\theta\right]^{1/2}}. \tag{7.91b}$$

However, conceptually it is preferable to use the kinematic relation (82) and the dynamic relations (91a) separately, because Eq. (91b) obscures the very important physical fact that the ratio of $k_\pm$, i.e. of the wave velocities of the two media, is only involved in the Snell law, while Eqs. (91b) explicitly include only the wave impedances – just as in the case of normal incidence.

In the opposite case of the linear polarization of the electric field within the plane of incidence (Fig. 11b), it is the magnetic field that does not have a normal component, so it is now the second of Eqs. (87) that does not participate in the solution. However, now the electric fields in the two media have not only tangential components,

$$E_x\big|_{z=-0} = \mathrm{Re}\left[E_\omega(1+R)\cos\theta\ e^{-i\omega t}\right], \qquad E_x\big|_{z=+0} = \mathrm{Re}\left[E_\omega T\cos r\ e^{-i\omega t}\right], \tag{7.92}$$

but also normal components (Fig. 11b):

$$E_z\big|_{z=-0} = E_\omega(-1+R)\sin\theta, \qquad E_z\big|_{z=+0} = -E_\omega T\sin r. \tag{7.93}$$

As a result, instead of Eqs. (90), the reflection and transmission coefficients are related as

$$(1+R)\cos\theta = T\cos r, \qquad \frac{1}{Z_-}(1-R) = \frac{1}{Z_+}T. \tag{7.94}$$

Again, the solution of this system may be immediately written using the analogy with Eq. (67):

[35] Note that we may calculate the reflection and transmission coefficients *R'* and *T'* for the wave traveling in the opposite direction just by making the following parameter swaps: $Z_+ \leftrightarrow Z_-$ and $\theta \leftrightarrow r$, and that the resulting coefficients satisfy the following *Stokes relations*: $R' = -R$, and $R^2 + TT' = 1$, for any $Z_\pm$.

$$R = \frac{Z_+ \cos r - Z_- \cos\theta}{Z_+ \cos r + Z_- \cos\theta}, \qquad T = \frac{2Z_+ \cos\theta}{Z_+ \cos r + Z_- \cos\theta}, \tag{7.95a}$$

or, alternatively, using the Snell law, in a more bulky form:

$$R = \frac{Z_+\left[1-(k_-/k_+)^2\sin^2\theta\right]^{1/2} - Z_-\cos\theta}{Z_+\left[1-(k_-/k_+)^2\sin^2\theta\right]^{1/2} + Z_-\cos\theta}, \qquad T = \frac{2Z_+\cos\theta}{Z_+\left[1-(k_-/k_+)^2\sin^2\theta\right]^{1/2} + Z_-\cos\theta}. \tag{7.95b}$$

For the particular case $\mu_+ = \mu_- = \mu_0$, when $Z_+/Z_- = (\varepsilon_-/\varepsilon_+)^{1/2} = k_-/k_+ = n_-/n_+$ (which is approximately correct for traditional optical media), Eqs. (91b) and (95b) are called the *Fresnel formulas*.[36] Most textbooks are quick to point out that there is a major difference between them: while for the electric field polarization within the plane of incidence (Fig. 11b), the reflected wave's amplitude (proportional to the coefficient R) turns to zero[37] at a special value of θ (called the *Brewster angle*):[38]

$$\theta_B = \tan^{-1}\frac{n_+}{n_-}, \tag{7.96}$$

while there is no such angle in the opposite case (shown in Fig. 11a). However, note that this statement, as well as Eq. (96), is true only for the case $\mu_+ = \mu_-$. In the general case of different ε and μ, Eqs. (91) and (95) show that the reflected wave vanishes at $\theta = \theta_B$ with

$$\tan^2\theta_B = \frac{\varepsilon_-\mu_+ - \varepsilon_+\mu_-}{\varepsilon_+\mu_+ - \varepsilon_-\mu_-} \times \begin{cases} (\mu_+/\mu_-), & \text{for } \mathbf{E} \perp \mathbf{n}_z \text{ (Fig.11a)}, \\ (-\varepsilon_+/\varepsilon_-), & \text{for } \mathbf{H} \perp \mathbf{n}_z \text{ (Fig.11b)}. \end{cases} \tag{7.97}$$

Brewster angle

Note the natural $\varepsilon \leftrightarrow \mu$ symmetry of these relations, resulting from the $\mathbf{E} \leftrightarrow \mathbf{H}$ symmetry for these two polarization cases (Fig. 11). These formulas also show that for any set of parameters of the two media (with $\varepsilon_\pm$, $\mu_\pm > 0$), $\tan^2\theta_B$ is positive (and hence a real Brewster angle θ_B exists) only for one of these two polarizations. In particular, if the interface is due to the change of μ alone (i.e. if $\varepsilon_+ = \varepsilon_-$), the first of Eqs. (97) is reduced to the simple form (96) again, while for the polarization shown in Fig. 11b, there is no Brewster angle, i.e. the reflected wave has a non-zero amplitude for any θ.

Such an account of both media parameters, ε and μ, on an equal footing is necessary to describe several interesting effects. The first of them is the so-called *negative refraction*.[39] As was shown in Sec.

[36] Named after Augustin-Jean Fresnel (1788-1827), one of the wave optics pioneers, who is credited, among many other contributions (see, in particular, discussions in Ch. 8), for the concept of light as a purely transverse wave.

[37] This effect is used in practice to obtain linearly polarized light, with the electric field vector perpendicular to the plane of incidence, from the natural light with its random polarization. An even more widespread application of the effect is a partial reduction of undesirable glare from wet pavement (for the water/air interface, $n_+/n_- \approx 1.33$, giving $\theta_B \approx 50°$) by covering glasses and car headlights with thin vertically-polarizing layers.

[38] A very simple interpretation of Eq. (96) is based on the fact that, together with the Snell law (82), it gives $r + \theta = \pi/2$. As a result, the vector $\mathbf{E}_+$ is parallel to the vector $\mathbf{k}'_-$, and hence the oscillating electric dipoles of the medium do not have the Cartesian component that could induce the transverse electric field $\mathbf{E}'_-$ of the potential reflected wave.

[39] Despite some important background theoretical work by A. Schuster (1904), L. Mandelstam (1945), D. Sivikhin (1957), and especially V. Veselago (1966-67), the negative refractivity effects became a subject of intensive scientific research and engineering development only in the 2000s.

2, in a medium with electric-field-driven resonances, the function $\varepsilon(\omega)$ may be almost real and negative, at least within limited frequency intervals – see, in particular, Eq. (34) and Fig. 5. As has already been discussed, if, at these frequencies, the function $\mu(\omega)$ is real and positive, then $k^2(\omega) = \omega^2\varepsilon(\omega)\mu(\omega) < 0$, and k may be represented as i/δ with a real δ, meaning the exponential field decay into the medium. However, let us consider the case when both $\varepsilon(\omega) < 0$ *and* $\mu(\omega) < 0$ at a certain frequency. (This is possible in a medium with both **E**-driven and **H**-driven resonances, at a proper choice of their resonance frequencies.) Since in this case $k^2(\omega) = \omega^2\varepsilon(\omega)\mu(\omega) > 0$, the wave vector is real, Eq. (79) describes a traveling wave, and one could think that there is nothing new in this case. Not so!

First of all, for a sinusoidal plane wave (79), the operator ∇ is equivalent to the multiplication by $i\mathbf{k}$. As the Maxwell equations (2a) show, this means that at a fixed direction of vectors **E** and **k**, the simultaneous reversal of signs of ε and μ means the reversal of the direction of the vector **H**. Namely, if both ε and μ are positive, these equations are satisfied with mutually orthogonal vectors {**E**, **H**, **k**} forming the usual, *right-hand* system (see Fig. 1 and Fig. 12a), the name stemming from the popular "right-hand rule" used to determine the vector product's direction. However, if both ε and μ are negative, the vectors form a *left-hand* system – see Fig. 12b. (Due to this fact, the media with $\varepsilon < 0$ and $\mu < 0$ are frequently called the *left-handed materials*, LHM for short.) According to the basic relation (6.114), which does not involve media parameters, this means that for a plane wave in a left-hand material, the Poynting vector $\mathbf{S} = \mathbf{E}\times\mathbf{H}$, i.e. the energy flow, is directed *opposite* to the wave vector **k**.

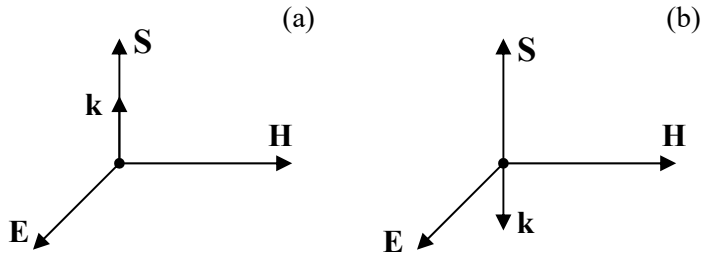


Fig. 7.12. Directions of the main vectors of a plane wave inside a medium with (a) positive and (b) negative values of ε and μ.

This fact may look strange but is in no contradiction with any fundamental principle. Let me remind you that, according to the definition of the vector **k**, its direction shows the direction of the *phase* velocity $v_{ph} = \omega/k$ of a sinusoidal (and hence infinitely long) wave, which cannot be used, for example, for signaling. Such signaling (by sending wave packets – see Fig. 13) is possible only with the *group* velocity $v_{gr} = d\omega/dk$. This velocity in left-hand materials is always directed (as in the right-hand materials) along the vector **S**, i.e. along the wave's energy flow.

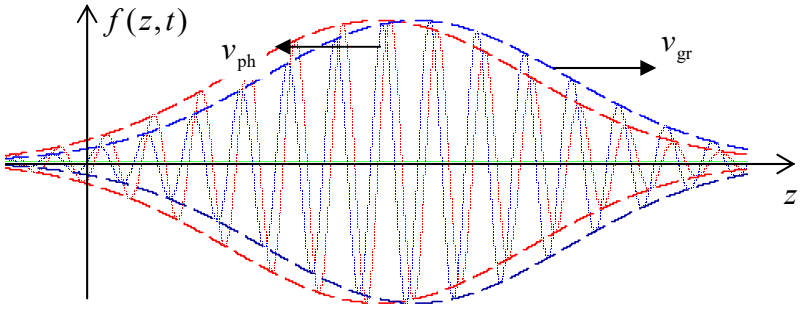


Fig. 7.13. An example of a wave packet moving along axis z with a negative phase velocity, but positive group velocity. Blue lines show a packet's snapshot a short time interval after the first snapshot (red lines).

Perhaps the most fascinating effect possible with left-hand materials is the wave refraction at their interfaces with the usual, right-handed materials – first predicted by V. Veselago in 1960. Consider the example shown in Fig. 14a. In the incident wave, arriving from a usual material, the directions of the vectors $\mathbf{k}_-$ and $\mathbf{S}_-$ coincide, and so they are in the reflected wave with vectors $\mathbf{k}'_-$ and $\mathbf{S}'_-$. This means that the electric and magnetic fields in the interface plane ($z = 0$) are, at our choice of the coordinate axes, proportional to $\exp\{ik_x x\}$, with a positive component $k_x = k_- \cos\theta$. To satisfy any linear boundary conditions, the refracted wave, propagating into the left-handed material, has to match that dependence, i.e. have a positive x-component of its wave vector $\mathbf{k}_+$. But in this medium, this vector has to be antiparallel to the vector $\mathbf{S}$, which in turn should be directed out of the interface, because it represents the power flow from the interface into the material's bulk. These conditions cannot be reconciled by the refracted wave propagating along the usual Snell-law direction (shown with the dashed line in Fig. 13a), but are all satisfied at refraction in the direction given by Snell's angle with the opposite sign. (Hence the term "negative refraction").[40]

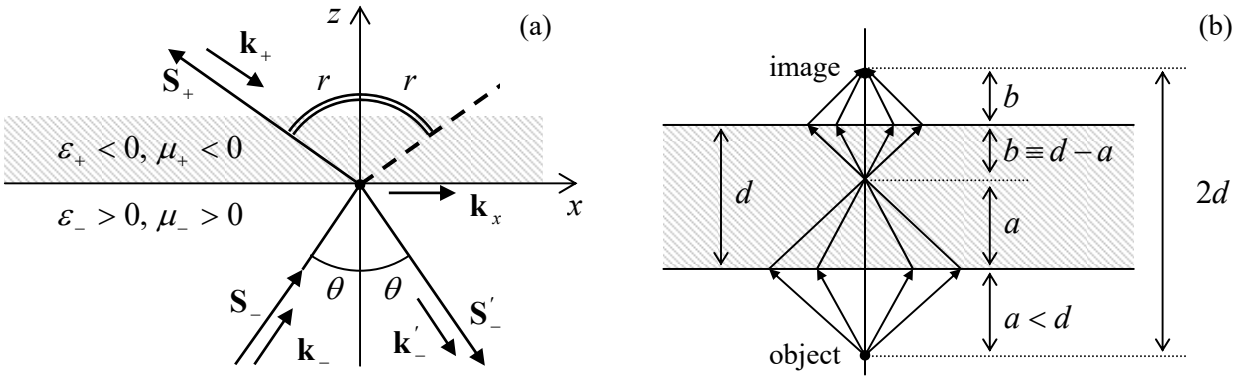


Fig. 7.14. Negative refraction: (a) waves at the interface between media with positive and negative values of $\varepsilon\mu$, and (b) the hypothetical *perfect lens*: a parallel plate made of a material with $\varepsilon = -\varepsilon_0$ and $\mu = -\mu_0$.

In order to understand how unusual the results of the negative refraction may be, let us consider a parallel slab of thickness d, made of a hypothetical left-handed material with exactly selected values $\varepsilon = -\varepsilon_0$, and $\mu = -\mu_0$ (see Fig. 14b). For such a material, placed in free space, the refraction angle $r = -\theta$, so the rays from a point source, located in free space at a distance $a < d$ from the slab's surface, propagate as shown on that panel, i.e. all meet again at the distance a beyond the surface, and then continue to propagate to the second surface of the slab. Repeating our discussion for this surface, we see that a point's image is also formed beyond the slab, at distance $2a + 2b = 2a + 2(d - a) = 2d$ from the object.

Superficially, this system looks like a usual lens, but the well-known lens formula, which relates a and b with the focal length f, is *not* satisfied. (In particular, a parallel beam is *not* focused into a point at any finite distance.) As an additional difference from the usual lens, the system shown in Fig. 14b *does not reflect* any part of the incident light. Indeed, it is straightforward to check that for all the above formulas for R and T to be valid, the sign of the wave impedance Z in left-handed materials has to be kept positive. Thus, for our particular choice of parameters ($\varepsilon = -\varepsilon_0$, $\mu = -\mu_0$), Eqs. (91a) and (95a) are

[40] In some publications inspired by this fact, the left-hand materials are prescribed a negative index of refraction n. However, this prescription should be treated with care. For example, it complies with the first form of Eq. (84), but not its second form, and the sign of n, in contrast to that of the wave vector $\mathbf{k}$, is a matter of convention.

valid with $Z_+ = Z_- = Z_0$ and $\cos r = \cos \theta = 1$, giving $R = 0$ for any linear polarization, and hence for any other wave polarization – circular, elliptic, natural, etc.

The perfect lens suggestion has triggered a wave of efforts to implement left-hand materials experimentally. (Attempts to find such materials in nature have failed so far.) Most progress in this direction has been achieved using the so-called *metamaterials*, which are essentially quasi-periodic arrays of specially designed electromagnetic resonators, ideally with high density $n >> \lambda^{-3}$. For example, Fig. 15 shows the metamaterial that was used for the first demonstration of negative refractivity in the microwave region – for ~10-GHz waves.[41] It combines straight strips of a metallic film, working as lumped resonators with a large electric dipole moment (and hence strongly coupled to the wave's electric field **E**), and several almost-closed film loops (so-called *split rings*), working as lumped resonators with large magnetic dipole moments, strongly coupled to the field **H**. The negative refractivity is achieved by designing the resonance frequencies close to each other. More recently, metamaterials with negative refractivity were demonstrated in the optical range as well,[42] although to the best of my knowledge, their relatively large absorption still prevents practical applications.

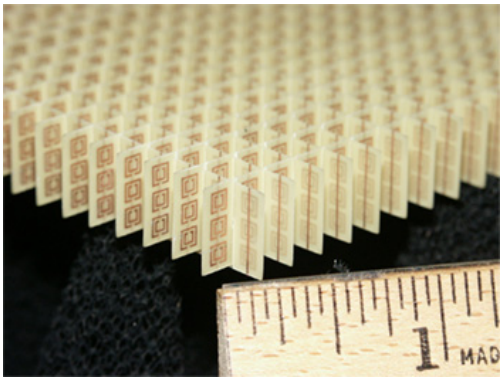


Fig. 7.15. An artificial left-hand material providing negative refraction at microwave frequencies ~10 GHz. The original by Jeffrey D. Wilson (in the public domain) is available at https://en.wikipedia.org/wiki/Metamaterial.

This progress has stimulated the development of other potential uses of metamaterials (not necessarily the left-handed ones), in particular, designs of nonuniform systems with handcrafted distributions $\varepsilon(\mathbf{r}, \omega)$ and $\mu(\mathbf{r}, \omega)$ that may provide electromagnetic wave propagation along the desired paths, e.g., around a certain region of space, making it virtually invisible for an external observer – so far, within very limited frequency ranges.[43]

As was mentioned in Sec. 5.5, another way to reach negative values of $\mu(\omega)$ is to place a ferromagnetic material into such an external dc magnetic field that the frequency ω_r of the ferromagnetic resonance is somewhat lower than ω. If thin layers of such a material (e.g., nickel) are interleaved with layers of a non-magnetic good conductor (such as copper), the average value of $\mu(\omega)$ of the resulting metamaterial may be positive but substantially below μ_0. According to Eq. (6.33), the skin-depth δ_s of such a material may be larger than that of the good conductor alone, enforcing a more uniform distribution of the ac current flowing along the layers, and hence making the energy losses lower than in the good conductor alone. This effect may be useful, in particular, for electronic circuit interconnects.[44]

[41] R. Shelby *et al.*, Science **292**, 77 (2001); J. Wilson and Z. Schwartz, *Appl. Phys. Lett.* **86**, 021113 (2005).
[42] See, e.g., J. Valentine *et al.*, *Nature* **455**, 376 (2008).
[43] For a review of such "invisibility cloaks", see, e.g., B. Wood, *Comptes Rendus Physique* **10**, 379 (2009).
[44] See, for example, N. Sato *et al.*, *J. Appl. Phys.* **111**, 07A501 (2012), and references therein.

7.5. Transmission lines: TEM waves

So far, we have analyzed plane electromagnetic waves, implying that their cross-section is infinite – evidently, an unrealistic assumption. The cross-section may be limited, still sustaining wave propagation, using *wave transmission lines*:[45] long, uniform structures made of either good conductors or dielectrics. Let us first discuss the first option, using the following simplifying assumptions:

(i) the structure is a cylinder (not necessarily with a round cross-section, see Fig. 16) filled with a usual (right-handed), uniform dielectric material with negligible energy losses ($\varepsilon'' = \mu'' = 0$), and

(ii) the wave attenuation due to the skin effect is also negligibly low. (As Eq. (78) indicates, for that the characteristic size a of the line's cross-section has to be much larger than the skin-depth δ_s of its wall material. The energy dissipation effects will be analyzed in Sec. 9 below.)

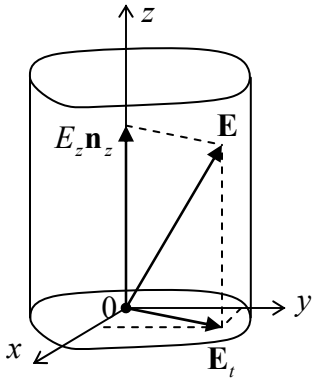


Fig. 7.16. Electric field's decomposition in a transmission line (in particular, a waveguide).

With such exclusion of energy losses, we may look for a particular solution of the macroscopic Maxwell equations in the form of a monochromatic wave traveling along the line:

$$\mathbf{E}(\mathbf{r},t) = \mathrm{Re}\left[\mathbf{E}_\omega(x,y)e^{i(k_z z-\omega t)}\right], \qquad \mathbf{H}(\mathbf{r},t) = \mathrm{Re}\left[\mathbf{H}_\omega(x,y)e^{i(k_z z-\omega t)}\right], \tag{7.98}$$

with real k_z, where the z-axis is directed along the transmission line – see Fig. 16. Note that this form allows a substantial coordinate dependence of the electric and magnetic field within the plane [x, y] of the transmission line's cross-section, as well as nonvanishing longitudinal components E_z and/or H_z of the fields, so the solution (98) is substantially more general than the plane waves discussed above. We will see in a minute that as a result, the parameter k_z may be very much different from its plane-wave value (13), $k \equiv \omega(\varepsilon\mu)^{1/2}$, in the same material, at the same frequency.

In order to describe these effects quantitatively, let us decompose the complex amplitudes of the wave's fields into their longitudinal and transverse components (Fig. 16):[46]

$$\mathbf{E}_\omega = E_z\mathbf{n}_z + \mathbf{E}_t, \quad \mathbf{H}_\omega = H_z\mathbf{n}_z + \mathbf{H}_t. \tag{7.99}$$

[45] Another popular term is the *waveguide*, but it is typically reserved for the transmission lines with singly-connected cross-sections, to be analyzed in the next section. The first structure for guiding waves was proposed by J. J. Thomson in 1893, and experimentally tested by O. Lodge in 1894.

[46] For the notation simplicity, I am dropping index ω in the complex amplitudes of the field components, and also have dropped the argument ω in k_z and Z, even though these parameters may depend on the wave's frequency rather substantially – see below.

Plugging Eqs. (98)-(99) into the source-free Maxwell equations (2), and requiring the longitudinal and transverse components to be balanced separately, we get

$$\begin{aligned} &ik_z \mathbf{n}_z \times \mathbf{E}_t - i\omega\mu \mathbf{H}_t = -\nabla_t \times (E_z \mathbf{n}_z), && ik_z \mathbf{n}_z \times \mathbf{H}_t + i\omega\varepsilon \mathbf{E}_t = -\nabla_t \times (H_z \mathbf{n}_z), \\ &\nabla_t \times \mathbf{E}_t = i\omega\mu H_z \mathbf{n}_z, && \nabla_t \times \mathbf{H}_t = -i\varepsilon\omega E_z \mathbf{n}_z, \\ &\nabla_t \cdot \mathbf{E}_t = -ik_z E_z, && \nabla_t \cdot \mathbf{H}_t = -ik_z H_z. \end{aligned} \tag{7.100}$$

where ∇_t is the 2D del operator acting in the transverse plane $[x, y]$ only, i.e. the usual ∇, but with $\partial/\partial z = 0$. The system (100) looks even bulkier than the original equations (2), but it is much simpler for analysis. Indeed, by eliminating the transverse components from these equations (or, even simpler, just by plugging Eq. (99) into Eqs. (3) and keeping only their z-components), we get a pair of self-consistent equations for the longitudinal components of the fields,[47]

2D Helmholtz equations for E_z and H_z

$$\left(\nabla_t^2 + k_t^2\right) E_z = 0, \qquad \left(\nabla_t^2 + k_t^2\right) H_z = 0, \tag{7.101}$$

where k is still defined by Eq. (13), $k \equiv (\varepsilon\mu)^{1/2}\omega$, while

Wave vector component balance

$$k_t^2 \equiv k^2 - k_z^2 = \omega^2 \varepsilon\mu - k_z^2. \tag{7.102}$$

After the distributions $E_z(x,y)$ and $H_z(x,y)$ have been found from these equations, they provide right-hand sides for the rather simple, closed system of equations (100) for the transverse components of field vectors. Moreover, as we will see below, each of the following three types of solutions:

(i) with $E_z = 0$ and $H_z = 0$ (called the *transverse electromagnetic*, or *TEM waves*),

(ii) with $E_z = 0$, but $H_z \neq 0$ (called either the *TE waves* or, more frequently, *H-modes*), and

(iii) with $E_z \neq 0$, but $H_z = 0$ (the so-called *TM waves* or *E-modes*),

has its own dispersion law and hence its own wave propagation velocity; as a result, these *modes* (i.e. the field distribution patterns) may be considered separately.

In the balance of this section, we will focus on the simplest, TEM waves (i), with *no longitudinal components* of either field. For them, the top two equations of the system (100) immediately give Eqs. (6) and (13), and $k_z = k$. In plain English, this means that $\mathbf{E} = \mathbf{E}_t$ and $\mathbf{H} = \mathbf{H}_t$ are proportional to each other and are mutually perpendicular (just as in the plane wave) at each point of the cross-section and that the TEM wave's impedance $Z \equiv E/H$ and dispersion law $\omega(k)$, and hence the propagation speed, are the same as in a plane wave in the same material. In particular, if ε and μ are frequency-independent within a certain frequency range, the dispersion law within this range is linear, $\omega = k/(\varepsilon\mu)^{1/2}$, and the wave's speed does not depend on its frequency. For practical applications to telecommunications, this is a very important advantage of the TEM waves over their TM and TE counterparts – to be discussed in the next sections.

Unfortunately for practice, such waves cannot propagate in every transmission line. To show this, let us have a look at the two last lines of Eqs. (100). For the TEM waves ($E_z = 0$, $H_z = 0$, $k_z = k$), they are reduced to merely

[47] The wave equation represented in the form (101), even with the 3D Laplace operator, is called the *Helmholtz equation*, named after Hermann von Helmholtz (1821-1894) – the mentor of H. Hertz and M. Planck, among many others.

$$\nabla_t \times \mathbf{E}_t = 0, \qquad \nabla_t \times \mathbf{H}_t = 0,$$
$$\nabla_t \cdot \mathbf{E}_t = 0, \qquad \nabla_t \cdot \mathbf{H}_t = 0. \tag{7.103}$$

Within the coarse-grain description of the conducting walls of the line (i. e., neglecting not only the screening depth but also the skin depth in comparison with the cross-section dimensions), we have to require that inside them, $\mathbf{E} = \mathbf{H} = 0$. Close to a wall but outside it, the normal component E_n of the electric field may be different from zero, because surface charges may sustain its jump – see Sec. 2.1, in particular Eq. (2.3). Similarly, the tangential component H_τ of the magnetic field may have a finite jump at the surface due to skin currents – see Sec. 6.3, in particular Eq. (6.38). However, the tangential component of the electric field and the normal component of the magnetic field cannot experience such jumps, and to have them equal to zero inside the walls they have to equal zero just outside the walls as well:

$$\mathbf{E}_\tau = 0, \qquad H_n = 0\,. \tag{7.104}$$

But the left columns of Eqs. (103)-(104) coincide with the formulation of the 2D boundary problem of electrostatics for the electric field induced by electric charges of the conducting walls, with the only difference that in our current case, the value of ε actually means $\varepsilon(\omega)$. Similarly, the right columns of those relations coincide with the formulation of the 2D boundary problem of magnetostatics for the magnetic field induced by currents in the walls, with $\mu \to \mu(\omega)$, with the only difference is that in our current coarse-grain approximation, the magnetic fields cannot penetrate into the conductors.

Now we immediately see that in waveguides with a singly-connected wall, for example, a hollow conducting tube (see, e.g., Fig. 16), the TEM waves are impossible, because there is no way to create a non-zero electrostatic field inside a conductor with such cross-section. However, such fields (and hence the TEM waves) are possible in structures with cross-sections consisting of two or more disconnected (galvanically-insulated) parts – see, e.g., Fig. 17.

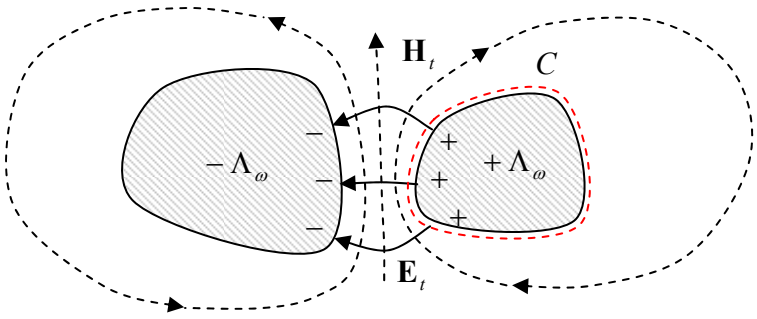


Fig. 7.17. An example of the cross-section of a transmission line that may support the TEM wave propagation.

In order to derive "global" relations for such a transmission line, let us consider the contour C drawn very close to the surface of one of its conductors – see, e.g., the red dashed line in Fig. 17. We can consider it, on one hand, as the cross-section of a cylindrically-shaped Gaussian volume of a certain elementary length $dz << \lambda \equiv 2\pi/k$. Using the generalized Gauss law (3.34), we get

$$\oint_C (\mathbf{E}_t)_n\, dr = \frac{\Lambda_\omega}{\varepsilon}, \tag{7.105}$$

where Λ_ω (not to be confused with the wavelength λ!) is the complex amplitude of the linear density of the electric charge of the conductor. On the other hand, the same contour C may be used in the generalized Ampère law (5.116) to write

$$\oint_C (\mathbf{H}_t)_\tau \, dr = I_\omega , \tag{7.106}$$

where I_ω is the total current flowing along the conductor (or rather its complex amplitude). But, as was mentioned above, in the TEM wave the ratio E_t/H_t of the field components participating in these two integrals is constant and equal to $Z = (\mu/\varepsilon)^{1/2}$, so Eqs. (105)-(106) give the following simple relation between the "global" variables of the conductor:

$$I_\omega = \frac{\Lambda_\omega / \varepsilon}{Z} \equiv \frac{\Lambda_\omega}{(\varepsilon\mu)^{1/2}} \equiv \frac{\omega}{k}\Lambda_\omega . \tag{7.107}$$

This important relation may be also obtained in a different way; let me describe it as well, because (as we will see below) it has an independent heuristic value. Let us consider a small segment $dz \ll \lambda = 2\pi/k$ of the line's conductor, and apply the electric charge conservation law (4.1) to the *instant* values of the linear charge density and current. The cancellation of dz in both parts yields

$$\frac{\partial \Lambda(z,t)}{\partial t} = -\frac{\partial I(z,t)}{\partial z} . \tag{7.108}$$

If we accept the sinusoidal waveform, $\exp\{i(kz - \omega t)\}$, for both these variables, we immediately recover Eq. (107) for their complex amplitudes, showing that this relation expresses just the charge continuity law.

The global equation (108) may be made more specific in the case when the frequency dependence of ε and μ is negligible, and the transmission line consists of just two isolated conductors – see, e.g., Fig. 17. In this case, to have the wave localized in the space near the two conductors, we need a sufficiently fast decrease of its electric field at large distances. For that, their linear charge densities for each value of z should be equal and opposite, and we can simply relate them to the potential difference V between the conductors:

$$\frac{\Lambda(z,t)}{V(z,t)} = C_0 , \tag{7.109}$$

where C_0 is the mutual capacitance of the conductors (per unit length) – which was repeatedly discussed in Chapter 2. Then Eq. (108) takes the following form:

$$C_0 \frac{\partial V(z,t)}{\partial t} = -\frac{\partial I(z,t)}{\partial z} . \tag{7.110}$$

Next, let us consider the contour shown with the red dashed line in Fig. 18 (which shows a different cross-section of the transmission line – by a plane containing the wave propagation axis z), and apply to it the Faraday induction law (6.3).

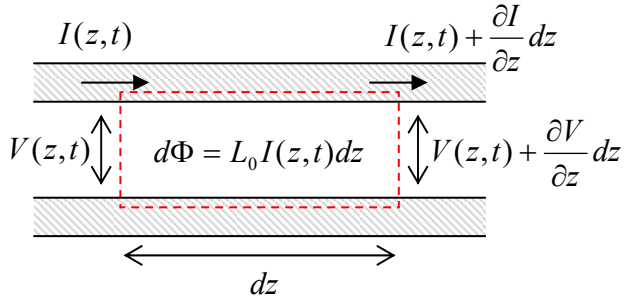


Fig. 7.18. Electric current, magnetic flux, and voltage in a two-conductor transmission line.

Since, in the coarse-grain approximation, the electric field inside the conductors (in Fig. 18, on the horizontal segments of the contour) vanishes, the total e.m.f. equals the difference of the voltages V at the ends of the segment dz, while the only sources of the magnetic flux through the area limited by the contour are the (equal and opposite) currents $\pm I$ in the conductors, we can use Eq. (5.70) to express the flux. As a result, by canceling dz in both parts of the equation, we get

$$L_0 \frac{\partial I(z,t)}{\partial t} = -\frac{\partial V(z,t)}{\partial z}, \tag{7.111}$$

where L_0 is the mutual inductance of the conductors per unit length. The only difference between this L_0 and the dc mutual inductances discussed in Chapter 5 is that at the high frequencies we are analyzing now, L_0 should be calculated neglecting the magnetic field penetration into the conductors. (In the dc case, we had the same situation for superconductor electrodes within their coarse-grain, ideal-diamagnet description.)

The system of Eqs. (110) and (111) is frequently called the *telegrapher's equations*. Combined, they give for any "global" variable f (either V, or I, or Λ) the usual 1D wave equation,

$$\frac{\partial^2 f}{\partial z^2} - L_0 C_0 \frac{\partial^2 f}{\partial t^2} = 0, \tag{7.112}$$

which describes dispersion-free TEM wave's propagation. Again, this equation is only valid within the frequency range where the frequency dependence of both ε and μ is negligible. If this is not so, the global approach may still be used for sinusoidal waves $f = \mathrm{Re}[f_\omega \exp\{i(kz - \omega t)\}]$. Repeating the above arguments, instead of Eqs. (110)-(111) we get a more general system of two algebraic equations

$$\omega C_0 V_\omega = k I_\omega, \qquad \omega L_0 I_\omega = k V_\omega, \tag{7.113}$$

in which $L_0 \propto \mu$ and $C_0 \propto \varepsilon$ may now depend on frequency. These equations are consistent only if

$$\boxed{L_0 C_0 = \frac{k^2}{\omega^2} \equiv \frac{1}{v^2} \equiv \varepsilon\mu}. \tag{7.114}$$

L_0C_0 product invariance

Besides the fact we have already known (that the TEM wave's speed is the same as that of the plane wave), Eq. (114) gives us the result that I confess was not emphasized enough in Chapter 5: the product L_0C_0 does not depend on the shape or size of line's cross-section, provided that the magnetic field's penetration into the conductors is negligible). Hence, if we have calculated the mutual capacitance C_0 of a system of two cylindrical conductors, the result immediately gives us their mutual inductance: $L_0 = \varepsilon\mu/C_0$. This relationship stems from the fact that both the electric and magnetic fields may be expressed via the solution of the same 2D Laplace equation for the system's cross-section.

With Eq. (114) satisfied, any of Eqs. (113) gives the same result for the following ratio:

$$\boxed{Z_W \equiv \frac{V_\omega}{I_\omega} = \left(\frac{L_0}{C_0}\right)^{1/2}}, \tag{7.115}$$

Transmission line's TEM Impedance

which is called the *transmission line's impedance*. This parameter has the same dimensionality (in SI units – ohms, denoted Ω) as the wave impedance (7),

$$Z \equiv \frac{E_\omega}{H_\omega} = \left(\frac{\mu}{\varepsilon}\right)^{1/2}, \tag{7.116}$$

but these parameters should not be confused, because Z_W depends on the cross-section's geometry, while Z does not. In particular, Z_W is the only important parameter of a transmission line for its matching with a lumped load circuit (Fig. 19) in the important case when both the cable cross-section's size and the load's linear dimensions are much smaller than the wavelength.[48]

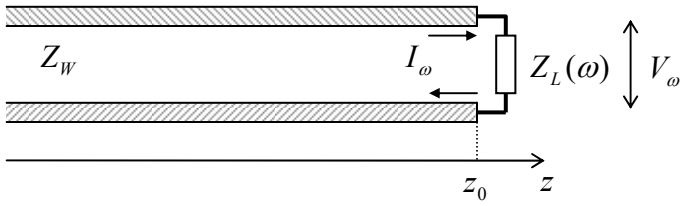


Fig. 7.19. Passive, lumped termination of a TEM transmission line.

Indeed, in this case, we may consider the load in the quasistatic limit and write

$$V_\omega(z_0) = Z_L(\omega) I_\omega(z_0) \ , \tag{7.117}$$

where $Z_L(\omega)$ is the (generally complex) impedance of the load. Taking $V(z,t)$ and $I(z,t)$ in the form similar to Eqs. (61) and (62), and writing the two Kirchhoff's circuit laws for the point $z = z_0$, we get for the reflection coefficient a result similar to Eq. (68):

$$R = \frac{Z_L(\omega) - Z_W}{Z_L(\omega) + Z_W}. \tag{7.118}$$

This formula shows that for the perfect matching (i.e. the total wave absorption in the load), the load's impedance $Z_L(\omega)$ should be real and equal to Z_W – but not necessarily to Z.

As an example, let us consider one of the simplest (and most practically important) transmission lines: the coaxial cable (Fig. 20).[49]

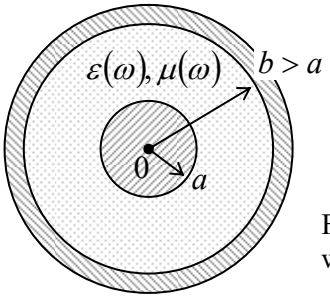


Fig. 7. 20. The cross-section of a coaxial cable with (possibly, dispersive) dielectric filling.

For this geometry, we already know the expressions for both C_0 and L_0,[50] though they have to be modified for the account of arbitrary dielectric and magnetic constants, and the magnetic field's non-penetration into the conductors. As a result of this (elementary) modification, we get the formulas,

[48] The ability of TEM lines to have such a small cross-section is another important practical advantage.
[49] It was invented by the same O. Heaviside in 1880.
[50] See, respectively, Eqs. (2.49) and (5.79).

$$C_0 = \frac{2\pi\varepsilon}{\ln(b/a)}, \qquad L_0 = \frac{\mu}{2\pi}\ln(b/a), \tag{7.119}$$

Coaxial cable's C_0 and L_0

illustrating that the universal relationship (114) is indeed valid. For the cable's impedance (115), Eqs. (119) yield a geometry-dependent value

$$Z_W = \left(\frac{\mu}{\varepsilon}\right)^{1/2}\frac{\ln(b/a)}{2\pi} \equiv Z\frac{\ln(b/a)}{2\pi} \neq Z\,. \tag{7.120}$$

For the standard TV antenna cables (such as RG-6/U, with $b/a \sim 3$, $\varepsilon/\varepsilon_0 \approx 2.2$), $Z_W = 75\ \Omega$, while for most computer component connections, coaxial cables with $Z_W = 50\ \Omega$ (such as RG-58/U) are prescribed by electronic engineering standards. Such cables are broadly used for the transmission of electromagnetic waves with frequencies up to 1 GHz over distances of a few km, and up to ~20 GHz on the tabletop scale (a few meters), limited by wave attenuation – see Sec. 9 below.

Moreover, the following two facts enable a wide application, in electrical engineering and physical experiment, of coaxial-cable-*like* systems. First, as Eq. (5.78) shows, in a cable with $a << b$, most energy of the wave is localized near the internal conductor. Second, the theory to be discussed in the next section shows that excitation of other (H- and E-) waves in the cable is impossible until the wavelength λ becomes smaller than $\sim\pi(a + b)$. As a result, the TEM mode propagation in a cable with $a << b < \lambda/\pi$ is not much affected even if the internal conductor is not straight, but bent – for example, into a helix – see, e.g., Fig. 21.

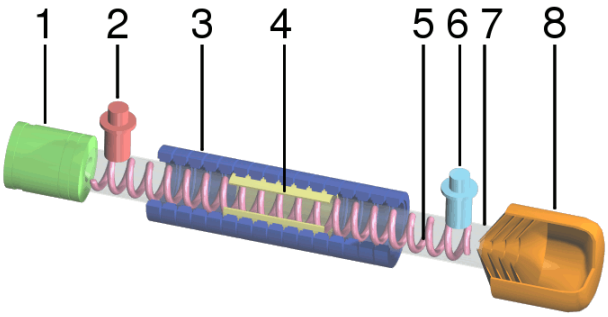


Fig. 7.21. A typical traveling-wave tube: (1) electron gun, (2) ac input, (3) beam-focusing magnets, (4) wave attenuator, (5) helix coil, (6) ac output, (7) vacuum tube, (8) electron collector. Adapted from https://en.wikipedia.org/wiki/Traveling-wave_tube under the Creative Commons BY-SA 3.0 license.

In such a system, called the *traveling-wave tube* (TWT), a quasi-TEM wave propagates with velocity $v \approx c$ along the helix's length, so the velocity's component along the cable's axis may be made close to the velocity $u << c$ of the electron beam moving ballistically along the tube's axis, enabling their effective interaction, and as a result, a length-accumulating amplification of the wave.[51]

Another important example of a TEM transmission line is a set of two parallel wires. In the form of *twisted pairs*,[52] they allow communications, in particular long-range telephone and DSL Internet

[51] Despite the current prevalence of semiconductor devices in electronics, TWTs are still used in satellite TV and radio systems, because they may work at very high microwave power – e.g., up to 200W at 20 GHz and pulsed 50W at 200 GHz. Very unfortunately, in this course, I will not have time/space to discuss even the (rather elegant) basic theory of such devices. The reader interested in this field may be referred, for example, to the detailed monograph by J. Whitaker, *Power Vacuum Tubes Handbook*, 3rd ed., CRC Press, 2017.

[52] Such twisting, around the line's direction axis, reduces the crosstalk between adjacent lines, and the parasitic radiation at their bends.

connections, at frequencies up to a few hundred kHz, as well as relatively short, multi-line Ethernet and TV cables at frequencies up to ~ 1 GHz, limited mostly by the mutual interference ("crosstalk") between the individual lines of the same cable, and the unintentional radiation of the wave into the environment.

7.6. Waveguides: *H* and *E* waves

Let us now return to Eqs. (100) and explore the *H*- and *E*-waves – with, respectively, either H_z or E_z different from zero. At the first sight, they may seem more complex. However, Eqs. (101), which determine the distribution of these longitudinal components over the cross-section, are just the 2D Helmholtz equations for scalar functions. For simple cross-section geometries, they may be readily solved using the methods discussed for the Laplace equation in Chapter 2, in particular the variable separation. After the solution of such an equation has been found, the transverse components of the fields may be calculated by differentiation, using the simple formulas,

$$\mathbf{E}_t = \frac{i}{k_t^2}\left[k_z \nabla_t E_z - kZ(\mathbf{n}_z \times \nabla_t H_z)\right], \qquad \mathbf{H}_t = \frac{i}{k_t^2}\left[k_z \nabla_t H_z + \frac{k}{Z}(\mathbf{n}_z \times \nabla_t E_z)\right], \tag{7.121}$$

which follow from the first line of Eqs. (100).[53]

In comparison with the boundary problems of electro- and magnetostatics, the only conceptually new feature of Eqs. (101) is that they form the so-called *eigenproblems*, with typically many solutions (*eigenfunctions*), each describing a specific wave mode, and corresponding to a specific *eigenvalue* of the parameter k_t. The good news here is that these values of k_t are determined by this 2D boundary problem and hence do not depend on k_z. As a result, the dispersion law $\omega(k_z)$ of any mode, which follows from the last form of Eq. (102),

Universal dispersion relation

$$\omega = \left(\frac{k_z^2 + k_t^2}{\varepsilon\mu}\right)^{1/2} \equiv \left(v^2 k_z^2 + \omega_c^2\right)^{1/2}, \tag{7.122}$$

is functionally similar for all modes. It is also similar to that of plane waves in plasma (see Eq. (38), Fig. 6, and their discussion in Sec. 2), with the only difference that the speed in light *c* is generally replaced with $v = 1/(\varepsilon\mu)^{1/2}$ – the speed of the plane or TEM waves in the medium filling the waveguide, and that ω_p is replaced with the so-called *cutoff frequency*

$$\omega_c \equiv v k_t, \tag{7.123}$$

specific for each mode. (As Eq. (101) implies, and as we will see from several examples below, k_t has the order of $1/a$, where *a* is the characteristic dimension of the waveguide's cross-section, so the critical value of the free-space wavelength $\lambda \equiv 2\pi c/\omega$ is of the order of *a*.) Below the cutoff frequency of each particular mode, this wave cannot propagate in the waveguide.[54] As a result, the modes with the *lowest*

[53] For the derivation of Eqs. (121), one of these two linear equations should be first vector-multiplied by $\mathbf{n}_z$. Note also that this approach could not be used to analyze the TEM waves, because for them $k_t = 0$, $E_z = 0$, $H_z = 0$, and Eqs. (121) yield uncertainty.

[54] An interesting twist in the ideas of electromagnetic metamaterials (mentioned in Sec. 5 above) is the so-called *ε-near-zero* materials, designed to have the effective product $\varepsilon\mu$ much lower than $\varepsilon_0\mu_0$ within certain frequency ranges. Since at these frequencies, the speed *v* (4) becomes much lower than *c*, the cutoff frequency (123)

values of ω_c present special practical interest, because the choice of the signal frequency ω between the two lowest values of the cutoff frequency (123) guarantees that the waves propagate in the form of only one mode, with the lowest k_t. Such a choice enables engineers to simplify the excitation of the desired mode by wave generators and to avoid the unintentional transfer of electromagnetic wave energy to undesirable modes by (virtually unavoidable) small inhomogeneities of the system.

The boundary conditions for the Helmholtz equations (101) depend on the propagating wave type. For the *E*-modes, with $H_z = 0$ but $E_z \neq 0$, the condition $E_\tau = 0$ immediately gives

$$E_z\big|_C = 0, \tag{7.124}$$

where *C* is the inner contour limiting the conducting wall's cross-section. For the *H*-modes, with $E_z = 0$ but $H_z \neq 0$, the boundary condition is slightly less obvious and may be obtained using, for example, the second equation of the system (100), vector-multiplied by $\mathbf{n}_z$. Indeed, for the component normal to the conductor surface, the result of such multiplication is

$$ik_z(\mathbf{H}_t)_n - i\frac{k}{Z}\left(\mathbf{n}_z \times \mathbf{E}_t\right)_n = \frac{\partial H_z}{\partial n}. \tag{7.125}$$

But the first term on the left-hand side of this relation must be zero on the wall surface, because of the second of Eqs. (104), while according to the first of Eqs. (104), the vector $\mathbf{E}_t$ in the second term cannot have a component tangential to the wall. As a result, the vector product in that term cannot have a normal component, so the term should equal zero as well, and Eq. (125) is reduced to

$$\frac{\partial H_z}{\partial n}\Big|_C = 0\,. \tag{7.126}$$

Let us see how all this machinery works for a simple but practically important case of a metallic-wall waveguide with a rectangular cross-section – see Fig. 22

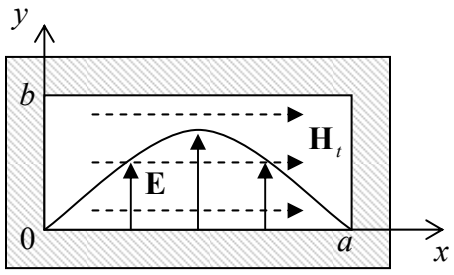


Fig. 7.22. A rectangular waveguide, and the transverse field distribution in its fundamental mode H_{10} (schematically).

In the natural Cartesian coordinates shown in this figure, both Eqs. (101) take the simple form

$$\left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + k_t^2\right) f = 0, \qquad \text{where } f = \begin{cases} E_z, & \text{for } E\text{-modes}, \\ H_z, & \text{for } H\text{-modes}. \end{cases} \tag{7.127}$$

From Chapter 2, we know that the most effective way of solving such equations in a rectangular region is the variable separation, in which the general solution is represented as a sum of partial solutions of the type

---

virtually vanishes. As a result, the waves may "tunnel" through very narrow sections of metallic waveguides filled with such materials – see, e.g., M. Silveirinha and N. Engheta, *Phys. Rev. Lett.* **97**, 157403 (2006).

$$f = X(x)Y(y)\,. \tag{7.128}$$

Plugging this expression into Eq. (127), and dividing each term by XY, we get the equation,

$$\frac{1}{X}\frac{d^2X}{dx^2} + \frac{1}{Y}\frac{d^2Y}{dy^2} + k_t^2 = 0\,, \tag{7.129}$$

which should be satisfied for all values of x and y within the waveguide's interior. This is only possible if each term of the sum equals a constant. Taking the X-term and Y-term constants in the form $(-k_x^2)$ and $(-k_y^2)$, respectfully, and solving the corresponding ordinary differential equations,[55] for the eigenfunction (128) we get

$$f = (c_x \cos k_x x + s_x \sin k_x x)(c_y \cos k_y y + s_y \sin k_y y), \qquad \text{with } k_x^2 + k_y^2 = k_t^2\,, \tag{7.130}$$

where the constants c and s should be found from the boundary conditions. Here the difference between the H-modes and E-modes kicks in.

For the H-modes, Eq. (130) is valid for H_z, and we should use the boundary condition (126) on all metallic walls of the waveguide, i.e. at $x = 0$ and a; and $y = 0$ and b – see Fig. 22. As a result, we get very simple expressions for eigenfunctions and eigenvalues:

$$(H_z)_{nm} = H_l \cos\frac{\pi n x}{a}\cos\frac{\pi m y}{b}\,, \tag{7.131}$$

$$k_x = \frac{\pi n}{a}, \quad k_y = \frac{\pi m}{b}, \quad \text{so that } (k_t)_{nm} = \left(k_x^2 + k_y^2\right)^{1/2} = \pi\left[\left(\frac{n}{a}\right)^2 + \left(\frac{m}{b}\right)^2\right]^{1/2}, \tag{7.132}$$

where H_l is the longitudinal field's amplitude, and n and m are two integer numbers – each of them arbitrary besides that they cannot be equal to zero simultaneously.[56] Assuming, just for certainty, that $a \geq b$ (as shown in Fig. 22), we see that the lowest eigenvalue of k_t and hence the lowest cutoff frequency (123) are achieved for the so-called H_{10} *mode* with $n = 1$ and $m = 0$, and hence with

Fundamental mode's cutoff

$$\boxed{(k_t)_{10} = \frac{\pi}{a}}, \tag{7.133}$$

thus confirming our prior estimate of k_t.

Depending on the a/b ratio, the second-lowest k_t (and hence ω_c) belongs to either the H_{11} mode with $n = 1$ and $m = 1$:

$$(k_t)_{11} = \pi\left(\frac{1}{a^2} + \frac{1}{b^2}\right)^{1/2} \equiv \left[1 + \left(\frac{a}{b}\right)^2\right]^{1/2}(k_t)_{10}\,, \tag{7.134}$$

or to the H_{20} mode with $n = 2$ and $m = 0$:

[55] Let me hope that the solution of equations of the type $d^2X/dx^2 + k_x^2X = 0$ does not present any problem for the reader, at least due to their prior experience with problems such as standing waves on a guitar string, wavefunctions in a flat 1D quantum well, or (with the replacement $x \to t$) a classical harmonic oscillator.

[56] Otherwise, the function $H_z(x,y)$ would be constant, so, according to Eq. (121), the transverse components of the electric and magnetic field would equal zero. As a result, as the last two lines of Eqs. (100) show, the whole field would be zero for any $k_z \neq 0$.

$$(k_t)_{20} = \frac{2\pi}{a} \equiv 2(k_t)_{10}. \tag{7.135}$$

These values become equal at $a/b = \sqrt{3} \approx 1.7$; in practical waveguides, the a/b ratio is not too far from this value. For example, in the standard X-band (~10-GHz) waveguide WR90, $a \approx 2.3$ cm ($f_c \equiv \omega_c/2\pi \approx$ 6.5 GHz), and $b \approx 1.0$ cm.

Now let us have a look at the alternative *E*-modes. For them, we still should use the general solution (130) with $f = E_z$, but now with the boundary condition (124). This gives us the eigenfunctions

$$(E_z)_{nm} = E_l \sin\frac{\pi n x}{a} \sin\frac{\pi m y}{b}, \tag{7.136}$$

and the same eigenvalue spectrum (132) as for the *H* modes. However, now neither n nor m can be equal to zero; otherwise, Eq. (136) would give the trivial solution $E_z(x,y) = 0$. Hence the lowest cutoff frequency of TM waves is achieved at the so-called E_{11} *mode* with n =1, m = 1, and with the eigenvalue given by Eq. (134), always higher than $(k_t)_{10}$.

Thus the fundamental H_{10} mode is certainly the most important wave in rectangular waveguides; let us have a better look at this field distribution. Plugging the corresponding solution (131) with n = 1 and m = 0 into the general relation (121), we easily get

$$(H_x)_{10} = -i\frac{k_z a}{\pi} H_l \sin\frac{\pi x}{a}, \qquad (H_y)_{10} = 0, \tag{7.137}$$

$$(E_x)_{10} = 0, \qquad (E_y)_{10} = i\frac{ka}{\pi} Z H_l \sin\frac{\pi x}{a}. \tag{7.138}$$

This field distribution is (schematically) shown in Fig. 22. Neither of the fields depends on the coordinate y – the feature very convenient, in particular, for microwave experiments with small samples. The electric field has only one (in Fig. 22, vertical) component that vanishes at the side walls and reaches its maximum at the waveguide's center; its field lines are straight, starting and ending on wall surface charges (whose distribution propagates along the waveguide together with the wave). In contrast, the magnetic field has two non-zero components (H_x and H_z), and its field lines are shaped as horizontal loops wrapped around the electric field maxima.

An important question is whether the H_{10} wave may be usefully characterized by a unique impedance introduced similarly to Z_W of the TEM modes – see Eq. (115). The answer is *not*, because the main value of Z_W is a convenient description of the impedance matching of a transmission line with a lumped load – see Fig. 19 and Eq. (118). As was discussed above, such a simple description is possible (i.e., does not depend on the exact geometry of the connection) only if both dimensions of the line's cross-section are much less than λ. But for the H_{10} wave (and more generally, any non-TEM mode) this is impossible – see, e.g., Eq. (129): its lowest frequency corresponds to the TEM wavelength $\lambda_{max} = 2\pi/(k_t)_{min} = 2\pi/(k_t)_{10} = 2a$. (The reader is challenged to find a simple interpretation of this equality.)

Now let us consider metallic-wall waveguides with a round cross-section (Fig. 23a). In this single-connected geometry, the TEM waves are impossible again, while for the analysis of *H*-modes and *E*-modes, the polar coordinates $\{\rho, \varphi\}$ are most natural. In these coordinates, the 2D Helmholtz equation (101) takes the following form:

$$\left[\frac{1}{\rho}\frac{\partial}{\partial\rho}\left(\rho\frac{\partial}{\partial\rho}\right)+\frac{1}{\rho^2}\frac{\partial^2}{\partial\varphi^2}+k_t^2\right]f=0, \qquad \text{where } f=\begin{cases}E_z, & \text{for } E\text{ - modes,}\\ H_z, & \text{for } H\text{ - modes.}\end{cases} \tag{7.139}$$

Separating the variables as $f = \mathcal{R}(\rho)\mathcal{F}(\varphi)$, we get

$$\frac{1}{\rho\mathcal{R}}\frac{d}{d\rho}\left(\rho\frac{d\mathcal{R}}{d\rho}\right)+\frac{1}{\rho^2\mathcal{F}}\frac{d^2\mathcal{F}}{d\varphi^2}+k_t^2=0\,. \tag{7.140}$$

But this is exactly the Eq. (2.127) that was studied in Sec. 2.7 in the context of electrostatics, just with a replacement of notation: $\gamma \to k_t$. So we already know that to have 2π-periodic functions $\mathcal{F}(\varphi)$ and finite values $\mathcal{R}(0)$ (which are evidently necessary for our current case – see Fig. 23a), the general solution must have the form given by Eq. (2.136), i.e. the eigenfunctions are expressed via integer-order Bessel functions of the first kind:

$$f_{nm}=J_n(k_{nm}\rho)(c_n\cos n\varphi+s_n\sin n\varphi)\equiv \text{const}\times J_n(k_{nm}\rho)\cos n(\varphi-\varphi_0), \tag{7.141}$$

with the eigenvalues k_{nm} of the transverse wave number k_t to be determined from appropriate boundary conditions, and an arbitrary constant φ_0.

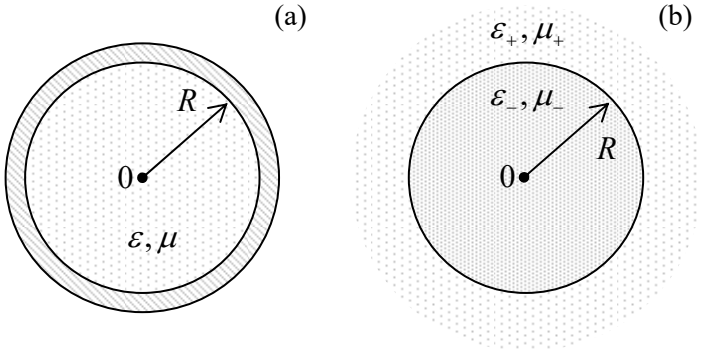


Fig. 7.23. (a) Metallic and (b) dielectric waveguides with circular cross-sections.

As for the rectangular waveguide, let us start with the H-modes ($f = H_z$). Then the boundary condition on the wall surface ($\rho = R$) is given by Eq. (126), which, for the solution (141), takes the form

$$\frac{d}{d\xi}J_n(\xi)=0, \qquad \text{where } \xi\equiv kR\,. \tag{7.142}$$

This means that the eigenvalues of Eq. (139) are

$$k_t=k_{nm}=\frac{\xi'_{nm}}{R}, \tag{7.143}$$

where ξ'_{nm} is the m^{th} zero of the function $dJ_n(\xi)/d\xi$. Approximate values of these zeros for several lowest n and m may be read out from Fig. 2.18; their more accurate values are given in Table 1 below.

Table 7.1. Zeros ξ'_{nm} of the function $dJ_n(\xi)/d\xi$ for a few lowest values of the Bessel function's index n and the root's number m.

	$m = 1$	2	3
$n = 0$	3.83171	7.015587	10.1735
1	1.84118	5.33144	8.53632
2	3.05424	6.70613	9.96947
3	4.20119	8.01524	11.34592

The table shows, in particular, that the lowest of the zeros is $\xi'_{11} \approx 1.84$.[57] Thus, perhaps a bit counter-intuitively, the fundamental mode, providing the lowest cutoff frequency $\omega_c = vk_{nm}$, is H_{11}, corresponding to $n = 1$ rather than $n = 0$:

$$H_z = H_l J_1\left(\xi'_{11}\frac{\rho}{R}\right)\cos(\varphi - \varphi_0). \tag{7.144}$$

It has the transverse wave number is $k_t = k_{11} = \xi'_{11}/R \approx 1.84/R$, and hence the cutoff frequency corresponding to the TEM wavelength $\lambda_{max} = 2\pi/k_{11} \approx 3.41\ R$. Thus the ratio of λ_{max} to the waveguide's diameter $2R$ is about 1.7, i.e. is close to the ratio $\lambda_{max}/a = 2$ for the rectangular waveguide. The origin of this proximity is clear from Fig. 24, which shows the transverse field distribution in the H_{11} mode. (It may be readily calculated from Eqs. (121) with $E_z = 0$ and H_z given by Eq. (144).)

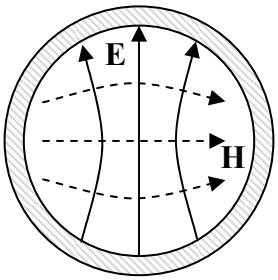


Fig. 7.24. Transverse field components in the fundamental H_{11} mode of a metallic, circular waveguide (schematically).

One can see that the field structure is actually very similar to that of the fundamental mode in the rectangular waveguide, shown in Fig. 22, despite the different nomenclature (which is due to the different coordinate system used for the solution). However, note the arbitrary constant angle φ_0, indicating that in circular waveguides, the transverse field's polarization is arbitrary. For some practical applications, such degeneracy of these "quasi-linearly-polarized" waves creates problems; some of them may be avoided by using waves with circular polarization.

As Table 1 shows, the next lowest H-mode is H_{21}, for which $k_t = k_{21} = \xi'_{21}/R \approx 3.05/R$, almost twice larger than that of the fundamental mode, and only then comes the first mode with no angular dependence of any field, H_{01}, with $k_t = k_{01} = \xi'_{01}/R \approx 3.83/R$,[58] followed by several angle-dependent modes: H_{31}, H_{12}, etc.

For the E modes, we may still use Eq. (141) (with $f = E_z$), but with the boundary condition (124) at $\rho = R$. This gives the following equation for the problem eigenvalues:

$$J_n(k_{nm}R) = 0, \quad \text{i.e. } k_{nm} = \frac{\xi_{nm}}{R}, \tag{7.145}$$

where ξ_{nm} is the m^{th} zero of function $J_n(\xi)$ – see Table 2.1. That table shows that the lowest k_t is equal to $\xi_{01}/R \approx 2.405/R$. Hence the corresponding mode (E_{01}), with no angular dependence of its fields, e.g.

[57] Mathematically, the lowest root of Eq. (142) with $n = 0$ equals 0. However, it would yield $k = 0$ and hence a constant field H_z, which, according to the first of Eqs. (121), would give a vanishing electric field.

[58] The electric field lines in the H_{01} mode (as well as all higher H_{0m} modes) are directed straight from the symmetry axis to the walls, reminding those of the TEM waves in the coaxial cable. Due to this property, these modes provide, at $\omega >> \omega_c$, much lower energy losses (see Sec. 9 below) than the fundamental H_{11} mode, and are sometimes used in practice, despite the inconvenience of working in the multimode frequency range.

$$E_z = E_l J_0\left(\xi_{01}\frac{\rho}{R}\right), \tag{7.146}$$

has the second-lowest cutoff frequency, ~30% higher than that of the fundamental mode H_{11}.

Finally, let us discuss one more topic of general importance – the number N of electromagnetic modes that may propagate in a waveguide within a certain range of relatively large frequencies $\omega >> \omega_c$. It is easy to calculate for a rectangular waveguide, with its simple expressions (132) for the eigenvalues of $\{k_x, k_y\}$. Indeed, these expressions describe a rectangular mesh on the $[k_x, k_y]$ plane, so each point corresponds to the plane area $\Delta A_k = (\pi/a)(\pi/b)$, and the number of modes in a large k-plane area $A_k >> \Delta A_k$ is $N = A_k/\Delta A_k = abA_k/\pi^2 = AA_k/\pi^2$, where A is the waveguide's cross-section area.[59] However, it is frequently more convenient to discuss transverse wave vectors $\mathbf{k}_t$ of arbitrary direction, i.e. with an arbitrary sign of their components k_x and k_y. Taking into account that the opposite values of each component actually give the same wave, the actual number of different modes of each type (E- or H-) is a factor of $2^2 \equiv 4$ lower than was calculated above. This means that the number of modes of *both* types is

$$N = 2\frac{A_k A}{(2\pi)^2}. \tag{7.147}$$

Let me leave it for the reader to find hand-waving (but convincing :-) arguments that this *mode counting rule* is valid for waveguides with cross-sections of any shape, and any boundary conditions on the walls, provided that $N >> 1$.

7.7. Dielectric waveguides, optical fibers, and paraxial beams

Now let us discuss electromagnetic wave propagation in *dielectric waveguides*. The simplest, *step-index* waveguide (see Figs. 23b and 25) consists of an inner *core* and an outer shell (in the optical fiber technology lingo, called *cladding*) with a higher wave propagation speed, i.e. a lower index of refraction:

$$v_+ > v_-, \quad \text{i.e. } n_+ < n_-, \quad k_+ < k_-, \quad \varepsilon_+\mu_+ < \varepsilon_-\mu_-. \tag{7.148}$$

at the same frequency. (In most cases the difference is achieved due to that in the electric permittivity, $\varepsilon_+ < \varepsilon_-$, while magnetically both materials are virtually passive: $\mu_- \approx \mu_+ \approx \mu_0$, so their refraction indices $n_\pm$, defined by Eq. (84), are very close to $(\varepsilon_\pm/\varepsilon_0)^{1/2}$; I will limit my discussion to this approximation.)

The basic idea of the waveguide's operation may be readily understood in the limit when the wavelength λ is much smaller than the characteristic size R of the core's cross-section. In this "geometric-optics" limit, at distances of the order of λ from the core-cladding interface, which provides the wave reflection, we can neglect the interface's curvature and approximate its geometry with a plane. As we know from Sec. 4, if the angle θ of the wave's incidence on such a plane interface is larger than the critical value θ_c specified by Eq. (85), the wave is totally reflected. As a result, the waves launched into the fiber core at such "grazing" angles, propagate inside the core, being repeatedly reflected from the cladding – see Fig. 25.

[59] This formula ignores the fact that, according to the above analysis, some modes (with $n = 0$ *and* $m = 0$ for the H modes, and $n = 0$ *or* $m = 0$ for the E modes) are forbidden. However, for $N >> 1$, the associated corrections of Eq. (147) are negligible.

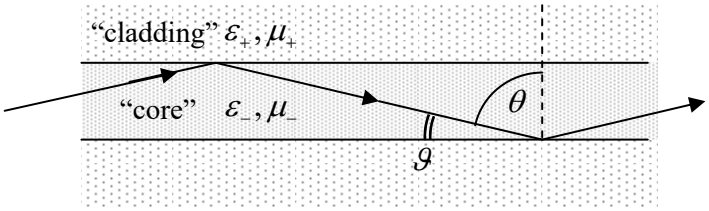


Fig. 7.25. Wave propagation in a thick optical fiber at $\theta > \theta_c$.

The most important type of dielectric waveguides is *optical fibers*.[60] Due to a heroic technological effort over three decades starting from the mid-1960s, the attenuation of such fibers has been decreased from values of the order of 20 db/km (typical for a window glass) to the fantastically low values of ~0.2 db/km (meaning virtually perfect transparency of 10-km-long fiber segments!), combined with the extremely low plane-wave ("chromatic") dispersion below 10 ps/km·nm.[61] In conjunction with the development of inexpensive erbium-based quantum amplifiers, this breakthrough has enabled inter-city and inter-continental (undersea), broadband[62] optical cables, which are the backbone of all modern telecommunication infrastructure.

The only bad news is that these breakthroughs were achieved for just one kind of materials (silica-based glasses)[63] within a very narrow range of their chemical composition. As a result, the dielectric constants $\kappa_\pm \equiv \varepsilon_\pm/\varepsilon_0$ of the cladding and core of practical optical fibers are both close to 2.2 (giving $n_\pm \approx 1.5$) and hence very close to each other, so the relative difference of the refraction indices,

$$\Delta \equiv \frac{n_- - n_+}{n_-} = \frac{\varepsilon_-^{1/2} - \varepsilon_+^{1/2}}{\varepsilon_-^{1/2}} \approx \frac{\varepsilon_- - \varepsilon_+}{2\varepsilon_\pm}, \tag{7.149}$$

is typically below 0.5%. This factor limits the fiber bandwidth. Indeed, let us use the geometric-optics picture to calculate the number of quasi-plane-wave modes that may propagate in the fiber. For the complementary angle (Fig. 25)

$$\vartheta \equiv \frac{\pi}{2} - \theta, \qquad \text{so that } \sin\theta = \cos\vartheta, \tag{7.150}$$

Eq. (85) gives the following propagation condition:

$$\cos\vartheta > \frac{n_+}{n_-} = 1 - \Delta \cdot \tag{7.151}$$

[60] For a comprehensive discussion of this vital technology see, e.g., A. Yariv and P. Yeh, *Photonics*, 6th ed., Oxford U. Press, 2007.

[61] Both these parameters have their best values not in the visible light range (with wavelengths from 380 to 740 nm), but in the near-infrared, with the attenuation lowest between approximately 1,500 and 1,630 nm. As a result, most modern communication systems use two spectral windows – the so-called C-band (1,530-1,565 nm) and L-band (1,570-1,610 nm) within that range.

[62] Each of the spectral bands mentioned above, at a typical signal-to-noise ratio $S/N > 10^5$, corresponds to the Shannon bandwidth $\Delta f \log_2(S/N)$ exceeding 10^{14} bits per second, some five orders of magnitude (!) higher than that of a modern Ethernet cable. The practically usable bandwidth of each fiber is somewhat lower, but a typical optical cable, with many fibers in parallel, has a proportionately higher aggregate bandwidth. A relatively recent (circa 2017) example is the C-band transatlantic (6,600-km-long) cable *Marea*, with eight fiber pairs and an aggregate useable bandwidth of 160 terabits per second.

[63] The silica-based fibers were developed in 1966 by an industrial research group led by Charles Kao (who shared the 2009 Nobel Prize in physics), but the very idea of using optical fibers for long-range communications may be traced back at least to the 1963 work by Jun-ichi Nishizawa – who also invented semiconductor lasers.

In the limit $\Delta << 1$, when the incidence angles $\theta > \theta_c$ of all propagating waves are very close to $\pi/2$, and hence the complementary angles are ϑ small, we may keep only two first terms in the Taylor expansion of the left-hand side of Eq. (151) and get

$$\vartheta_{\max}^2 \approx 2\Delta . \tag{7.152}$$

(Even for the higher-end value $\Delta = 0.005$, this critical angle is only ~0.1 radian, i.e. is close to 5°.) Due to this smallness, we may approximate the maximum transverse component of the wave vector as

$$(k_t)_{\max} = k(\sin\vartheta)_{\max} \approx k\vartheta_{\max} \approx \sqrt{2}k\Delta , \tag{7.153}$$

and use Eq. (147) to calculate the number N of propagating modes:

$$N \approx 2\frac{(\pi R^2)(\pi k^2 \vartheta_{\max}^2)}{(2\pi)^2} = (kR)^2\Delta . \tag{7.154}$$

For typical values $k = 0.73\times10^7$ m^{-1} (corresponding to the free-space wavelength $\lambda_0 = n\lambda = 2\pi n/k \approx 1.3$ μm), $R = 25$ μm, and $\Delta = 0.005$, this formula gives $N \approx 150$.

Now we can calculate the *geometric dispersion* of such a fiber, i.e. the difference in the mode propagation speed, which is commonly characterized in terms of the difference between the wave delay times (traditionally measured in picoseconds per kilometer) of the fastest and slowest modes. Within the geometric optics approximation, the difference in time delays of the fastest mode (with $k_z = k$) and the slowest mode (with $k_z = k\sin\theta_c$) at distance l is

$$\Delta t = \Delta\left(\frac{l}{v_z}\right) = \Delta\left(\frac{k_z l}{\omega}\right) = \frac{l}{\omega}\Delta k_z = \frac{l}{v}(1-\sin\theta_c) = \frac{l}{v}\left(1-\frac{n_+}{n_-}\right) \equiv \frac{l}{v}\Delta . \tag{7.155}$$

For the example considered above, the TEM wave's speed in the glass, $v = c/n \approx 2\times10^8$ m/s, and the geometric dispersion $\Delta t/l$ is close to 25 ps/m, i.e. 25,000 ps/km. (This means, for example, that a 1-ns pulse, being distributed between the modes, would spread to a ~25-ns pulse after passing a just 1-km fiber segment.) This result should be compared with the chromatic dispersion mentioned above, below 10 ps/km·nm, which gives dt/l is of the order of only 1,000 ps/km in the whole communication band $d\lambda$ ~ 100 nm. Due to this high geometric dispersion, such relatively thick ($2R$ ~ 50 nm) *multi-mode fibers* are used for the transfer of signals over only short distances below ~ 100 m. (As compensation, they may carry relatively large power, beyond 10 mW, without being damaged by the field.)

Long-range telecommunications are based on *single-mode fibers*, with thin cores (typically with diameters $2R$ ~ 5 μm, i. e. of the order of $\lambda/\Delta^{1/2}$). For such structures, Eq. (154) yields $N \sim 1$, but in this case, the geometric optics approximation is not quantitatively valid, and for the fiber analysis, we should get back to the Maxwell equations. In particular, this analysis should take into explicit account the evanescent wave in the cladding, because its penetration depth may be comparable with R.[64]

[64] The following quantitative analysis of the single-mode fibers is very valuable – both for practice and as a very good example of Maxwell equations' solution. However, I have to confess that its results will not be used in the following parts of the course. So, if the reader is not interested in this topic, they may safely jump to the text following Eq. (181). (I believe that the discussion of the angular momentum of electromagnetic radiation, starting at that point, is compulsory for every professional physicist.)

Since the cross-section of an optical fiber lacks metallic walls, the Maxwell equations describing them cannot be exactly satisfied with either TEM-wave, or *H*-mode, or *E*-mode solutions. Instead, the fibers can carry the so-called *HE* and *EH* modes, with both vectors **H** and **E** having longitudinal components simultaneously. In such modes, both E_z and H_z inside the core ($\rho \le R$) have a form similar to Eq. (141):

$$f_- = f_l J_n(k_t \rho)\cos n(\varphi - \varphi_0), \qquad \text{where } k_t^2 = k_-^2 - k_z^2 > 0, \text{ and } k_-^2 \equiv \omega^2 \varepsilon_- \mu_-, \tag{7.156}$$

where the constant angles φ_0 may be different for each field. On the other hand, for the evanescent wave in the cladding, we may rewrite Eqs. (101) as

$$\left(\nabla^2 - \kappa_t^2\right) f_+ = 0, \qquad \text{where } \kappa_t^2 \equiv k_z^2 - k_+^2 > 0, \quad \text{and } k_+^2 \equiv \omega^2 \varepsilon_+ \mu_+ . \tag{7.157}$$

Figure 26 illustrates these relations between k_t, κ_t, k_z, and $k_\pm$; note that the following sum,

Universal relation between k_t and κ_t

$$\boxed{k_t^2 + \kappa_t^2 = \omega^2(\varepsilon_- - \varepsilon_+)\mu_0 = 2k^2\Delta,} \tag{7.158}$$

is fixed (at a given frequency) and, for typical fibers, is very small ($<< k^2$). In particular, Fig. 26 shows that neither k_t nor κ_t can be larger than $\omega[(\varepsilon_- - \varepsilon_+)\mu_0]^{1/2} = (2\Delta)^{1/2} k$. This means that the depth $\delta = 1/\kappa_t$ of the wave penetration into the cladding is at least $1/k(2\Delta)^{1/2} = \lambda/2\pi(2\Delta)^{1/2} >> \lambda/2\pi$. This is why the cladding layers in practical optical fibers are made as thick as ~50 μm, so only a negligibly small tail of this evanescent wave field reaches their outer surfaces.

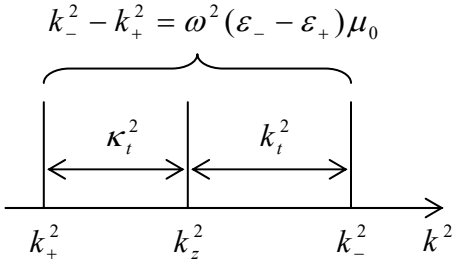


Fig. 7.26. The relation between the transverse exponents k_t and κ_t for waves in optical fibers.

In the polar coordinates, Eq. (157) becomes

$$\left[\frac{1}{\rho}\frac{\partial}{\partial\rho}\left(\rho\frac{\partial}{\partial\rho}\right) + \frac{1}{\rho^2}\frac{\partial^2}{\partial\varphi^2} - \kappa_t^2\right] f_+ = 0, \tag{7.159}$$

- the equation to be compared with Eq. (139) for the circular metallic-wall waveguide. From Sec. 2.7, we know that the eigenfunctions of Eq. (159) are the products of the sine and cosine functions of $n\varphi$ by a linear combination of the modified Bessel functions I_n and K_n, shown in Fig. 2.22, now of the argument $\kappa_t\rho$. The fields have to vanish at $\rho \to \infty$, so only the latter functions (of the second kind) can participate in the solution:

$$f_+ \propto K_n(\kappa_t \rho)\cos n(\varphi - \varphi_0). \tag{7.160}$$

Now we have to reconcile Eqs. (156) and (160), using the boundary conditions at $\rho = R$ for both longitudinal and transverse components of both fields, with the latter components first calculated using Eqs. (121). Such a conceptually simple, but a bit bulky calculation (which I am leaving for the reader's exercise), yields a system of two homogeneous linear equations for the complex amplitudes E_l and H_l, which are compatible if

$$\left(\frac{k_-^2}{k_t}\frac{J_n'}{J_n}+\frac{k_+^2}{\kappa_t}\frac{K_n'}{K_n}\right)\left(\frac{1}{k_t}\frac{J_n'}{J_n}+\frac{1}{\kappa_t}\frac{K_n'}{K_n}\right)=\frac{n^2}{R^2}\left(\frac{k_-^2}{k_t^2}+\frac{k_+^2}{\kappa_t^2}\right)\left(\frac{1}{k_t^2}+\frac{1}{\kappa_t^2}\right), \tag{7.161}$$

where the prime signs denote the derivatives of each Bessel function over its full argument: $k_t\rho$ for J_n, and $\kappa_t\rho$ for K_n, and the functions and their derivatives are taken at $\rho = R$.

For any given frequency ω, the system of equations (158) and (161) determines the values of k_t and κ_t, and hence k_z. Actually, for any $n > 0$, this system provides two different solutions: one corresponding to the so-called *HE* wave, with a larger ratio of E_z/H_z, and the *EH* wave, with a smaller value of that ratio. For angular-symmetric modes with $n = 0$ (for whom we might naively expect the lowest cutoff frequency), the equations may be satisfied by the fields having just one non-zero longitudinal component (either E_z or H_z), so the *HE* wave are the usual *E*-modes, while the *EH* modes are the *H*-waves. For the *H*-modes, the characteristic equation is reduced to the requirement that the expression in the second parentheses on the left-hand side of Eq. (161) is equal to zero. Using the Bessel function identities $J_0' = -J_1$ and $K_0' = -K_1$, this equation may be rewritten in a simpler form:

$$\frac{1}{k_t}\frac{J_1(k_tR)}{J_0(k_tR)}=-\frac{1}{\kappa_t}\frac{K_1(\kappa_tR)}{K_0(\kappa_tR)}. \tag{7.162}$$

Using the universal relation between k_t and κ_t given by Eq. (158), we may plot both sides of Eq. (162) as functions of the same argument, say, $\xi \equiv k_tR$ – see Fig. 27.

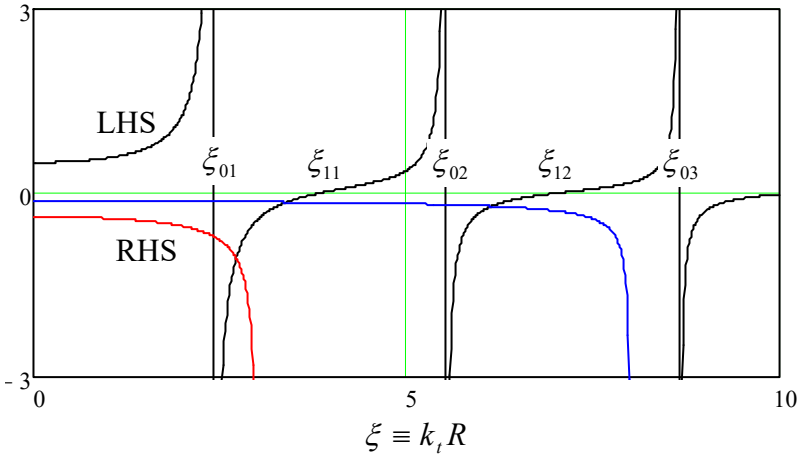


Fig. 7.27. Two sides of the characteristic equation (162), plotted as functions of k_tR, for two values of its dimensionless parameter: $\mathscr{V} = 8$ (blue line) and $\mathscr{V} = 3$ (red line). Note that according to Eq. (158), the argument of the functions K_0 and K_1 is $\kappa_tR = [\mathscr{V}^2 - (k_tR)^2]^{1/2} \equiv (\mathscr{V}^2 - \xi^2)^{1/2}$.

The right-hand side of Eq. (162) depends not only on ξ, but also on the dimensionless parameter $\mathscr{V}$ defined as the normalized right-hand side of Eq. (158):

$$\mathscr{V}^2 \equiv \omega^2(\varepsilon_- - \varepsilon_+)\mu_0R^2 \approx 2\Delta\, k_\pm^2R^2. \tag{7.163}$$

(According to Eq. (154), if $\mathscr{V} >> 1$, it gives twice the number N of the fiber modes – the conclusion confirmed by Fig. 27, taking into account that it describes only the *H*-modes.) Since the ratio K_1/K_0 is positive for all values of the functions' argument (see, e.g., the right panel of Fig. 2.22), the right-hand side of Eq. (162) is always negative, so the equation may have solutions only in the intervals where the ratio J_1/J_0 is negative, i.e. at

$$\xi_{01} < k_tR < \xi_{11}, \quad \xi_{02} < k_tR < \xi_{12}, \ldots, \tag{7.164}$$

where ξ_{nm} is the m-th zero of the function $J_n(\xi)$ – see Table 2.1. The right-hand side of the characteristic equation (162) diverges at $\kappa_t R \to 0$, i.e. at $k_t R \to \mathscr{V}$, so no solutions are possible if $\mathscr{V}$ is below the critical value $\mathscr{V}_c = \xi_{01} \approx 2.405$. At this cutoff point, Eq. (163) yields $k_\pm \approx \xi_{01}/R(2\Delta)^{1/2}$. Hence, the cutoff frequency of the lowest H mode corresponds to the TEM wavelength

$$\lambda_{\max} = \frac{2\pi R}{\xi_{01}}(2\Delta)^{1/2} \approx 3.7 R\Delta^{1/2}. \tag{7.165}$$

For typical parameters $\Delta = 0.005$ and $R = 2.5$ μm, this result yields $\lambda_{\max} \sim 0.65$ μm, corresponding to the free-space wavelength $\lambda_0 \sim 1$ μm. A similar analysis of the first parentheses on the left-hand side of Eq. (161) shows that at $\Delta \to 0$, the cutoff frequency for the E modes is similar.

This situation may look exactly like that in metallic-wall waveguides, with no waves possible at frequencies below ω_c, but this is not so. The basic reason for the difference is that in the metallic waveguides, the approach to ω_c results in the divergence of the longitudinal wavelength $\lambda_z \equiv 2\pi/k_z$. On the other hand, in dielectric waveguides, the approach leaves λ_z finite ($k_z \to k_+$). Due to this difference, a certain linear superposition of HE and EH waves with $n = 1$ can propagate at frequencies well below the cutoff frequency for $n = 0$, which we have just calculated.[65] This mode, in the limit $\varepsilon_+ \approx \varepsilon_-$ (i.e. $\Delta << 1$) allows a very interesting and simple description using the *Cartesian* (rather than polar) components of the fields, but still expressed as functions of the *polar* coordinates ρ and φ. The reason is that this mode is very close to a linearly polarized TEM wave. (Due to this reason, this mode is referred to as LP_{01}.)

Let us select the x-axis parallel to the transverse component of the magnetic field vector at $\rho = 0$, so $E_x|_{\rho=0} = 0$, but $E_y|_{\rho=0} \neq 0$, and $H_x|_{\rho=0} \neq 0$, but $H_y|_{\rho=0} = 0$. The only suitable solutions of the 2D Helmholtz equation (that should be obeyed not only by the z-components of the fields but also their x- and y-components) are proportional to $J_0(k_t\rho)$, with zero coefficients for E_x and H_y:

$$E_x = 0, \quad E_y = E_0 J_0(k_t\rho), \quad H_x = H_0 J_0(k_t\rho), \quad H_y = 0, \quad \text{for } \rho \le R. \tag{7.166}$$

LP_{01} mode

Now we can use the last two equations of Eqs. (100) to calculate the longitudinal components of the fields:

$$E_z = \frac{1}{-ik_z}\frac{\partial E_y}{\partial y} = -i\frac{k_t}{k_z}E_0 J_1(k_t\rho)\sin\varphi, \qquad H_z = \frac{1}{-ik_z}\frac{\partial H_x}{\partial x} = -i\frac{k_t}{k_z}H_0 J_1(k_t\rho)\cos\varphi, \tag{7.167}$$

where I have used the following mathematical identities: $J_0 = -J_1'$, $\partial\rho/\partial x = x/\rho = \cos\varphi$, and $\partial\rho/\partial y = y/\rho = \sin\varphi$. As a sanity check, we see that the longitudinal component or each field is a (legitimate!) eigenfunction of the type (141), with $n = 1$. Note also that if $k_t << k_z$ (this relation is always true if $\Delta << 1$ – see either Eq. (158) or Fig. 26), the longitudinal components of the fields are much smaller than their transverse counterparts, so the wave is indeed very close to the TEM one. Because of that, the ratio of the electric and magnetic field amplitudes is also close to that in the TEM wave: $E_0/H_0 \approx Z_- \approx Z_+$.

Now to satisfy the boundary conditions at the core-to-cladding interface ($\rho = R$), we need to have a similar angular dependence of these components at $\rho \ge R$. The longitudinal components of the fields

[65] This fact becomes less surprising if we recall that in the circular metallic waveguide, discussed in Sec. 6, the fundamental mode (H_{11}, see Fig. 23) also corresponded to $n = 1$ rather than $n = 0$.

are tangential to the interface and thus should be continuous. Using the solutions similar to Eq. (160) with $n = 1$, we get

$$E_z = -i\frac{k_t}{k_z}\frac{J_1(k_tR)}{K_1(\kappa_tR)}E_0K_1(\kappa_t\rho)\sin\varphi, \quad H_z = -i\frac{k_t}{k_z}\frac{J_1(k_tR)}{K_1(\kappa_tR)}H_0K_1(\kappa_t\rho)\cos\varphi, \quad \text{for } \rho \geq R. \quad (7.168)$$

For the transverse components, we should require the continuity of the normal magnetic field μH_n, for our simple field structure equal to just $\mu H_x\cos\varphi$, of the tangential electric field $E_\tau = E_y\sin\varphi$, and of the normal component of $D_n = \varepsilon E_n = \varepsilon E_y\cos\varphi$. Assuming that $\mu_- = \mu_+ = \mu_0$, and $\varepsilon_+ \approx \varepsilon_-$[66] we can satisfy these conditions with the following solutions:

$$E_x = 0, \quad E_y = \frac{J_0(k_tR)}{K_0(\kappa_tR)}E_0K_0(\kappa_t\rho), \quad H_x = \frac{J_0(k_tR)}{K_0(k_tR)}H_0K_0(\kappa_t\rho), \quad H_y = 0, \quad \text{for } \rho \geq R. \quad (7.169)$$

From here, we can calculate components from E_z and H_z, using the same approach as for $\rho \leq R$:

$$E_z = \frac{1}{-ik_z}\frac{\partial E_y}{\partial y} = -i\frac{\kappa_t}{k_z}\frac{J_0(k_tR)}{K_0(\kappa_tR)}E_0K_1(\kappa_t\rho)\sin\varphi,$$
$$H_z = \frac{1}{-ik_z}\frac{\partial H_x}{\partial x} = -i\frac{\kappa_t}{k_z}\frac{J_0(k_tR)}{K_0(\kappa_tR)}H_0K_1(\kappa_t\rho)\cos\varphi, \quad \text{for } \rho \geq R. \quad (7.170)$$

These relations provide the same functional dependence of the fields as Eqs. (167), i.e. the internal and external fields are compatible, but their amplitudes at the interface coincide only if

LP_{01} mode: characteristic equation

$$k_t\frac{J_1(k_tR)}{J_0(k_tR)} = \kappa_t\frac{K_1(\kappa_tR)}{K_0(\kappa_tR)}. \quad (7.171)$$

This characteristic equation (which may be also derived from Eq. (161) with $n = 1$ in the limit $\Delta \to 0$) looks close to Eq. (162), but functionally is much different from it – see Fig. 28. Indeed, its right-hand side is always positive, and the left-hand side tends to zero at $k_tR \to 0$. As a result, Eq. (171) may have a solution for arbitrary small values of the parameter $\mathscr{V}$ defined by Eq. (163), i.e. for *arbitrary low frequencies* (large wavelengths). This is why this mode is used in practical single-mode fibers: there are no other modes with wavelengths larger than the λ_{max} given by Eq. (165), so they cannot be unintentionally excited on small inhomogeneities of the fiber.

It is easy to use the Bessel function approximations by the first terms of the Taylor expansions (2.132) and (2.157) to show that in the limit $\mathscr{V} \to 0$, κ_tR tends to zero much faster than $k_tR \approx \mathscr{V}$: $\kappa_tR \to 2\exp\{-1/\mathscr{V}\} << \mathscr{V}$. This means that the scale $\rho_c \equiv 1/\kappa_t$ of the radial distribution of the LP_{01} wave's fields in the cladding becomes very large. In this limit, this mode may be interpreted as a virtually TEM wave propagating in the cladding, just slightly deformed (and guided) by the fiber's core. The drawback of this feature is that it requires very thick cladding, to avoid energy losses in its outer ("buffer" and "jacket") layers that defend the silica layers from the elements and mechanical damages, but lack their

[66] This is the core assumption of this approximate theory, which accounts only for the most important effect of the small difference of the dielectric constants ε_+ and ε_-: the difference between $(k_+^2 - k_z^2) = k_t^2 > 0$ and $(k_-^2 - k_z^2) = -\kappa_t^2 < 0$. For more discussion of the accuracy of this approximation and some exact results, the interested reader may be referred either to the monograph by A. Snyder and D. Love, *Optical Waveguide Theory*, Chapman and Hill, 1983, or to Chapter 3 and Appendix B in the monograph by Yariv and Yeh, that was cited above.

low optical absorption. Due to this reason, the core radius is usually selected so that the parameter $\mathscr{V}$ is just slightly less than the critical value $\mathscr{V}_c = \xi_{01} \approx 2.4$ for higher modes, thus ensuring the single-mode operation.

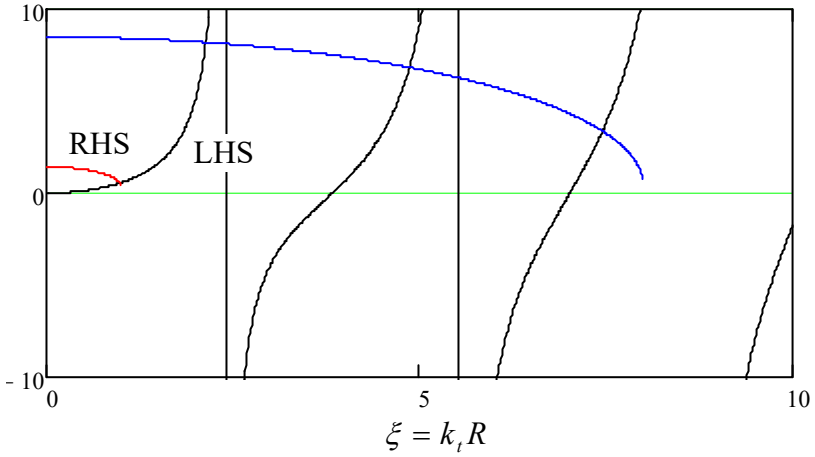


Fig. 7.28. Two sides of the characteristic equation (171) for the LP_{01} mode, plotted as a function of $k_t R$, for two values of the dimensionless parameter: $\mathscr{V} = 8$ (blue line) and $\mathscr{V} = 1$ (red line).

In order to reduce the field spread into the cladding, the step-index fibers discussed above may be replaced with *graded-index* fibers whose dielectric constant ε is gradually and slowly decreased from the center to the periphery.[67] Keeping only the main two terms in the Taylor expansion of the function $\varepsilon(\rho)$ at $\rho = 0$, we may approximate such reduction as[68]

$$\varepsilon(\rho) \approx \varepsilon(0)\left(1 - \zeta\rho^2\right), \tag{7.172}$$

where $\zeta \equiv -[(d^2\varepsilon/d\rho^2)/2\varepsilon]_{\rho=0}$ is a positive constant characterizing the fiber composition gradient. Moreover, if this constant is sufficiently small ($\zeta << k^2$), the field distribution across the fiber's cross-section may be described by the same 2D Helmholtz equation (101), but with a space-dependent transverse wave vector:[69]

$$\left[\nabla_t^2 + k_t^2(\rho)\right] f = 0, \tag{7.173}$$

where

$$k_t^2(\rho) = k^2(\rho) - k_z^2 \equiv k_t^2(0) - k^2(0)\zeta\rho^2, \quad \text{and} \quad k^2(0) \equiv \omega^2\varepsilon(0)\mu_0 .$$

Surprisingly for such an axially-symmetric problem, because of its special dependence on the radius, this equation may be most readily solved in the Cartesian coordinates. Indeed, rewriting it as

$$\left[\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + k_t^2(0) - k^2(0)\zeta\left(x^2 + y^2\right)\right] f = 0 , \tag{7.174}$$

and separating the variables as $f = X(x)Y(y)$, we get

[67] Due to the difficulty of fabrication of graded-index fibers with wave attenuation below a few dm/km, they are not used as broadly as the step-index ones.

[68] For an axially-symmetric smooth function $\varepsilon(\rho)$, the *first* derivative $d\varepsilon/d\rho$ always vanishes at $\rho = 0$, so Eq. (172) does not have a term linear in ρ.

[69] This approach is invalid at arbitrary (large) ζ because in the macroscopic Maxwell equations, $\varepsilon(\mathbf{r})$ is under the differentiation sign, and the exact Helmholtz-type equations for fields have additional terms containing $\nabla\varepsilon$.

$$\frac{1}{X}\frac{d^2X}{dx^2}+\frac{1}{Y}\frac{d^2Y}{dy^2}+k_t^2(0)-k^2(0)\zeta\left(x^2+y^2\right)=0, \tag{7.175}$$

so the functions X and Y obey similar differential equations, for example

$$\frac{d^2X}{dx^2}+\left[k_x^2-k^2(0)\zeta x^2\right]X=0, \tag{7.176}$$

with the separation constants satisfying the following condition:

$$k_x^2+k_y^2=k_t^2(0)\equiv k^2(0)-k_z^2. \tag{7.177}$$

The ordinary differential equation (176) is well known in quantum mechanics, because the stationary Schrödinger equation for one of the most important basic quantum systems, a 1D harmonic oscillator, may be rewritten in this form. Its eigenvalues are very simple:

$$(k_x^2)_n=k(0)\zeta^{1/2}(2n+1),\qquad (k_y^2)_m=k(0)\zeta^{1/2}(2m+1),\qquad \text{with } n,m=0,1,2,..., \tag{7.178}$$

but the corresponding eigenfunctions $X_n(x)$ and $Y_m(y)$ are expressed via not quite elementary functions – the Hermite polynomials.[70] For most practical purposes, however, the lowest eigenfunctions $X_0(x)$ and $Y_0(y)$ are sufficient, because they correspond to the lowest $k_{x,y}$, and hence the lowest

$$\left[k_t^2(0)\right]_{\min}=(k_x^2)_0+(k_y^2)_0=2k(0)\zeta^{1/2}, \tag{7.179}$$

and the lowest cutoff frequency. As may be readily verified by the substitution to Eq. (176), the eigenfunctions corresponding to this fundamental mode are also simple:

$$X_0(x)=\text{const}\times\exp\left\{-\frac{k(0)\zeta^{1/2}x^2}{2}\right\}, \tag{7.180}$$

and similarly for $Y_0(y)$, so the field distribution follows the Gaussian function

$$f_0(\rho)=f_0(0)\exp\left\{-\frac{k(0)\zeta^{1/2}\rho^2}{2}\right\}\equiv f_0(0)\exp\left\{-\frac{\rho^2}{2a^2}\right\},\quad \text{with } a\equiv 1/k^{1/2}(0)\zeta^{1/4}, \tag{7.181}$$

where $a >> 1/k(0)$ has the sense of the effective width of the field's extension in the radial direction, normal to the wave's propagation axis z. This is the so-called *Gaussian beam*, very convenient for some applications.

The Gaussian beam (181) is just one example of the so-called *paraxial beams*, which may be represented as a result of modulation of a plane wave with a wave number k, by an axially-symmetric *envelope function* $f(\rho)$, where $\boldsymbol{\rho}\equiv\{x, y\}$, with a relatively large effective radius $a >> 1/k$.[71] Such beams give me a convenient opportunity to deliver on the promise made in Sec. 1: calculate the angular momentum $\mathbf{L}$ of a circularly polarized wave propagating in free space, and prove its fundamental relation to the wave's energy U. Let us start from the calculation of U for a paraxial beam (with an

[70] See, e.g., QM Sec. 2.9.

[71] Note that propagating in a uniform medium, i.e. outside of grade-index fibers or other focusing systems, such beams gradually increase their width a due to diffraction – the effect to be analyzed in the next chapter.

arbitrary, but spatially localized envelope *f*) of a circularly polarized wave, with the transverse electric field components given by Eq. (19):

$$E_x = E_0 f(\rho)\cos\psi, \qquad E_y = \mp E_0 f(\rho)\sin\psi, \tag{7.182a}$$

where E_0 is the real amplitude of the wave's electric field at the propagation axis, $\psi \equiv kz - \omega t + \varphi$ is its total phase, and the two signs correspond to two possible directions of the circular polarization.[72] According to Eq. (6), the corresponding transverse components of the magnetic field are

$$H_x = \pm\frac{E_0}{Z_0} f(\rho)\sin\psi, \qquad H_y = \frac{E_0}{Z_0} f(\rho)\cos\psi\,. \tag{7.182b}$$

These expressions are sufficient to calculate the energy density (6.113) of the wave,[73]

$$u = \frac{\varepsilon_0\left(E_x^2 + E_y^2\right)}{2} + \frac{\mu_0\left(H_x^2 + H_y^2\right)}{2} = \frac{\varepsilon_0 E_0^2 f^2}{2} + \frac{\mu_0 E_0^2 f^2}{2Z_0^2} \equiv \varepsilon_0 E_0^2 f^2, \tag{7.183}$$

and hence the full energy (per unit length in the direction *z* of the wave's propagation) of the beam:

$$U = \int u d^2 r \equiv 2\pi\int_0^\infty u\rho d\rho = 2\pi\varepsilon_0 E_0^2 \int_0^\infty f^2 \rho d\rho\,. \tag{7.184}$$

However, the transverse fields (182) are insufficient to calculate a non-zero average of **L**. Indeed, following the angular moment's definition in mechanics,[74] $\mathbf{L} \equiv \mathbf{r}\times\mathbf{p}$, where **p** is the particle's (linear) momentum, we may use Eq. (6.115) for the electromagnetic field momentum's density **g** in free space, to define the field's angular momentum's density as

$$\boxed{\mathbf{l} \equiv \mathbf{r}\times\mathbf{g} \equiv \frac{1}{c^2}\mathbf{r}\times\mathbf{S} \equiv \frac{1}{c^2}\mathbf{r}\times(\mathbf{E}\times\mathbf{H}).} \tag{7.185}$$

EM field's angular momentum

Let us use the familiar *bac minus cab* rule of the vector algebra[75] to transform this expression to

$$\mathbf{l} = \frac{1}{c^2}\left[\mathbf{E}(\mathbf{r}\cdot\mathbf{H}) - \mathbf{H}(\mathbf{r}\cdot\mathbf{E})\right] \equiv \frac{1}{c^2}\left\{\mathbf{n}_z\left[E_z(\mathbf{r}\cdot\mathbf{H}) - H_z(\mathbf{r}\cdot\mathbf{E})\right] + \left[\mathbf{E}_t(\mathbf{r}\cdot\mathbf{H}) - \mathbf{H}_t(\mathbf{r}\cdot\mathbf{E})\right]\right\}. \tag{7.186}$$

If the field is purely transverse ($E_z = H_z = 0$), as it is in a strictly plane wave, the first square brackets in the last expression vanish, while the second bracket gives an azimuthal component of **l**, which oscillates in time and vanishes at its time averaging. (This is exactly the reason why I have not tried to calculate **L** during our first discussion of the circularly polarized waves in Sec. 1.)

---

[72] For our task of calculating two *quadratic* forms of the fields (**L** and *U*), their real representation (182) is more convenient than the complex-exponent one. However, for *linear* manipulations, the latter representation of the circularly polarized waves, $\mathbf{E}_t = E_0 f(\rho)\mathrm{Re}[(\mathbf{n}_x \pm i\mathbf{n}_y)\exp\{i\psi\}]$, $\mathbf{H}_t = (E_0/Z_0)f(\rho)\mathrm{Re}[(\mp i\mathbf{n}_x + \mathbf{n}_y)\exp\{i\psi\}]$, is usually more convenient, and is broadly used.

[73] Note that, in contrast to a linearly-polarized wave (16), the energy density of a circularly-polarized wave does not depend on the full phase ψ – in particular, on *t* at fixed *z*, or vice versa. This is natural because its field vectors rotate (keeping their magnitude) rather than oscillate – see Fig. 3b.

[74] See, e.g., CM Eq. (1.31).

[75] See, e.g., MA Eq. (7.5).

Fortunately, our discussion of optical fibers, in particular, the derivation of Eqs. (167), (168), and (170) gives us a clear clue on how to resolve this paradox. If the envelope function $f(\rho)$ differs from a constant, the transverse wave components (182) alone do *not* satisfy the Maxwell equations (2b), which necessitates longitudinal components E_z and H_z of the fields, with[76]

$$\frac{\partial E_z}{\partial z} = -\frac{\partial E_x}{\partial x} - \frac{\partial E_y}{\partial y}, \qquad \frac{\partial H_z}{\partial z} = -\frac{\partial H_x}{\partial x} - \frac{\partial H_y}{\partial y}. \tag{7.187}$$

However, as these expressions show, if the envelope function f changes very slowly in the sense $df/d\rho \sim f/a << kf$, the longitudinal components are very small and do not have a back effect on the transverse components. Hence, the above calculation of U is still valid (asymptotically, at $ka \to 0$), and we may still use Eqs. (182) on the right-hand side of Eqs. (187),

$$\frac{\partial E_z}{\partial z} = E_0\left(-\frac{\partial f}{\partial x}\cos\psi \pm \frac{\partial f}{\partial x}\sin\psi\right), \qquad \frac{\partial H_z}{\partial z} = \frac{E_0}{Z_0}\left(\mp\frac{\partial f}{\partial x}\sin\psi - \frac{\partial f}{\partial x}\cos\psi\right), \tag{7.188}$$

and integrate them over z as

$$\begin{aligned} E_z &= E_0\int\left(-\frac{\partial f}{\partial x}\cos\psi \pm \frac{\partial f}{\partial x}\sin\psi\right)dz = \frac{E_0}{k}\left(-\frac{\partial f}{\partial x}\int\cos\psi\, d\psi \pm \frac{\partial f}{\partial x}\int\sin\psi\, d\psi\right) \\ &\equiv \frac{E_0}{k}\left(-\frac{\partial f}{\partial x}\sin\psi \mp \frac{\partial f}{\partial x}\cos\psi\right). \end{aligned} \tag{7.189a}$$

Here the integration constant is taken for zero because no wave field component may have a time-independent part. Integrating, absolutely similarly, the second of Eqs. (188), we get

$$H_z = \frac{E_0}{kZ_0}\left(\pm\frac{\partial f}{\partial x}\cos\psi - \frac{\partial f}{\partial y}\sin\psi\right). \tag{7.189b}$$

With the same approximation, we may calculate the longitudinal (z–) component of $\mathbf{l}$, given by the first term of Eq. (186), keeping only the dominating, transverse fields (182) in the scalar products:

$$l_z = E_z\left(\mathbf{r}\cdot\mathbf{H}_t\right) - H_z\left(\mathbf{r}\cdot\mathbf{E}_t\right) \equiv E_z\left(xH_x + yH_y\right) - H_z\left(xE_x + yE_y\right). \tag{7.190}$$

Plugging in Eqs. (182) and (189), and taking into account that in free space, $k = \omega/c$, and hence $1/Z_0c^2k = \varepsilon_0/\omega$, we get:

$$l_z = \mp\frac{\varepsilon_0E_0^2}{\omega}\left(xf\frac{\partial f}{\partial x} + y\frac{\partial f}{\partial y}\right) \equiv \mp\frac{\varepsilon_0E_0^2}{2\omega}\left[x\frac{\partial\left(f^2\right)}{\partial x} + y\frac{\partial\left(f^2\right)}{\partial y}\right] \equiv \mp\frac{\varepsilon_0E_0^2}{2\omega}\boldsymbol{\rho}\cdot\boldsymbol{\nabla}\left(f^2\right) \equiv \mp\frac{\varepsilon_0E_0^2}{2\omega}\rho\frac{d\left(f^2\right)}{d\rho}. \tag{7.191}$$

Hence the total angular momentum of the beam (per unit length), is

$$L_z = \int l_z d^2r \equiv 2\pi\int_0^\infty l_z\rho d\rho = \mp\pi\frac{\varepsilon_0E_0^2}{\omega}\int_0^\infty \rho^2\frac{d\left(f^2\right)}{d\rho}d\rho \equiv \mp\pi\frac{\varepsilon_0E_0^2}{\omega}\int_{\rho=0}^{\rho=\infty}\rho^2 d\left(f^2\right). \tag{7.192}$$

Taking this integral by parts, with the assumption that $\rho f \to 0$ at $\rho \to 0$ and $\rho \to \infty$ (as it is true for the Gaussian beam (181) and all realistic paraxial beams), we finally get

[76] The complex-exponential versions of these equalities are given by the bottom line of Eq. (100).

$$L_z = \pm\pi \frac{\varepsilon_0 E_0^2}{\omega} \int_0^\infty f^2 d\left(\rho^2\right) \equiv \pm 2\pi \frac{\varepsilon_0 E_0^2}{\omega} \int_0^\infty f^2 \rho d\rho \,. \tag{7.193}$$

Now comparing this expression with Eq, (184), we see that remarkably, the ratio L_z/U does not depend on the shape and the width of the beam (and of course on the wave's amplitude E_0), so these parameters are very simply and universally related:

$$\boxed{L_z = \pm \frac{U}{\omega}} \,. \tag{7.194}$$

Angular momentum at circular polarization

Since this relation is valid in the plane-wave limit $a \to \infty$, it may be attributed to plane waves as well, with the understanding that in reality, they always have some width ("aperture") restriction.

As the reader certainly knows, in quantum mechanics the energy excitations of any harmonic oscillator of frequency ω are quantized in the units of $\hbar\omega$, while the internal angular momentum of a particle is quantized in the units of $s\hbar$, where s is its spin. In this context, the classical relation (194) is used in quantum electrodynamics as the basis for treating the electromagnetic field excitation quanta (*photons*) as some sort of quantum particles with spin $s = 1$. (Such integer spin also fits the Bose-Einstein statistics of the electromagnetic radiation.)

Unfortunately, I do not have time/space for a further discussion of the (very interesting) physics of paraxial beams but cannot help noticing, at least in passing, the very curious effect of *helical waves* – the beams carrying not only the "spin" momentum (194), but also an additional "orbital" angular momentum. The distribution of their energy in space is not monotonic, as it is in the Gaussian beam (181), but reminds several threads twisted around the propagation axis – hence the term "helical".[77] Mathematically, their field structure is described by the *associate Laguerre polynomials* – the same special functions that are used for the quantum-mechanical description of hydrogen-like atoms.[78] Presently, there are efforts to use such beams for the so-called *orbital angular momentum* (OAM) multiplexing for high-rate information transmission.[79]

7.8. Resonant cavities

Generally, *resonators* are structures that may sustain oscillations (in electrodynamics, of the electromagnetic field) even without an external source, until the oscillation amplitude slowly decreases in time due to unavoidable energy losses. If the resonator quality (described by the so-called *Q-factor*, which will be defined and discussed in the next section) is high, $Q >> 1$, this decay takes many oscillation periods. Alternatively, high-Q resonators may sustain high oscillating fields permanently, if driven by relatively weak incident waves. In contrast to lumped-element resonators, say, the well-known *LC tank circuit*, the subject of this section is *resonant cavities* (or "distributed resonators") limited by either conducting or dielectric walls that contain distributed standing waves inside them.

[77] Noticing such solutions of the Maxwell equations may be traced back to at least a 1943 theoretical work by J. Humblet; however, this issue had not been discussed in literature too much until experiments carried out in 1992 – see, e.g. L. Allen *et al.*, *Optical Angular Momentum;* IOP, 2003.

[78] See, e.g., QM Sec. 3.7.

[79] See, e.g., J. Wang *et al.*, *Nature Photonics* **6**, 488 (2012).

Conceptually the simplest resonant cavity is the *Fabry-Pérot interferometer*[80] that may be obtained by placing two well-conducting planes parallel to each other.[81] Indeed, in Sec. 3 we have seen that if a plane wave is normally incident on such a "perfect mirror", located at $z = 0$, its reflection, at negligible skin depth, results in a standing wave described by Eq. (61b):

$$E(z,t) = \mathrm{Re}\left(2E_\omega e^{-i\omega t + i\pi/2}\right)\sin kz\,. \tag{7.195}$$

This wave would not change if we place the second mirror (isolating the segment of length l from the external wave source) at any position $z = l$ with $\sin kl = 0$, i.e. with

$$kl = p\pi, \quad \text{where } p = 1, 2, ... \tag{7.196}$$

This condition, which determines the spectrum of *own* (or *resonance*, or *eigen-*) *frequencies* of the resonator of fixed length l,

$$\omega_p = vk_p = \frac{\pi v}{a}p, \qquad \text{with } v = \frac{1}{(\varepsilon\mu)^{1/2}}, \tag{7.197}$$

has a simple physical sense: the resonator's length l equals exactly p half-waves of the frequency ω_p. Though this is all very simple, please note a considerable change of philosophy from what we have been doing in the previous sections: the main task of the resonator's analysis is finding its own frequencies ω_p, which are now determined by the system's geometry rather than by an external wave source.

Before we move to cavities of more complex shapes, let us use Eq. (62) to represent the magnetic field in the Fabry-Pérot interferometer:

$$H(z,t) = \mathrm{Re}\left(2\frac{E_\omega}{Z}e^{-i\omega t}\right)\cos kz\,. \tag{7.198}$$

Expressions (195) and (198) show that in contrast to traveling waves, each field of the standing wave changes simultaneously (proportionately) at all points of the Fabry-Pérot resonator, turning to zero everywhere twice a period. At these instants, the energy of the corresponding field vanishes, but the total energy of the two fields stays constant because the counterpart field oscillates with the phase shift $\pi/2$. Such behavior is typical for all electromagnetic resonators.

A more technical remark is that we can readily get the same results (195)-(198) by solving the Maxwell equations from scratch. For example, we already know that in the absence of dispersion, losses, and sources, they are reduced to wave equations (3) for any field components. For the Fabry-Pérot resonator's analysis, we can use the 1D form of these equations, say, for the transverse component of the electric field:

$$\left(\frac{\partial^2}{\partial z^2} - \frac{1}{v^2}\frac{\partial^2}{\partial t^2}\right)E = 0, \tag{7.199}$$

and solve it as a part of an eigenvalue problem with the corresponding boundary conditions. Indeed, by separating time and space variables as $E(z, t) = Z(z)\mathcal{T}(t)$, we obtain

[80] This device, named after its inventors, Charles Fabry and Alfred Pérot; is also called the *Fabry-Pérot etalon* (meaning "gauge"), because of its initial usage for light wavelength measurements.

[81] The resonators formed by well-conducting (usually, metallic) walls are frequently called *resonant cavities*.

$$\frac{1}{Z}\frac{d^2Z}{dz^2} - \frac{1}{v^2}\frac{1}{\mathcal{T}}\frac{d^2\mathcal{T}}{dt^2} = 0. \tag{7.200}$$

Calling the separation constant k^2, we get two similar ordinary differential equations,

$$\frac{d^2Z}{dz^2} + k^2 Z = 0, \qquad \frac{d^2\mathcal{T}}{dt^2} + k^2 v^2 \mathcal{T} = 0, \tag{7.201}$$

both with sinusoidal solutions, so the product $Z(z)\mathcal{T}(t)$ is a standing wave with the wave vector k and frequency $\omega = kv$. (In this form, the equations are valid even in the presence of dispersion, but with a frequency-dependent wave speed: $v^2 = 1/\varepsilon(\omega)\mu(\omega)$.) Now using the boundary conditions $E(0, t) = E(l, t) = 0$,[82] we get the eigenvalue spectrum for k_p and hence for $\omega_p = vk_p$, given by Eqs. (196) and (197).

Lessons from this simple case study may be readily generalized to any cavity formed as a transmission line's section:[83] there are two approaches to finding the resonant frequency spectrum:

(i) We may look at a traveling wave solution and find where reflecting mirrors may be inserted without affecting the wave's structure.

(ii) We may solve the general 3D wave equations,

$$\left(\nabla^2 - \frac{1}{v^2}\frac{\partial^2}{\partial t^2}\right) f(\mathbf{r}, t) = 0, \tag{7.202}$$

for field components, as an eigenvalue problem with appropriate boundary conditions. If the system's parameters (and hence the coefficient v) do not change in time, the spatial and temporal variables of Eq. (202) may be *always* separated by taking

$$f(\mathbf{r}, t) = \sum_k \mathcal{f}_k(\mathbf{r})\mathcal{T}_k(t), \tag{7.203}$$

where each function $\mathcal{T}_k(t)$ *always* obeys the same equation as in Eq. (201), having the sinusoidal solution of frequency $\omega_k = vk$. Plugging this solution back into Eqs. (202), for the spatial distribution of the field, we get the *3D Helmholtz equation*,

$$\boxed{\left(\nabla^2 + k^2\right)\mathcal{f}_k(\mathbf{r}) = 0,} \tag{7.204}$$

3D Helmholtz equation

whose eigenfunctions $\mathcal{f}_k(\mathbf{r})$ may be much more involved, especially for non-symmetric geometries.

Let us use these approaches to find the resonant frequency spectrum of a few simple, but practically important cavities. First of all, the first method is completely sufficient for the analysis of any resonator formed as a fragment of a uniform TEM transmission line (e.g., a coaxial cable), confined with two conducting lids normal to the line's direction. Indeed, since in such lines $k_z = k = \omega/v$, and the electric field is perpendicular to the propagation axis, e.g., parallel to the lid surface, the boundary conditions are exactly the same as in the Fabry-Pérot resonator, and we again arrive at the eigenfrequency spectrum (197).

[82] This is of course the expression of the first of the general boundary conditions (104). The second of these conditions (for the magnetic field) is satisfied automatically for the transverse waves we are considering.

[83] The resonators may have different geometries as well, and in many cases, only the second approach may be used.

Now let us analyze a slightly more complex system: a rectangular metallic-wall cavity of volume $a \times b \times l$ – see Fig. 29. To use the first approach outlined above, let us consider the resonator as a finite-length ($\Delta z = l$) section of the rectangular waveguide extended along the z-axis, which was analyzed in detail in Sec. 6. As a reminder, at $a < b$, in the fundamental H_{10} traveling wave mode, both vectors **E** and **H** do not depend on y, with **E** having only a y-component. In contrast, **H** has two components, H_x and H_z, with the phase shift $\pi/2$ between them, and with H_x having the same phase as E_y – see Eqs. (131), (137), and (138). Hence, if a plane perpendicular to the z-axis, is placed so that the electric field vanishes on it, H_x also vanishes, so both boundary conditions (104), pertinent to a perfect metallic wall, are fulfilled simultaneously.

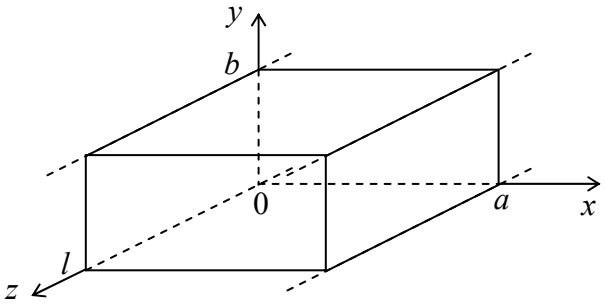


Fig. 7.29. Rectangular metallic-wall resonator as a finite section of a waveguide with the cross-section shown in Fig. 22.

As a result, the H_{10} wave would not be perturbed by two metallic walls separated by an integer number of half-wavelengths $\lambda_z/2$ corresponding to the wave number given by the combination of Eqs. (102) and (133):

$$k_z = \left(k^2 - k_t^2\right)^{1/2} = \left(\frac{\omega^2}{v^2} - \frac{\pi^2}{a^2}\right). \tag{7.205}$$

Using this expression, we see that the smallest of these distances, $l = \lambda_z/2 = \pi/k_z$, gives the resonance frequency[84]

$$\omega_{101} = v\left[\left(\frac{\pi}{a}\right)^2 + \left(\frac{\pi}{l}\right)^2\right]^{1/2}, \tag{7.206}$$

where the indices of ω show the numbers of half-waves along each dimension of the system, in the order [a, b, l]. This is the lowest ("fundamental") frequency of the resonator (if $b < a, l$).

The field distribution in this mode is close to that in the corresponding waveguide mode H_{10} (Fig. 22), with the important difference that the magnetic and electric fields are now shifted by phase $\pi/2$ both in space and time, just as in the Fabry-Pérot resonator – see Eqs. (195) and (198). Such a time shift allows for a very simple interpretation of the H_{101} mode, which is especially adequate for very flat resonators, with $b << a, l$. At the instant when the electric field reaches its maximum (Fig. 30a), i.e. when the magnetic field vanishes in the whole volume, the surface electric charge of the broadest (in Fig. 30, horizontal) walls of the resonator is largest, being localized mostly near the centers of the walls. At the immediate later times, the walls start to recharge via surface currents, whose density J is largest in the side walls, and reaches its maximal value in a quarter of the oscillation period $\mathcal{T} = 2\pi/\omega_{101}$ – see Fig. 30b. The currents generate the vortex magnetic field, with looped field lines in the plane of the

[84] In most electrical engineering handbooks, the index corresponding to the shortest side of the resonator is listed last, so the fundamental mode is nominated as H_{110} and its eigenfrequency as ω_{110}.

broadest face of the resonator. The surface currents continue to flow in this direction until (in one more quarter period) the broader walls of the resonator are fully recharged in the polarity opposite to that shown in Fig. 30a. After that, the surface currents start to flow in the direction opposite to that shown in Fig. 30b. This process, which repeats again and again, is conceptually similar to the well-known oscillations in a lumped *LC* circuit, with the role of (now, distributed) capacitance played mostly by the broadest walls of the resonator, and that of (now, distributed) inductance, by its narrower walls.

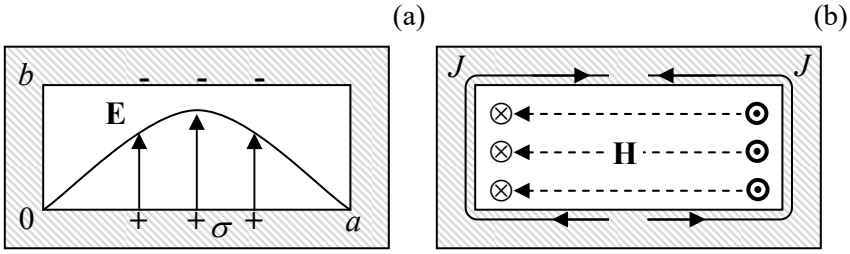


Fig. 7.30. Fields, charges, and currents in the fundamental (H_{101}) mode of a rectangular metallic resonator, at two instants separated by $\Delta t = \pi/2\omega_{101}$ – schematically.

In order to generalize Eq. (206) to higher oscillation modes, the second of the approaches discussed above is more prudent. Separating the variables in the Helmholtz equation (204) as $\mathcal{R}(\mathbf{r}) = X(x)Y(y)Z(z)$, we see that X, Y, and Z have to be either sinusoidal or cosinusoidal functions of their arguments, with the wave vector components satisfying the characteristic equation

$$k_x^2 + k_y^2 + k_z^2 = k^2 \equiv \frac{\omega^2}{v^2}. \tag{7.207}$$

In contrast to the wave propagation problem, now we are dealing with standing waves along all three dimensions, and have to satisfy the macroscopic boundary conditions (104) on all sets of parallel walls. It is straightforward to check that these conditions ($E_\tau = 0$, $H_n = 0$) are fulfilled at the following field component distribution:

$$\begin{aligned} E_x &= E_1 \cos k_x x \sin k_y y \sin k_z z, & H_x &= H_1 \sin k_x x \cos k_y y \cos k_z z, \\ E_y &= E_2 \sin k_x x \cos k_y y \sin k_z z, & H_y &= H_2 \cos k_x x \sin k_y y \cos k_z z, \\ E_z &= E_3 \sin k_x x \sin k_y y \cos k_z z, & H_z &= H_3 \cos k_x x \cos k_y y \sin k_z z, \end{aligned} \tag{7.208}$$

with each of the wave vector components having an equidistant spectrum similar to Eq. (196):

$$k_x = \frac{\pi n}{a}, \quad k_y = \frac{\pi m}{b}, \quad k_z = \frac{\pi p}{l}, \tag{7.209}$$

so the full spectrum of resonance frequencies is given by the following formula:

$$\omega_{nmp} = vk = v\left[\left(\frac{\pi n}{a}\right)^2 + \left(\frac{\pi m}{b}\right)^2 + \left(\frac{\pi p}{l}\right)^2\right]^{1/2}, \tag{7.210}$$

which is a natural generalization of Eq. (206). Note, however, that of the three integers m, n, and p, at least two have to be different from zero to keep the fields (208) from vanishing at all points.

We may use Eq. (210), in particular, to evaluate the number of different modes in a relatively small range $d^3k \ll k^3$ of the wave vector space volume that is, on the other hand, much larger than the reciprocal volume, $1/V = 1/abl$, of the cavity. Taking into account that each eigenfrequency (210), with

$nml \neq 0$, corresponds to two field modes with different polarizations,[85] the argumentation absolutely similar to the one used for the 2D case at the end of Sec. 7, yields

Oscillation mode density

$$dN = 2V\frac{d^3k}{(2\pi)^3}. \tag{7.211}$$

This property, valid for resonators of arbitrary shape, is broadly used in classical and quantum statistical physics,[86] in the following form. If some electromagnetic mode functional $f(\mathbf{k})$ is a smooth function of the wave vector $\mathbf{k}$, and the volume V is large enough, then Eq. (211) may be used to approximate the sum of the functional's values over the modes by an integral:

$$\sum_{\mathbf{k}} f(\mathbf{k}) \approx \int_N f(\mathbf{k})dN \equiv \int_{\mathbf{k}} f(\mathbf{k})\frac{dN}{d^3k}d^3k = 2\frac{V}{(2\pi)^3}\int_{\mathbf{k}} f(\mathbf{k})d^3k\,. \tag{7.212}$$

Leaving similar analyses of resonant cavities of some other simple shapes for the reader's exercises, let me finish this section by noting that low-loss resonators may be also formed by finite-length sections of not only metallic-wall waveguides of various cross-sections but also of dielectric waveguides. Moreover, even a simple slab of a dielectric material with a μ/ε ratio substantially different from that of its environment (say, of the free space) may be used as a high-Q Fabry-Pérot interferometer (Fig. 31), due to an effective wave reflection from its surfaces at the normal and especially an inclined incidence – see, respectively, Eqs. (68), and Eqs. (91) and (95).

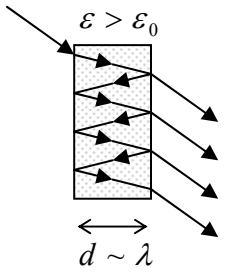


Fig. 7.31. A dielectric Fabry-Pérot interferometer.

Actually, such dielectric Fabry-Pérot interferometers are frequently more convenient for practical purposes than metallic-wall resonators, not only due to possibly lower losses (especially in the optical range) but also due to a natural coupling to the environment, which offers a ready way of wave insertion and extraction – see Fig. 31 again. The backside of the same medal is that this coupling to the environment provides an additional mechanism of power losses, limiting the resonance's quality factor – see the next section.

7.9. Energy loss effects

The inevitable energy losses ("dissipation") in passive media lead, in two basic situations, to two different effects. In a long transmission line fed by a constant wave source, the losses lead to a gradual

[85] This fact becomes evident from plugging Eqs. (208) into the Maxwell equation $\nabla\cdot\mathbf{E} = 0$. The resulting equation, $k_xE_1 + k_yE_2 + k_zE_3 = 0$, with the discrete, equidistant spectrum (209) for each wave vector component, may be always satisfied by two linearly independent sets of the constants $E_{1,2,3}$.

[86] See, e.g., QM Sec. 1.1 and SM Sec. 2.6.

attenuation of the wave, i.e. to a decrease of its amplitude, and hence its power $\mathscr{P}$, with the growing distance z from the source. In linear materials, the power losses are proportional to the power $\mathscr{P}$ carried by the wave, so the energy balance on a small segment dz takes the form

$$d\mathscr{P}_{\text{loss}} \equiv -d\mathscr{P} \equiv -\frac{d\mathscr{P}}{dz}dz = \alpha\mathscr{P}dz\,. \tag{7.213}$$

The coefficient α participating in the last form of Eq. (213), and hence defined as

$$\alpha \equiv -\frac{d\mathscr{P}/dz}{\mathscr{P}}, \tag{7.214}$$

is called the *attenuation constant*.[87] Comparing the solution of Eq. (213),

$$\boxed{\mathscr{P}(z) = \mathscr{P}(0)e^{-\alpha z}\,,} \tag{7.215}$$

Wave's attenuation

with Eq. (29), where k is replaced with k_z, we see that α may be expressed as

$$\alpha = 2\,\text{Im}\,k_z\,, \tag{7.216}$$

where k_z is the component of the wave vector along the transmission line. In the most important limit when the losses are low in the sense $\alpha << \left|\ k_z\ \right| \approx \text{Re}\ k_z$, its effects on the field distribution along the line's cross-section are negligible, making the calculation of α rather straightforward. In particular, in this limit, the contributions to attenuation from two major sources, the energy losses in the filling dielectric and the skin-effect losses in conducting walls, are independent and additive.

The dielectric losses are especially simple to describe. Indeed, a review of our calculations in Secs. 5-7 shows that all of them remain valid if either $\varepsilon(\omega)$, or $\mu(\omega)$, or both, and hence $k(\omega)$, have small imaginary parts:

$$k'' = \omega\,\text{Im}\left[\varepsilon^{1/2}(\omega)\mu^{1/2}(\omega)\right] << k'\,. \tag{7.217}$$

In TEM transmission lines, $k_z = k$, and hence Eq. (216) yields

$$\boxed{\alpha_{\text{filling}} = 2k'' = 2\omega\,\text{Im}\left[\varepsilon^{1/2}(\omega)\mu^{1/2}(\omega)\right].} \tag{7.218}$$

Attenuation due to filling

For dielectric waveguides, in particular optical fibers, these losses are the main attenuation mechanism. As was discussed in Sec. 7, in practical optical fibers $\kappa_t R >> 1$, i.e. most of the field propagates (as an evanescent wave) in the cladding, with a field distribution very close to the TEM wave. This is why Eq. (218) is approximately valid if it is applied to the cladding material alone. In waveguides with non-TEM waves, we can use the relations between k_z and k, derived in the previous sections, to re-calculate k'' into Im k_z. (Note that at this recalculation, the values of k_t have to be kept real, because they are just the eigenvalues of the Helmholtz equation (101), which does not include the filling media parameters.).

---

[87] In engineering, wave attenuation is most frequently measured in *decibels per meter*, abbreviated as db/m (the term not to be confused with dBm standing for decibel-milliwatt):

$$\alpha_{\text{db/m}} \equiv 10\,\log_{10}\frac{\mathscr{P}(z)}{\mathscr{P}(z+1\,\text{m})} = 10\,\log_{10}e^{\alpha[1/\text{m}]} \equiv \frac{10}{\ln 10}\alpha\left[\text{m}^{-1}\right] \approx 4.343\,\alpha\left[\text{m}^{-1}\right].$$

Alternatively, it is sometimes measured in *nepers per meter* (Np/m) defined as $\alpha_{\text{np/m}} \equiv \alpha/2$, so $\alpha_{\text{db/m}} \approx 8.686\ \alpha_{\text{Np/m}}$.

In transmission lines and waveguides and with metallic walls, higher energy losses may come from the skin effect. If the wavelength λ is much larger than δ_s, as it usually is,[88] the losses may be readily evaluated using Eq. (6.36):

$$\frac{d\mathscr{P}_{\rm loss}}{dA} \equiv -\frac{d\mathscr{P}}{dA} = H_{\rm wall}^2 \frac{\mu\omega\delta_s}{4}, \tag{7.219}$$

where $H_{\rm wall}$ is the real amplitude of the tangential component of the magnetic field at the wall's surface. The total power loss $\mathscr{P}_{\rm loss}/dz$ per unit length of a waveguide, i.e. the right-hand side of Eq. (213), now may be calculated by the integration of this expression along the contour(s) limiting the cross-section of all conducting walls. Since our calculation is only valid for low losses, we may ignore their effect on the field distribution, so the unperturbed distributions may be used both in Eq. (219), i.e. in the numerator of Eq. (214), and also for the calculation of the average propagating power, i.e. the denominator of Eq. (214) – as the integral of the Poynting vector over the cross-section of the waveguide.

Let us see how this approach works for the TEM mode in one of the simplest transmission lines, the coaxial cable (Fig. 20). As we already know from Sec. 5, in the coarse-grain approximation, implying negligible power loss, the TEM mode field distributions between the two conductors are the same as in statics, namely:

$$H_z = 0, \qquad H_\rho = 0, \qquad H_\varphi(\rho) = H_0 \frac{a}{\rho}, \tag{7.220}$$

where H_0 is the field's amplitude on the surface of the inner conductor, and

$$E_z = 0, \qquad E_\rho(\rho) = ZH_\varphi(\rho) = ZH_0 \frac{a}{\rho}, \qquad E_\varphi = 0, \qquad \text{where } Z \equiv \left(\frac{\mu}{\varepsilon}\right)^{1/2}. \tag{7.221}$$

Neglecting the power losses for a minute, we may plug these expressions into Eq. (42) to calculate the time-averaged Poynting vector:

$$\overline{S} = \frac{Z\left|H_\varphi(\rho)\right|^2}{2} = \frac{Z\left|H_0\right|^2}{2}\left(\frac{a}{\rho}\right)^2, \tag{7.222}$$

and from it, the total wave power flow through the cross-section:

$$\mathscr{P} = \int_A \overline{S} d^2 r = \frac{Z\left|H_0\right|^2 a^2}{2} 2\pi \int_a^b \frac{\rho d\rho}{\rho^2} = \pi Z \left|H_0\right|^2 a^2 \ln\frac{b}{a}. \tag{7.223}$$

Next, for this particular system (Fig. 20), the contours limiting the wall cross-section are circles of radii $\rho = a$ (where the surface field amplitude $H_{\rm walls}$ equals, in our notation, H_0), and $\rho = b$ (where, according to Eq. (214), the field is a factor of b/a lower). As a result, for the power loss per unit length, Eq. (219) yields

$$\frac{d\mathscr{P}_{\rm loss}}{dz} = \oint_{C_a + C_b} \frac{d\mathscr{P}_{\rm loss}}{dA} dl = \left(2\pi a\left|H_0\right|^2 + 2\pi b\left|H_0\frac{a}{b}\right|^2\right)\frac{\mu_0\omega\delta_s}{4} = \frac{\pi}{2} a\left(1+\frac{a}{b}\right)\mu\omega\delta_s\left|H_0\right|^2. \tag{7.224}$$

[88] As follows from Eq. (78), which may be used for crude estimates even in cases of arbitrary wave incidence, this condition is necessary for low attenuation: $\alpha << k$ only if $\ell << 1$.

Note that at $a << b$, the losses in the inner conductor dominate, despite its smaller surface, because of the higher surface field.

Now we may plug Eqs. (223) and (224) into the definition (214) of α, to calculate the skin-effect contribution to the attenuation constant:

$$\alpha_{\rm skin} \equiv \frac{d\mathscr{P}_{\rm loss}/dz}{\mathscr{P}} = \frac{1}{2\ln(b/a)}\left(\frac{1}{a}+\frac{1}{b}\right)\frac{\mu\omega\delta_{\rm s}}{Z} \equiv \frac{k\delta_{\rm s}}{2\ln(b/a)}\left(\frac{1}{a}+\frac{1}{b}\right). \tag{7.225}$$

This result shows that the relative (dimensionless) attenuation, α/k, scales approximately as the ratio $\delta_{\rm s}/\min[a, b]$, in a semi-quantitative agreement with the plane-wave result (78).

Let us use this result to evaluate α for the standard TV cable RG-6/U, with copper conductors of diameters $2a = 1$ mm, $2b = 4.7$ mm, and $\varepsilon \approx 2.2\varepsilon_0$ and $\mu \approx \mu_0$. According to Eq. (6.33), for $f = 100$ MHz (i.e. $\omega \approx 6.3\times10^8$ s^{-1}) the skin depth of pure copper at room temperature (with $\sigma \approx 6.0\times10^7$ S/m) is close to 6.5×10^{-6} m, while $k = \omega(\varepsilon\mu)^{1/2} = (\varepsilon/\varepsilon_0)^{1/2}(\omega/c) \approx 3.1$ m^{-1}. As a result, the attenuation is rather low: $\alpha_{\rm skin} \approx 0.016$ m^{-1}, so the attenuation length scale $l_{\rm d} \equiv 1/\alpha$ is about 60 m. Hence the attenuation in a cable connecting a roof TV antenna or a cable distribution box to a TV set is not a big problem, though using a worse conductor, e.g., steel, would make the losses rather noticeable. (Hence the current worldwide shortage of copper.) However, the use of such cable in the X-band ($f \sim 10$ GHz) is more problematic. Indeed, though the skin depth $\delta_{\rm s} \propto \omega^{-1/2}$ decreases with frequency, the wavelength drops, i.e. k increases, even faster ($k \propto \omega$), so the attenuation $\alpha_{\rm skin} \propto \omega^{1/2}$ becomes close to 0.16 m^{-1}, i.e. $l_{\rm d}$ to ~6 m. This is why at such frequencies, it may be necessary to use rectangular waveguides, with their larger internal dimensions a, $b \sim 1/k$, and hence lower attenuation. Let me leave the calculation of this attenuation, using Eq. (219) and the results derived in Sec. 7, for the reader's exercise.

The main effect of dissipation on free oscillations in *resonators* is different: here it leads to a gradual decay of the oscillating fields' energy U in time. A useful dimensionless measure of this decay, called the *Q factor*, is commonly defined by writing the following temporal analog of Eq. (213):[89]

$$-dU \equiv \mathscr{P}_{\rm loss}dt = \frac{\omega}{Q}Udt, \tag{7.226}$$

where ω in the resonance frequency in the loss-free limit, and

$$\boxed{\frac{\omega}{Q} \equiv \frac{\mathscr{P}_{\rm loss}}{U}}. \tag{7.227}$$ Q-factor

The solution of Eq. (226),

$$U(t) = U(0)e^{-t/\tau}, \qquad \text{with } \tau \equiv \frac{Q}{\omega} \equiv \frac{Q/2\pi}{\omega/2\pi} = \frac{Q\mathcal{T}}{2\pi}, \tag{7.228}$$

which is the temporal analog of Eq. (215), shows the physical meaning of the Q-factor: the characteristic time τ of the oscillation energy's decay is $(Q/2\pi)$ times longer than the oscillation period $\mathcal{T} = 2\pi/\omega$. (Another useful interpretation of Q comes from the universal relation[90]

[89] As losses grow, the oscillation waveform deviates from the sinusoidal one, and the very notion of "oscillation frequency" becomes vague. As a result, the parameter Q is well-defined only if it is much higher than 1.

[90] See, e.g., CM Sec. 5.1.

$$Q = \frac{\omega}{\Delta\omega}, \tag{7.229}$$

where $\Delta\omega$ is the so-called *FWHM* [91] *bandwidth* of the resonance, namely the difference between the two values of the external signal frequency, one above and one below ω, at which the energy of the oscillations induced in the resonator by an input signal is twice lower than its resonance value.)

In the important particular case of a resonant cavity formed by the insertion of metallic walls into a TEM transmission line of a small cross-section (with the linear size scale a much less than the wavelength λ), there is no need to calculate the Q-factor directly, provided that the line attenuation coefficient α is already known. In fact, as was discussed in Sec. 8 above, the standing waves in such a resonator, of the length given by Eq. (196): $l = p(\lambda/2)$ with $p = 1, 2,\ldots$, may be understood as an overlap of two TEM waves propagating in opposite directions, or in other words, a traveling wave plus its reflection from one of the ends, the whole roundtrip taking time $\Delta t = 2l/v = p\lambda/v = 2\pi p/\omega = p\mathcal{T}$. According to Eq. (215), at this distance, the wave's power drops by a factor of $\exp\{-2\alpha l\} = \exp\{-p\alpha\lambda\}$. On the other hand, the same decay may be viewed as taking place in time, and according to Eq. (228), results in the drop by a factor of $\exp\{-\Delta t/\tau\} = \exp\{-(p\mathcal{T})/(Q/\omega)\} = \exp\{-2\pi p/Q\}$. Comparing these two exponents, we get

Q vs. α

$$Q = \frac{2\pi}{\alpha\lambda} = \frac{k}{\alpha}. \tag{7.230}$$

This simple relation neglects the losses at the wave reflection from the walls limiting the resonator length. This approximation is indeed legitimate at $l >> \lambda$; if this relation is violated, or if we are dealing with more complex resonator modes (such as those based on the reflection of E or H waves), the Q-factor may be different from that given by Eq. (230), and needs to be calculated directly from Eq. (227). A substantial relief for such a direct calculation is that, just at the calculation of small attenuation in waveguides, in the low-loss limit ($Q >> 1$), both the numerator and denominator of the right-hand side of that formula may be calculated neglecting the effects of the energy loss on the field distribution in the resonator. I am leaving such a calculation, for a few simple resonant cavities, including the rectangular and the circular ones, for the reader's exercise.

To conclude this chapter, the last remark: in some distributed resonators (including certain dielectric resonators[92] and metallic cavities with holes in their walls), additional energy losses due to the wave radiation into the environment are also possible. In some simple cases (say, the Fabry-Pérot interferometer shown in Fig. 31), the calculation of these *radiative losses* is straightforward, but sometimes it requires more elaborated approaches that will be discussed in the next chapter.

7.10. Exercise problems

7.1.* Find the temporal Green's function of a medium whose complex permittivity $\varepsilon(\omega)$ obeys the Lorentz oscillator model given by Eq. (32), by using:

(i) the Fourier transform of the underlying Eq. (7.30), and

[91] FWHM is the acronym for *Full Width at Half-Maximum*.

[92] Due to the absence of metallic walls and the associated skin-effect losses, some microwave dielectric resonators (e.g., those based on pure sapphire crystals cooled to helium temperatures) may feature Q-factors as high as 10^9 – see, e.g., D. Creedon *et al.*, *Appl. Phys.* **98**, 222903 (2011).

(ii) the direct solution of that equation.

Hint: For the Fourier-transform approach, you may like to use the Cauchy integral.[93]

7.2. The electric polarization of some material responds to an electric field step[94] in the following way:

$$P(t) = \varepsilon_1 E_0 \left(1 - e^{-t/\tau}\right), \qquad \text{if } E(t) = E_0 \times \begin{cases} 0, & \text{for } t < 0, \\ 1, & \text{for } 0 < t, \end{cases}$$

where $\tau > 0$ and ε_1 are some constants. Calculate the complex permittivity $\varepsilon(\omega)$ of this material, and discuss a possible simple physical model giving such dielectric response.

7.3. Calculate the complex permittivity $\varepsilon(\omega)$ of a material whose dielectric-response Green's function defined by Eq. (23), is

$$G(\theta) = G_0 \left(1 - e^{-\theta/\tau}\right),$$

with some positive constants G_0 and τ. What is the difference between this dielectric response and the apparently similar one considered in the previous problem?

7.4. Use the oscillator model of an atom, given by Eq. (30), to calculate its average potential energy in a uniform, sinusoidal ac electric field, and use the result to calculate the potential profile created for the atom by a standing electromagnetic wave with the electric field amplitude $E_\omega(\mathbf{r})$.

7.5. The solution of the previous problem shows that a standing electromagnetic wave may exert a time-averaged force on an otherwise free non-relativistic charged particle. Reveal the physics of this force by writing and solving the equations of motion of such a particle in:

(i) a linearly-polarized monochromatic plane traveling wave, and
(ii) a similar but standing wave.

7.6. Use the first of Eqs. (54) to relate the integral $\int_0^\infty \varepsilon''(\Omega)\Omega d\Omega$ to the plasma frequency for the Lorentz oscillator model of a system of non-interacting particles.

7.7. Prove that Eq. (6.42) cannot be correct for all frequencies, and suggest its correction making the result compatible with both the causality principle and the physical model (6.39).

7.8. Calculate, sketch, and discuss the dispersion relation for electromagnetic waves propagating in a medium described by the Lorentz oscillator model (32), for the case of negligible damping.

7.9. As was briefly discussed in Sec. 2,[95] a wave pulse of a finite but relatively large spatial extension $\Delta z >> \lambda \equiv 2\pi/k$ may be formed as a *wave packet* – a sum of sinusoidal waves with wave

[93] See, e.g., MA Eq. (15.2).
[94] This function $E(t)$ is of course proportional to the well-known Heaviside step function $\theta(t)$ – see, e.g., MA Eq. (14.3). I am not using this notion here just to avoid confusion between two different uses of the Greek letter θ.
[95] For even more detail, see CM Sec. 5.3 and especially QM Sec. 2.2.

vectors **k** within a relatively narrow interval. Consider an electromagnetic plane wave packet of this type, with the electric field distribution

$$\mathbf{E}(\mathbf{r},t) = \mathrm{Re}\int_{-\infty}^{+\infty} \mathbf{E}_k e^{i(kz-\omega_k t)}dk, \qquad \text{with } k = \omega_k\left[\varepsilon(\omega_k)\mu(\omega_k)\right]^{1/2},$$

propagating along the z-axis in an isotropic, linear, and dissipation-free (but not necessarily dispersion-free) medium. Express the full energy of the packet (per unit area of the wave's front) via the complex amplitudes $\mathbf{E}_k$, and discuss its dependence on time.

7.10. Prove the Lorentz reciprocity relation (6.121) for a linear isotropic medium.

7.11.* A plane wave of frequency ω is normally incident, from free space, on a plane surface of a collision-free plasma with the electron density growing linearly and slowly with the distance from the surface: $n = \gamma z$ for $z \geq 0$, where $\gamma > 0$ is a small constant. Calculate the functional form of the resulting standing wave's "tail" inside the plasma.

7.12.* Analyze the effect of a time-independent uniform magnetic field $\mathbf{B}_0$, parallel to the direction **n** of an electromagnetic wave propagation, on the wave's dispersion in plasma, within the same simple model that was used in Sec. 2 for the derivation of Eq. (38). (Limit your analysis to relatively weak waves, whose magnetic field is much smaller than $\mathbf{B}_0$.)

Hint: You may like to represent the incident wave as a linear superposition of two circularly polarized waves, with opposite polarization directions.

7.13. A monochromatic plane electromagnetic wave is normally incident, from free space, on a uniform slab with electric permittivity ε and magnetic permeability μ, with the slab's thickness d comparable with the wavelength.

(i) Calculate the power transmission coefficient $\mathcal{T}$, i.e. the fraction of the incident wave's power, that is transmitted through the slab.

(ii) Assuming that ε and μ are frequency-independent and positive, analyze in detail the frequency dependence of $\mathcal{T}$. In particular, how does the function $\mathcal{T}(\omega)$ depend on the slab's thickness d and the wave impedance $Z \equiv (\mu/\varepsilon)^{1/2}$ of its material?

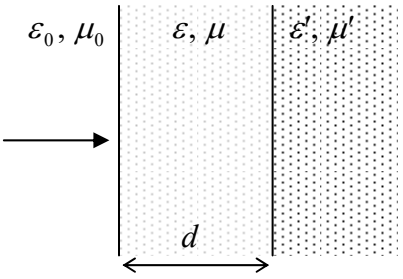
7.14. A plane electromagnetic wave with a free-space wave number k_0 is normally incident on a planar conducting film of thickness $d \sim \delta_s << 1/k_0$. Calculate the power transmission coefficient of the system and analyze the result in the limits of small and large values of the ratio d/δ_s.

7.15. One of the results of the previous problem's solution was the following expression for the coefficient of power transmission of a plane electromagnetic wave through a thin conducting film of thickness $d << \delta_s, \lambda$, at normal incidence:

$$\mathcal{T} = \frac{1}{\left(1 + Z_0/2R_\square\right)^2},$$

where $R_\square \equiv 1/\sigma d$ is the sheet resistance ("resistance per square") of the film. Derive this formula in a simpler way, utilizing the smallness of d from the very beginning. Also, calculate the power reflection coefficient $\mathcal{R}$, compare it with $\mathcal{T}$, and comment.

7.16. A plane wave of frequency ω is normally incident, from free space, on a plane surface of a material with real electric permittivity ε' and magnetic permeability μ'. To minimize the wave's reflection from the surface, it may be covered with a layer, of thickness d, of another transparent material – see the figure on the right. Calculate the optimal values of ε, μ, and d.



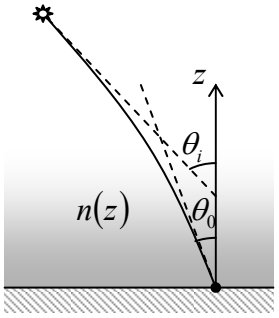
7.17. A monochromatic plane wave is incident from inside a medium with $\varepsilon\mu > \varepsilon_0\mu_0$ on its planar surface, at an incidence angle θ larger than the critical angle $\theta_c = \sin^{-1}(\varepsilon_0\mu_0/\varepsilon\mu)^{1/2}$. Calculate the depth δ of the evanescent wave penetration into the free space, and analyze its dependence on θ. Does the result depend on the wave's polarization?

7.18. Calculate the critical angle θ_c for a wave of frequency ω, incident from free space upon a planar surface of a plasma with electron density n, and discuss the implications of the result for ultraviolet and X-ray optics.

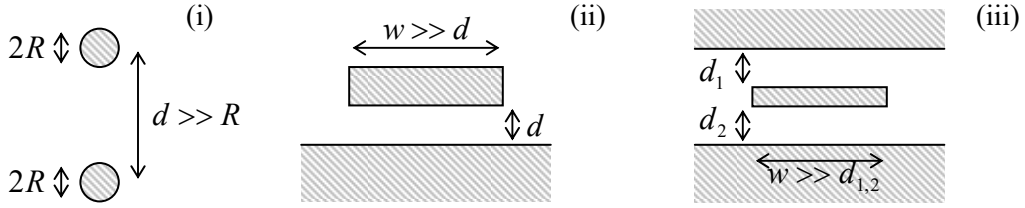
7.19. Analyze the possibility of propagation of surface electromagnetic waves along a planar boundary between plasma and free space. In particular, calculate and analyze the dispersion relation of the waves.

Hint: Assume that the magnetic field of the wave is parallel to the boundary and perpendicular to the wave's propagation direction. (After solving the problem, justify this mode choice.)

7.20. Light from a very distant source arrives to an observer through a plane layer of nonuniform medium with a certain 1D gradient of its refraction index $n(z)$, at angle θ_0 – see the figure on the right. What is the genuine direction θ_i to the source, if $n(z) \to 1$ at $z \to \infty$? (This problem is evidently important for high-precision astronomical measurements at the Earth's surface.)



7.21. Calculate the TEM impedance Z_W of uniform TEM transmission lines with well-conducting electrodes and the cross-sections shown in the figure below:



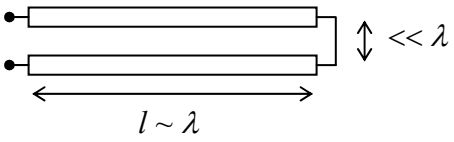
(i) two parallel round wires separated by distance $d >> R$,
(ii) a *microstrip line* of width $w >> d$,
(iii) a *stripline* with $w >> d_1 \sim d_2$,

in all cases using the coarse-grain boundary conditions on conductor surfaces. Assume that the conductors are embedded into a linear dielectric with constant ε and μ.

7.22. Modify the solution of Task (ii) of the previous problem for a superconductor microstrip line, taking into account the magnetic field's penetration into both the strip and the ground plane.

7.23.* What lumped ac circuit would be equivalent to the TEM-line system shown in Fig. 19, with an incident wave's power $\mathscr{P}_i$? Assume that the wave reflected from the lumped load circuit does not return to it.

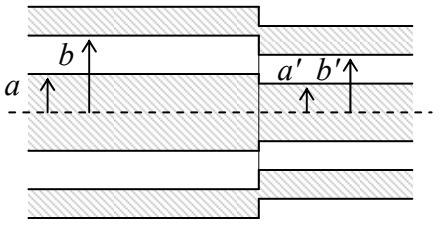
7.24. Find the lumped ac circuit equivalent to a loss-free TEM transmission line of length $l \sim \lambda$, with a small cross-section area $A << \lambda^2$, as "seen" (measured) from one end, if the line's conductors are galvanically connected ("shortened") at the other end – see the figure on the right. Discuss the result's dependence on the signal frequency.



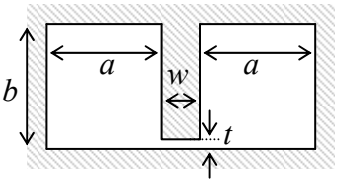
7.25. Represent the fundamental H_{10} wave in a rectangular waveguide (Fig. 22) with a sum of two plane waves, and discuss the physics behind such a representation.

7.26.* For the coaxial cable (see, e.g., Fig. 20), find the lowest non-TEM mode and calculate its cutoff frequency.

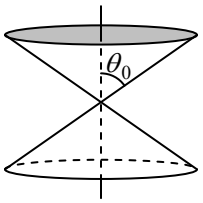
7.27. Two coaxial cable sections are connected coaxially – see the figure on the right, which shows the system's cut along its symmetry axis. Relations (118) and (120) seem to imply that if the ratios b/a of these sections are equal, their impedance matching is perfect, i.e. a TEM wave incident from one side on the connection would pass it without any reflection at all: $R = 0$. Is this statement correct?



7.28. Calculate the cutoff frequencies ω_c of the fundamental mode and the next lowest mode in the so-called *ridge waveguide* with the cross-section shown in the figure on the right, in the limit $t << a, b, w$. Briefly discuss possible advantages and drawbacks of such waveguides for signal transfer and physical experiment.



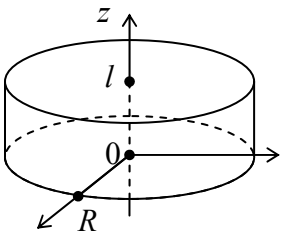
7.29. Prove that TEM-like waves may propagate, in the radial direction, in the free space between two coaxial, round, well-conducting cones – see the figure on the right. Can this system be characterized by a certain transmission line impedance Z_W, as defined by Eq. (115)?



7.30. Use the recipe outlined in Sec. 7 to prove the characteristic equation (161) for the *HE* and *EH* waves in step-index optical fibers with a round cross-section.

7.31. Derive an approximate equation describing spatial variations of the complex amplitude of a general monochromatic paraxial beam propagating in a uniform medium, for the case when these variations are sufficiently slow. Is the Gaussian beam described by Eq. (181) one of the possible solutions of this equation? Give your interpretation of the last result.

7.32. Calculate the lowest resonance frequencies and the corresponding field distributions of standing electromagnetic waves inside a round cylindrical cavity with well-conducting walls (see the figure on the right), neglecting the skin depth δ_s in comparison with l and R.



7.33. Analyze electromagnetic waves that may propagate inside a relatively narrow gap between two well-conducting concentric spherical shells of radii R and $R + d$, in the limit $d << R$.

(i) Within the coarse-grain approximation, derive the 2D equation describing such waves with relatively large wavelengths $\lambda \sim R >> d$.

(ii) Calculate the lowest resonance frequencies of the system.

7.34. A molecule with an electric polarizability α is placed inside an otherwise empty macroscopic cavity with well-conducting walls. Express the resulting shifts of its resonance frequencies via the unperturbed field distribution in the corresponding mode.

7.35. A plane monochromatic wave propagates through a medium with an Ohmic conductivity σ and negligible electric and magnetic polarization effects. Calculate the wave's attenuation and relate the result to a certain calculation carried out in Chapter 6.

7.36. Generalize the telegrapher's equations (110)-(111) by accounting for small energy losses:

(i) in the transmission line's conductors, and

(ii) in the medium separating the conductors,

using their simplest (Ohmic) models. Formulate the conditions of validity of the resulting equations.

7.37. Calculate the skin-effect contribution to the attenuation constant α of a TEM wave in the microstrip line discussed in Problem 21 (ii).

7.38. Calculate the skin-effect contribution to the attenuation coefficient α defined by Eq. (214), for the fundamental (H_{10}) mode propagating in a waveguide with well-conducting walls, of a rectangular cross-section – see Fig.22. Use the results to evaluate the wave decay length $l_d \equiv 1/\alpha$ of a 10 GHz wave in the standard X-band waveguide WR-90 (with copper walls, $a = 23$ mm, $b = 10$ mm, and no dielectric filling), at room temperature. Compare the result with that (obtained in Sec. 9) for the standard TV coaxial cable, at the same frequency.

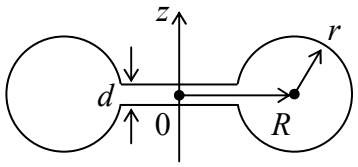
7.39. Calculate the skin-effect contribution to the attenuation coefficient α of

(i) the fundamental (H_{11}) wave, and
(ii) the H_{01} wave,

in a conductor-wall waveguide with the circular cross-section (see Fig. 23a), and analyze the low-frequency ($\omega \to \omega_c$) and high-frequency ($\omega >> \omega_c$) behaviors of α for each of these modes.

7.40. For a rectangular cavity of dimensions $a \times b \times l$, with $b \le a, l$, calculate the Q-factor of the fundamental oscillation mode, due to the skin-effect losses in its conducting walls. Evaluate the factor for a $23 \times 23 \times 10$ mm^3 cavity with copper walls, at room temperature.

7.41.* Calculate the lowest eigenfrequency and the Q-factor (due to the skin-effect losses) of the axially symmetric toroidal cavity with well-conducting walls and the interior's cross-section shown in the figure on the right, for the case $d << r, R$.[96]



7.42. Express the contribution to the damping coefficient (the reciprocal Q-factor) of a resonant cavity, by small energy losses in the dielectric that fills it, via the complex functions $\varepsilon(\omega)$ and $\mu(\omega)$ of the material.

7.43. For the dielectric Fabry-Pérot resonator (Fig. 31) with the normal wave incidence, calculate the Q-factor due to radiation losses, in the limit of a strong impedance mismatch ($Z >> Z_0$), by using two approaches:

(i) from the energy balance, using Eq. (227), and
(ii) from the frequency dependence of the power transmission coefficient, using Eq. (229).

Compare the results.

[96] Such resonators are broadly used in particle accelerators and also in vacuum electron devices for high-power microwave amplification and generation (e.g., the so-called *klystrons*), where the electric field has to be concentrated in the region of charged particle passage – typically, along the symmetry axis (in the figure above, the z-axis), through a pair of small holes in the cavity's walls, which do not affect the field distribution substantially.

This page is intentionally left blank

Chapter 8. Radiation, Scattering, Interference, and Diffraction

This chapter continues the discussion of electromagnetic wave propagation, now focusing on the results of wave incidence on various objects of more complex shapes. Depending on the shape, the resulting wave pattern is called either "scattering", or "diffraction", or "interference". However, as the reader will see, the boundaries between these effects may be blurry, and their basic mathematical description may be conveniently based on the same key calculation – the electric-dipole radiation of a spherical wave by a localized source. Naturally, I will start the chapter from this calculation, deriving it from an even more general result – the "retarded-potential" solution of the Maxwell equations.

8.1. Retarded potentials

Let us start by finding the general solution of the macroscopic Maxwell equations (6.99) in a dispersion-free, linear, uniform, isotropic medium characterized by frequency-independent real ε and μ.[1] The easiest way to perform this calculation is to use the scalar (ϕ) and vector (**A**) potentials defined by Eqs. (6.7):

$$\mathbf{E} = -\nabla\phi - \frac{\partial \mathbf{A}}{\partial t}, \qquad \mathbf{B} = \nabla \times \mathbf{A}. \tag{8.1}$$

As was discussed in Sec. 6.8, by imposing upon the potentials the Lorenz gauge condition (6.117),

$$\nabla \cdot \mathbf{A} + \frac{1}{v^2}\frac{\partial \phi}{\partial t} = 0, \qquad \text{with } v^2 \equiv \frac{1}{\varepsilon\mu}, \tag{8.2}$$

which does not affect the fields **E** and **B**, the Maxwell equations are reduced to a pair of very similar, simple equations (6.118) for the potentials:

$$\nabla^2\phi - \frac{1}{v^2}\frac{\partial^2\phi}{\partial t^2} = -\frac{\rho}{\varepsilon}, \tag{8.3a}$$

$$\nabla^2\mathbf{A} - \frac{1}{v^2}\frac{\partial^2\mathbf{A}}{\partial t^2} = -\mu\,\mathbf{j}. \tag{8.3b}$$

Let us find the general solution of these equations, for now thinking of the densities $\rho(\mathbf{r}, t)$ and $\mathbf{j}(\mathbf{r}, t)$ of the stand-alone charges and currents as known functions. (This will not prevent the results from being valid for the cases when $\rho(\mathbf{r}, t)$ and $\mathbf{j}(\mathbf{r}, t)$ should be calculated self-consistently.) The idea of such a solution may be borrowed from electro- and magnetostatics. Indeed, for the stationary case ($\partial/\partial t = 0$), the solutions of Eqs. (8.3) are given by a ready generalization of, respectively, Eqs. (1.38) and (5.28) to a uniform, linear medium:

$$\phi(\mathbf{r}) = \frac{1}{4\pi\varepsilon}\int \rho(\mathbf{r}')\frac{d^3r'}{|\mathbf{r}-\mathbf{r}'|}, \tag{8.4a}$$

$$\mathbf{A}(\mathbf{r}) \equiv \frac{\mu}{4\pi}\int \mathbf{j}(\mathbf{r}')\frac{d^3r'}{|\mathbf{r}-\mathbf{r}'|}. \tag{8.4b}$$

[1] When necessary (e.g., at the discussion of the Cherenkov radiation in Sec. 10.5), it will be not too hard to generalize these results to a dispersive medium.

As we know, these expressions may be derived by, first, calculating the potential of a point source, and then using the linear superposition principle for a system of such sources.

Let us do the same for the time-dependent case, starting from the field induced by a time-dependent point charge at the origin:[2]

$$\rho(\mathbf{r},t) = q(t)\delta(\mathbf{r}), \tag{8.5}$$

In this case, Eq. (3a) is homogeneous everywhere but the origin:

$$\nabla^2\phi - \frac{1}{v^2}\frac{\partial^2\phi}{\partial t^2} = 0, \quad \text{for } r \neq 0. \tag{8.6}$$

Due to the spherical symmetry of the problem, it is natural to look for a spherically symmetric solution to this equation.[3] Thus, we may simplify the Laplace operator correspondingly (as was repeatedly done earlier in this course), so Eq. (6) becomes

$$\left[\frac{1}{r^2}\frac{\partial}{\partial r}\left(r^2\frac{\partial}{\partial r}\right) - \frac{1}{v^2}\frac{\partial^2}{\partial t^2}\right]\phi = 0, \quad \text{for } r \neq 0. \tag{8.7}$$

By introducing a new variable $\chi(r, t) \equiv r\phi(r, t)$, Eq. (7) is reduced to the 1D wave equation

$$\left(\frac{\partial^2}{\partial r^2} - \frac{1}{v^2}\frac{\partial^2}{\partial t^2}\right)\chi = 0, \quad \text{for } r \neq 0. \tag{8.8}$$

From discussions in Chapter 7,[4] we know that its general solution may be represented as

$$\chi(r,t) = \chi_{\text{out}}\left(t - \frac{r}{v}\right) + \chi_{\text{in}}\left(t + \frac{r}{v}\right), \tag{8.9}$$

where χ_{in} and χ_{out} are (so far) arbitrary functions of one variable. The physical sense of $\phi_{\text{out}} = \chi_{\text{out}}/r$ is a spherical wave propagating from our source (located at $r = 0$) to outer space, i.e. exactly the solution we are looking for. On the other hand, $\phi_{\text{in}} = \chi_{\text{in}}/r$ describes a spherical wave that could be created by some distant spherically-symmetric source, that converged exactly on our charge located at the origin – evidently not the effect we want to consider here. Discarding this term, and returning to $\phi = \chi/r$, we get

$$\phi(r,t) = \frac{1}{r}\chi_{\text{out}}\left(t - \frac{r}{v}\right), \quad \text{for } r \neq 0. \tag{8.10}$$

In order to calculate the function χ_{out}, let us consider the solution (10) at distances r so small ($r << vt$) that the time-derivative term in Eq. (3a), with the right-hand side (5),

$$\nabla^2\phi - \frac{1}{v^2}\frac{\partial^2\phi}{\partial t^2} = -\frac{q(t)}{\varepsilon}\delta(\mathbf{r}), \tag{8.11}$$

[2] Admittedly, this expression does *not* satisfy the continuity equation (4.5), but this deficiency will be corrected imminently, at the linear superposition stage – see Eq. (17) below.

[3] Let me emphasize that this is *not* the general solution to Eq. (6). For example, it does not describe the possible waves created by other sources, that pass by the considered charge $q(t)$. However, such fields are irrelevant to our current task: to calculate the field *induced* by the charge $q(t)$. The solution becomes general when it is integrated (as it will be) over all relevant charges.

[4] See also CM Sec. 6.3.

is much smaller than the spatial derivative term (which diverges at $r \to 0$) . Then Eq. (11) is reduced to the Poisson equation, whose solution (4a), for the source (5), is

$$\phi(r \to 0, t) = \frac{q(t)}{4\pi\varepsilon r}. \tag{8.12}$$

Now requiring the two solutions, Eqs. (10) and (12), to coincide at $r << vt$, we get $\chi_{out}(t) = q(t)/4\pi\varepsilon r$, so Eq. (10) becomes

$$\phi(r,t) = \frac{1}{4\pi\varepsilon r} q\left(t - \frac{r}{v}\right). \tag{8.13}$$

Just as was repeatedly done in statics, this result may be readily generalized for the arbitrary position $\mathbf{r}'$ of the point charge:

$$\rho(\mathbf{r},t) = q(t)\delta(\mathbf{r} - \mathbf{r}') \equiv q(t)\delta(\mathbf{R}), \tag{8.14}$$

where R is the distance between the field observation point $\mathbf{r}$ and the source position point $\mathbf{r}'$, i.e. the length of the vector,

$$\mathbf{R} \equiv \mathbf{r} - \mathbf{r}', \tag{8.15}$$

connecting these points – see Fig. 1.

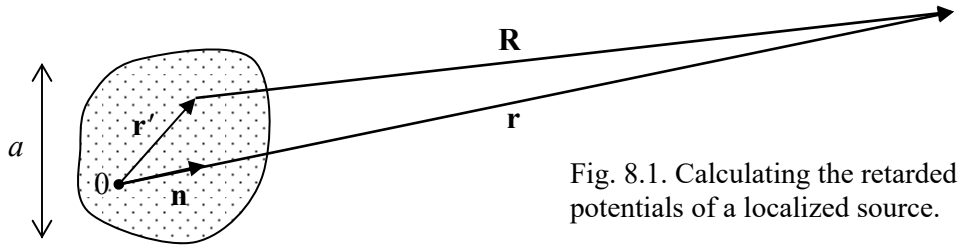


Fig. 8.1. Calculating the retarded potentials of a localized source.

Obviously, now Eq. (13) becomes

$$\phi(\mathbf{r},t) = \frac{1}{4\pi\varepsilon R} q\left(t - \frac{R}{v}\right). \tag{8.16}$$

Finally, we may use the linear superposition principle to write, for the arbitrary charge distribution,

Retarded scalar potential

$$\phi(\mathbf{r},t) = \frac{1}{4\pi\varepsilon}\int \rho\left(\mathbf{r}', t - \frac{R}{v}\right)\frac{d^3r'}{R}, \tag{8.17a}$$

where the integration is extended over all charges of the system under analysis. Solving Eq. (4b) absolutely similarly, for the vector potential we get[5]

Retarded vector potential

$$\mathbf{A}(\mathbf{r},t) = \frac{\mu}{4\pi}\int \mathbf{j}\left(\mathbf{r}', t - \frac{R}{v}\right)\frac{d^3r'}{R}. \tag{8.17b}$$

[5] As should be clear from the analogy of Eqs. (17) with their stationary forms (4), which were discussed, respectively, in Chapters 1 and 5, in the Gaussian units the retarded potential formulas are valid with the coefficient $1/4\pi$ dropped in Eq. (17a), and replaced with $1/c$ in Eq. (17b).

(Now nothing prevents the functions $\rho(\mathbf{r}, t)$ and $\boldsymbol{j}(\mathbf{r}, t)$ from satisfying the continuity equation.)

The solutions expressed by Eqs. (17) are traditionally called the *retarded potentials*, the name signifying the fact that the observed fields are "retarded" (in the meaning "delayed") in time by $\Delta t = R/v$ relative to the source variations – physically, because of the finite speed v of the electromagnetic wave propagation. Note that, very remarkably, these simple expressions are *exact* solutions of the macroscopic Maxwell equations (again, in a uniform, linear, dispersion-free medium) for an *arbitrary* distribution of stand-alone charges and currents. They also may be considered as the *general* solutions of these equations, provided that the integration has been extended over all field sources in the Universe – or at least over those ones that affect our observations.

Note also that due to the mathematical similarity of the microscopic and macroscopic Maxwell equations, Eqs. (17) are valid, with the coefficient replacement $\varepsilon \to \varepsilon_0$ and $\mu \to \mu_0$, for the exact, rather than the macroscopic fields, provided that the functions $\rho(\mathbf{r}, t)$ and $\mathbf{j}(\mathbf{r}, t)$ describe not only stand-alone but all charges and currents in the system. (Alternatively, this statement may be formulated as the validity of Eqs. (17), with the same coefficient replacement, in free space.)

Finally, note that Eqs. (17) may be plugged into Eqs. (1), giving (after an explicit differentiation) the so-called *Jefimenko equations*[6] for fields $\mathbf{E}$ and $\mathbf{B}$ – similar in structure to Eqs. (17), but more cumbersome. Conceptually, the existence of such equations is good news, because they are free from the gauge ambiguity pertinent to the potentials ϕ and $\mathbf{A}$. However, the practical value of these explicit expressions for the fields is not overly high: for all applications I am aware of, it is easier to use Eqs. (17) to calculate the particular expressions for the potentials first, and only then calculate the fields from Eqs. (1). Let me now present an (arguably, the most important) example of this approach.

8.2. Electric dipole radiation

Consider again the problem that was discussed in electrostatics (Sec. 3.1), namely the field of a localized source with linear dimensions $a << r$ (see Fig. 1 again), but now with time-dependent charge and/or current distributions. Using all the arguments of that discussion, in particular the condition expressed by Eq. (3.1), $r' << r$, we may apply the Taylor expansion (3.3), truncated to two leading terms,

$$f(\mathbf{R}) = f(\mathbf{r}) - \mathbf{r}' \cdot \nabla f(\mathbf{r}) + \dots , \tag{8.18}$$

to the scalar function $f(\mathbf{R}) \equiv R$ (for which $\nabla f(\mathbf{r}) = \nabla R = \mathbf{n}$, where $\mathbf{n} \equiv \mathbf{r}/r$ is the unit vector directed toward the observation point – see Fig. 1) to approximate the distance R as

$$R \approx r - \mathbf{r}' \cdot \mathbf{n} . \tag{8.19}$$

In each of the retarded potential formulas (17), R participates in two places: in the denominator and in the source's time argument. If ρ and $\mathbf{j}$ change in time on the scale $\sim 1/\omega$, where ω is some characteristic frequency, then any change of the argument $(t - R/v)$ on that time scale, for example due to a change of R on the spatial scale $\sim v/\omega = 1/k$, may substantially change these functions. Thus, the expansion (19) may be applied to R in the argument $(t - R/v)$ only if $ka << 1$, i.e. if the system's size a

[6] They were published by O. D. Jefimenko only in 1966, but the Fourier representation of the same result was obtained much earlier (in 1912) by G. A. Scott.

is much smaller than the radiation wavelength $\lambda = 2\pi/k$. On the other hand, the function $1/R$ changes relatively slowly, and for it even the first term of the expansion (19) gives a good approximation as soon as $a << r, R$. In the latter approximation alone, Eq. (17a) yields

$$\phi(\mathbf{r},t) \approx \frac{1}{4\pi\varepsilon r}\int \rho\left(\mathbf{r}', t-\frac{R}{v}\right)d^3r' \equiv \frac{1}{4\pi\varepsilon r}Q\left(t-\frac{R}{v}\right), \tag{8.20}$$

where $Q(t)$ is the net electric charge of the localized system. Due to the charge conservation, this charge cannot change with time, so the approximation (20) describes just a static Coulomb field of our localized source, rather than a radiated wave.

Let us, however, apply a similar approximation to the vector potential (17b):

$$\mathbf{A}(\mathbf{r},t) \approx \frac{\mu}{4\pi r}\int \mathbf{j}\left(\mathbf{r}', t-\frac{R}{v}\right)d^3r'. \tag{8.21}$$

According to Eq. (5.87), the right-hand side of this expression vanishes in statics, but in dynamics, this is no longer true. For example, if the current is due to some non-relativistic motion[7] of a system of point charges q_k, we can write

$$\int \mathbf{j}(\mathbf{r}',t)d^3r' = \sum_k q_k \dot{\mathbf{r}}_k(t) = \frac{d}{dt}\sum_k q_k \mathbf{r}_k(t) \equiv \dot{\mathbf{p}}(t), \tag{8.22}$$

where $\mathbf{p}(t)$ is the dipole moment of the localized system, defined by Eq. (3.6). Now, after the integration, we may keep only the first term of the approximation (19) in the argument $(t - R/v)$ as well, getting

$$\mathbf{A}(\mathbf{r},t) \approx \frac{\mu}{4\pi r}\dot{\mathbf{p}}\left(t-\frac{r}{v}\right), \qquad \text{for } a << R, \frac{1}{k}. \tag{8.23}$$

Let us analyze what exactly this result describes. The second of Eqs. (1) allows us to calculate the magnetic field by the spatial differentiation of $\mathbf{A}$. At large distances $r >> \lambda$ (i.e. in the so-called *far-field zone*), where Eq. (23) describes a locally-plane wave, the dominating contribution to this derivative is given by the dipole moment factor:

Far-field wave

$$\mathbf{B}(\mathbf{r},t) = \frac{\mu}{4\pi r}\nabla\times\dot{\mathbf{p}}\left(t-\frac{r}{v}\right) = -\frac{\mu}{4\pi rv}\mathbf{n}\times\ddot{\mathbf{p}}\left(t-\frac{r}{v}\right). \tag{8.24}$$

This expression means that the magnetic field, at the observation point, is perpendicular to the vectors $\mathbf{n}$ and (the retarded value of) $\ddot{\mathbf{p}}$, and its magnitude is

$$B = \frac{\mu}{4\pi rv}\ddot{p}\left(t-\frac{r}{v}\right)\sin\Theta, \qquad \text{i.e. } H = \frac{1}{4\pi rv}\ddot{p}\left(t-\frac{r}{v}\right)\sin\Theta, \tag{8.25}$$

where Θ is the angle between those two vectors – see Fig. 2.[8]

[7] For relativistic particles, moving with velocities of the order of speed of light, one has to be more careful. As the result, I will postpone the discussion of their radiation until Chapter 10, i.e. until after the detailed discussion of special relativity in Chapter 9.

[8] From the first of Eqs. (1) for the electric field, in the first approximation (23), we would get $-\partial\mathbf{A}/\partial t = -(1/4\pi\varepsilon vr)\ddot{\mathbf{p}}(t - r/v) = -(Z/4\pi r)\ddot{\mathbf{p}}(t - r/v)$. The transverse component of this vector (see Fig. 2) is the proper electric field $\mathbf{E}$

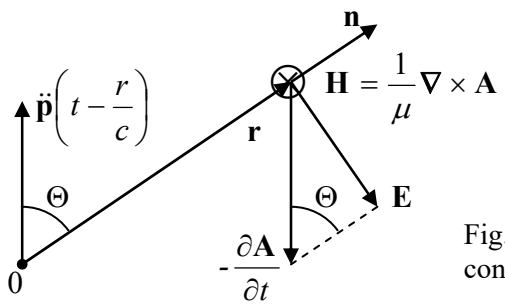


Fig. 8.2. Far-fields of a localized source, contributing to its electric dipole radiation.

The most important feature of this result is that the time-dependent field decreases very slowly (only as $1/r$) with the distance from the source, so the radial component of the corresponding Poynting vector (7.9b),[9]

$$S_r = ZH^2 = \frac{Z}{(4\pi vr)^2}\left[\ddot{p}\left(t-\frac{r}{v}\right)\right]^2 \sin^2\Theta\,, \tag{8.26}$$

Instant power density

drops only as $1/r^2$, i.e. the full instant power $\mathscr{P}$ of the emitted wave,

$$\mathscr{P} \equiv \oint_{4\pi} S_r r^2 d\Omega = \frac{Z}{(4\pi v)^2}\ddot{p}^2 2\pi\int_0^{\pi}\sin^3\Theta\, d\Theta = \frac{Z}{6\pi v^2}\ddot{p}^2\,. \tag{8.27}$$

Larmor formula

does not depend on the distance from the source – as it should for radiation. [10]

This is the famous *Larmor formula*[11] for the *electric dipole radiation*; it is the dominating component of radiation by a localized system of charges – unless $\ddot{\mathbf{p}} = 0$. Please notice its angular dependence: the radiation vanishes at the axis of the retarded vector $\ddot{\mathbf{p}}$ (where $\Theta = 0$), and reaches its maximum in the plane normal to that axis.

In order to find the average power, Eq. (27) has to be averaged over a sufficiently long time. In particular, if the source is monochromatic, $\mathbf{p}(t) = \mathrm{Re}[\mathbf{p}_\omega \exp\{-i\omega t\}]$, with a time-independent vector amplitude $\mathbf{p}_\omega$, such averaging may be carried out just over one period, giving an extra factor 2 in the denominator:

$$\overline{\mathscr{P}} = \frac{Z\omega^4}{12\pi v^2}|p_\omega|^2. \tag{8.28}$$

Average radiation power

The easiest application of this formula is to a point charge oscillating, with frequency ω, along a straight line (which we may take for the z-axis), with amplitude a. In this case, $\mathbf{p} = qz(t)\mathbf{n}_z = qa\,\mathrm{Re}[\exp\{-i\omega t\}]\mathbf{n}_z$, and if the charge velocity amplitude, $a\omega$, is much less than the electromagnetic wave's speed v, we may use Eq. (28) with $p_\omega = qa$, giving

---

$= Z\mathbf{H}\times\mathbf{n}$ of the radiated wave, while its longitudinal component is exactly compensated by $(-\nabla\phi)$ in the *next* term of the Taylor expansion of Eq. (17a) in small parameter $ka \sim a/\lambda << 1$.

[9] Note the "doughnut" dependence of S_r on the direction $\mathbf{n}$, frequently used to visualize the dipole radiation.

[10] In the Gaussian units, for free space ($v = c$), Eq. (27) reads $\mathscr{P} = (2/3c^3)\ddot{p}^2$.

[11] Named after Joseph Larmor, who was the first to derive this formula (in 1897) for the particular case of a single point charge q moving with acceleration $\ddot{\mathbf{r}}$, when $\ddot{\mathbf{p}} = q\ddot{\mathbf{r}}$.

$$\overline{\mathscr{P}} = \frac{Zq^2a^2\omega^4}{12\pi v^2}. \tag{8.29}$$

Applied to a classical picture of an electron (with $q = -e \approx 1.6\times10^{-19}$C), initially rotating about an atom's nucleus at an atomic distance $a \sim 10^{-10}$ m, Eq. (29) shows[12] that the energy loss due to the dipole radiation is so large that it would cause the electron to collapse on the nucleus in just $\sim10^{-11}$ s. In the beginning of the 1900s, this result was one of the main arguments for the development of quantum mechanics, which prevents such a collapse of electrons for their lowest-energy (ground) quantum state.

Another useful application of Eq. (28) is the radio wave radiation by a short, straight, symmetric antenna which is fed, for example, by a TEM transmission line such as a coaxial cable – see Fig. 3.

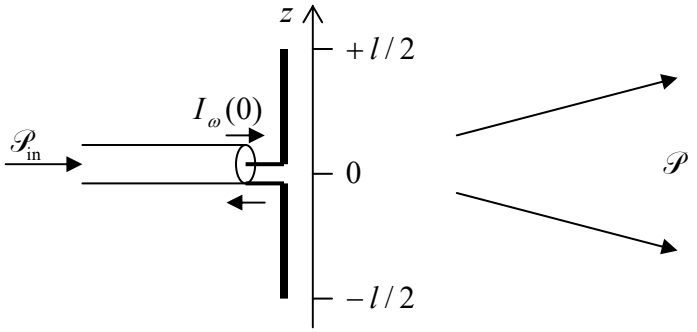


Fig. 8.3. The dipole antenna.

The exact solution of this problem is rather complicated because the law $I_\omega(z)$ of the current variation along the antenna's length should be calculated self-consistently with the distribution of the electromagnetic field induced by the current in the surrounding space. (Unfortunately, this fact is not mentioned in some textbooks.) However, the current should be largest in the feeding point (in Fig. 3, taken for $z = 0$) and vanish at the antenna's ends ($z = \pm l/2$), and hence we may guess that at $l << \lambda$, the linear function

$$I_\omega(z) = I_\omega(0)\left(1 - \frac{2}{l}|z|\right), \tag{8.30}$$

should be a good approximation of the actual distribution – as it indeed is. Now we can use the continuity equation $\partial Q/\partial t = I$, i.e. $-i\omega Q_\omega = I_\omega$, to calculate the complex amplitude $Q_\omega(z) = iI_\omega(z)\mathrm{sgn}(z)/\omega$ of the electric charge $Q(z, t) = \mathrm{Re}[Q_\omega \exp\{-i\omega t\}]$ of the wire's segment $[0, z]$, and from it, the amplitude of the charge's linear density:

$$\lambda_\omega(z) \equiv \frac{dQ_\omega(z)}{d|z|} = -i\frac{2I_\omega(0)}{\omega l}\mathrm{sgn}\, z\,. \tag{8.31}$$

From here, the dipole moment's amplitude is

$$p_\omega = 2\int_0^{l/2}\lambda_\omega(z)z\,dz = -i\frac{I_\omega(0)}{2\omega}l\,, \tag{8.32}$$

so Eq. (28) yields

[12] Actually, the formula needs a numerical coefficient adjustment to account for the electron's orbital (rather than linear) motion – the task left for the reader's exercise. However, this adjustment does not affect the order-of-magnitude estimate given above.

$$\overline{\mathscr{P}} = Z\frac{\omega^4}{12\pi v^2}\frac{|I_\omega(0)|^2}{4\omega^2}l^2 = \frac{Z(kl)^2}{24\pi}\frac{|I_\omega(0)|^2}{2}, \tag{8.33}$$

where $k = \omega/v$. The analogy between this result and the dissipation power $\mathscr{P} = \mathrm{Re}Z\,|\,I_\omega|^2/2$ in a lumped linear circuit element, enables the interpretation of the first fraction in the last form of Eq. (33) as the real part of the antenna's impedance:

$$\mathrm{Re}\,Z_A = Z\frac{(kl)^2}{24\pi}, \tag{8.34}$$

as felt by the transmission line.

According to Eq. (7.118), the wave traveling along the line toward the antenna is fully radiated, i.e. not reflected back, only if Z_A equals to Z_W of the line. As we know from Sec. 7.5 (and the solution of the related problems), for typical TEM lines, $Z_W \sim Z_0$, while Eq. (34), which is only valid in the limit $kl << 1$, shows that for the radiation into free space ($Z = Z_0$), $\mathrm{Re}Z_A$ is much less than Z_0. Hence to reach the impedance matching condition $Z_W = Z_A$, the antenna's length should be increased – as a more involved theory shows, to $l \approx \lambda/2$. However, in many cases, practical considerations make short antennas necessary. The example most often met nowadays is the cell phone antennas, which use frequencies close to 1 or 2 GHz, with free-space wavelengths λ between 15 and 30 cm, i.e. much larger than the phone size.[13] The quadratic dependence of the antenna's efficiency on l, following from Eq. (34), explains why every millimeter counts in the design of such antennas, and why the designs are carefully optimized using software packages for the (virtually exact) numerical solution of the Maxwell equations for the specific shape of the antenna and other phone parts.[14]

To conclude this section, let me note that if the wave source is not monochromatic, so $\mathbf{p}(t)$ should be represented as a Fourier series,

$$\mathbf{p}(t) = \mathrm{Re}\sum_\omega \mathbf{p}_\omega e^{-i\omega t}, \tag{8.35}$$

the terms corresponding to the interference of spectral components with different frequencies ω are averaged out at the time averaging of the Poynting vector, and the *average* radiated power is just a sum of contributions (28) from all substantial frequency components.

8.3. Wave scattering

The Larmor formula may be used as the basis of the theory of *scattering* – the phenomenon illustrated by Fig. 4. Generally, scattering is a complex problem. However, in many cases it allows the so-called *Born approximation*,[15] in which the scattered wave field's effect on the scattering object is assumed to be much weaker than that of the incident wave, and is neglected.

[13] The situation will be partly remedied by the planned transfer of wireless mobile technology to the next generations, with the signal frequencies gradually moving up.

[14] A partial list of popular software packages of this kind includes both publicly available codes such as Nec2 (whose various versions are available online, e.g., at http://www.qsl.net/4nec2/), and proprietary packages – such as *Momentum* from Agilent Technologies (now owned by Hewlett-Packard), *FEKO* from EM Software & Systems, and *XFdtd* from Remcom.

[15] Named after Max Born, one of the founding fathers of quantum mechanics. However, the basic idea of this approach was developed much earlier (in 1881) by Lord Rayleigh – born John William Strutt.

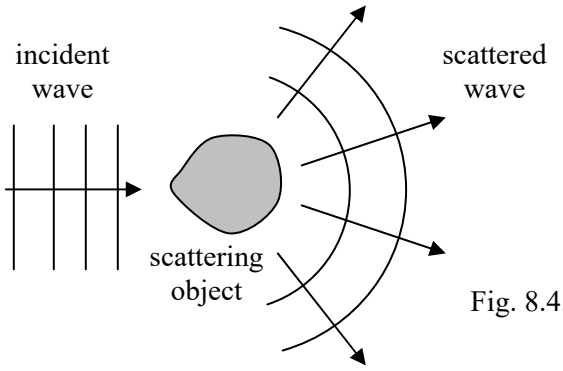


Fig. 8.4. Wave scattering (schematically).

As the first example of this approach, let us consider the scattering of a plane wave, propagating in free space ($Z = Z_0$, $v = c$), by an otherwise free[16] charged particle whose motion may be described by non-relativistic classical mechanics. (This requires, in particular, the incident wave to be not too powerful, so the speed of the charge's motion induced by the wave remains much lower than c.) As was already discussed at the derivation of Eq. (7.32), in this case, the magnetic component of the Lorentz force (5.10) is negligible in comparison with the force $\mathbf{F}_e = q\mathbf{E}$ exerted by the wave's electric field. Thus, assuming that the incident wave is linearly polarized along some axis x, the equation of the particle's motion in the Born approximation is just $m\ddot{x} = qE(t)$, so for the x-component $p_x = qx$ of its dipole moment we can write

$$\ddot{p} = q\ddot{x} = \frac{q^2}{m}E(t). \tag{8.36}$$

As we already know from Sec. 2, oscillations of the dipole moment lead to radiation of a wave with a wide angular distribution of intensity; in our case, this is the scattered wave – see Fig. 4. Its full power may be found by plugging Eq. (36) into Eq. (27):

$$\mathscr{P} = \frac{Z_0}{6\pi c^2}\ddot{p}^2 = \frac{Z_0 q^4}{6\pi c^2 m^2}E^2(t), \tag{8.37}$$

so for the average power, we get

$$\overline{\mathscr{P}} = \frac{Z_0 q^4}{12\pi c^2 m^2}\left|E_\omega\right|^2. \tag{8.38}$$

Since this power is proportional to the incident wave's intensity S, it is customary to characterize the scattering ability of an object by the ratio,

Full cross-section: definition

$$\sigma \equiv \frac{\overline{\mathscr{P}}}{\overline{S_{\text{incident}}}} \equiv \frac{\overline{\mathscr{P}}}{\left|E_\omega\right|^2/2Z_0}, \tag{8.39}$$

which has the dimensionality of area, and is called the *total cross-section* of scattering.[17] For this measure, Eq. (38) yields the famous result

[16] As Eq. (7.30) shows, this calculation is also valid for an oscillator with a low own frequency, $\omega_0 << \omega$.

[17] This definition parallels those accepted in the classical and quantum theories of *particle* scattering – see, e.g., respectively, CM Sec. 3.5 and QM Sec. 3.3.

$$\sigma = \frac{Z_0^2 q^4}{6\pi c^2 m^2} = \frac{\mu_0^2 q^4}{6\pi m^2}, \tag{8.40}$$

which is called the *Thomson scattering formula*,[18] especially when applied to an electron. This relation is most frequently represented in the form[19]

$$\sigma = \frac{8\pi}{3} r_c^2, \qquad \text{with } r_c \equiv \frac{q^2}{4\pi\varepsilon_0}\cdot\frac{1}{mc^2}. \tag{8.41}$$

Thomson scattering

This constant r_c is called the *classical radius of the particle* (or sometimes the "Thomson scattering length"); for the electron ($q = -e$, $m = m_e$) it is close to 2.82×10^{-15} m. Its possible interpretation is evident from Eq. (41) for r_c: at that distance between two similar particles, the potential energy $q^2/4\pi\varepsilon_0 r$ of their electrostatic interaction is equal to the particle's rest-mass energy mc^2.[20]

Now we have to go back and establish the conditions at which the Born approximation, when the field of the scattered wave is negligible, is indeed valid for a point-object scattering. Since the scattered wave's intensity described by Eq. (26) diverges at $r \to 0$ as $1/r^2$, according to the definition (39) of the cross-section, it may become comparable to S_{incident} at $r^2 \sim \sigma$. However, Eq. (38) itself is only valid if $r >> \lambda$, so the Born approximation does not lead to a contradiction only if

$$\sigma << \lambda^2. \tag{8.42}$$

For the Thompson scattering by an electron, this condition means $\lambda >> r_c \sim 3\times10^{-15}$ m and is fulfilled for all frequencies up to very hard γ-rays with photon energies ~100 MeV.

Possibly the most notable feature of the result (40) is its independence of the wave frequency. As it follows from its derivation, particularly from Eq. (37), this independence is intimately related to the unbound character of charge motion. For bound charges, say for electrons in gas molecules, this result is only valid if the wave frequency ω is much higher than the frequencies ω_j of most important quantum transitions. In the opposite limit, $\omega << \omega_j$, the result is dramatically different. Indeed, in this limit we may approximate the molecule's dipole moment by its static value (3.48):

$$\mathbf{p} = \alpha\mathbf{E}. \tag{8.43}$$

In the Born approximation, and in the absence of the molecular field effects mentioned in Sec. 3.3, **E** in this expression is just the incident wave's field, and we can use Eq. (28) to calculate the power of the wave scattered by a single molecule:

---

[18] Named after Sir Joseph John ("JJ") Thomson, the discoverer of the electron – and isotopes as well! He should not be confused with his son, G. P. Thomson, who discovered (simultaneously with C. Davisson and L. Germer) quantum-mechanical wave properties of the same electron.

[19] In the Gaussian units, this formula looks like $r_c = q^2/mc^2$ (giving, of course, the same numerical values: for the electron, $r_c \approx 2.82\times10^{-13}$ cm). This *classical* quantity should not be confused with the particle's *Compton wavelength* $\lambda_C \equiv 2\pi\hbar/mc$ (for the electron, close to 2.24×10^{-12} m), which naturally arises in *quantum* electrodynamics – see a brief discussion in the next chapter, and also QM Sec. 1.1.

[20] It is fascinating how smartly the *relativistic* expression mc^2 sneaked into the result (40)-(41), which was obtained using the *non-relativistic* equation (36) of the particle motion. This was possible because the calculation engaged electromagnetic waves, which propagate with the speed of light, and whose quanta (*photons*), as a result, may be frequently treated as relativistic (moreover, ultra-relativistic) particles – see the next chapter.

$$\overline{\mathcal{P}} = \frac{Z_0 \omega^4}{4\pi c^2} \alpha^2 \left|E_\omega\right|^2 . \tag{8.44}$$

Now, using the last form of the definition (39) of the cross-section, we get a very simple result,

$$\sigma = \frac{Z_0^2 \omega^4}{6\pi c^2} \alpha^2 , \tag{8.45}$$

showing that in contrast to Eq. (40), at low frequencies σ changes as fast as ω^4.

Now let us explore the effect of such *Rayleigh scattering* on wave propagation in a gas, with a relatively low volumic density n. We may expect (and will prove in the next section) that due to the randomness of molecule positions, the waves scattered by individual molecules may be treated as *incoherent* ones, so the total scattering power may be calculated just as the sum of those scattered by each molecule. We can use this fact to write the balance of the incident's wave intensity in a small volume dV of length (along the incident wave direction) dz, and area A across it. Since such a segment includes $ndV = nAdz$ molecules, and according to Eq. (39), each of them scatters power $S\sigma = \mathcal{P}\sigma/A$, the total scattered power is $n\mathcal{P}\sigma dz$; hence the incident power's change is

$$d\mathcal{P} \equiv -n\sigma\mathcal{P}\, dz . \tag{8.46}$$

Comparing this equation with the definition (7.213) of the wave attenuation constant, applied to the scattering, [21]

$$d\mathcal{P} \equiv -\alpha_{\text{scat}}\mathcal{P}\, dz . \tag{8.47}$$

we see that this effect gives the following contribution to attenuation: $\alpha_{\text{scat}} = n\sigma$. From here, using Eq. (3.50) to write $\alpha = \varepsilon_0(\kappa - 1)/n$, where κ is the dielectric constant, and Eq. (45) for σ, we get

Rayleigh scattering

$$\boxed{\alpha_{\text{scat}} = \frac{k^4}{6\pi n}(\kappa - 1)^2, \qquad \text{where } k \equiv \frac{2\pi}{\lambda_0} = \frac{\omega}{c}.} \tag{8.48}$$

This is the famous *Rayleigh scattering formula*, which in particular explains the colors of blue sky and red sunsets. Indeed, through the visible light spectrum, ω changes almost two-fold; as a result, the scattering of blue components of sunlight is an order of magnitude higher than that of its red components. For the air near the Earth's surface, $\kappa - 1 \approx 6\times10^{-4}$, and $n \sim 2.5\times10^{25}$ m^{-3} – see Sec. 3.3. Plugging these numbers into Eq. (48), we see that the effective length $l_{\text{scat}} \equiv 1/\alpha_{\text{scat}}$ of scattering is ~30 km for the blue light and ~200 km for the red light.[22] The effective thickness h of the Earth's atmosphere is ~10 km, so the Sun looks just a bit yellowish during most of the day. However, an elementary geometry shows that at sunset, the light has to pass the length $l \sim (R_E h)^{1/2} \approx 300$ km to reach an Earth-surface observer; as a result, the blue components of the Sun's light spectrum are almost completely scattered out, and even the red components are weakened very substantially.

[21] I am sorry for using the same letter (α) for both the molecular polarizability and the wave attenuation, but both notations are traditional. Hopefully, the subscript "scat" marking α in the latter meaning minimizes the possibility of confusion.

[22] These values are approximate because both n and $(\kappa - 1)$ vary through the atmosphere's thickness.

8.4. Interference and diffraction

Now let us discuss scattering by objects with a size of the order of, or even larger than λ. For such extended objects, the phase difference factors (neglected above) step in, leading in particular to the important effects of *interference* and *diffraction*. These effects show up not as much in the total power of the scattered radiation, as in its angular distribution. It is common to characterize this distribution by the *differential cross-section* defined as

$$\frac{d\sigma}{d\Omega} \equiv \frac{\overline{S}_r r^2}{S_{\text{incident}}}, \tag{8.49}$$

Differential cross-section

where r is the distance from the scatterer, at which the scattered wave is observed.[23] Both the definition and the notation may become clearer if we notice that according to Eq. (26), at large distances ($r >> a$), the numerator of the right-hand side of Eq. (49), and hence the differential cross-section as a whole, do not depend on r, and that its integral over the total solid angle $\Omega = 4\pi$ coincides with the total cross-section defined by Eq. (39):

$$\oint_{4\pi} \frac{d\sigma}{d\Omega} d\Omega = \frac{1}{S_{\text{incident}}} r^2 \oint_{4\pi} \overline{S}_r d\Omega = \frac{1}{S_{\text{incident}}} \oint_{r=\text{const}} \overline{S}_r d^2 r = \frac{\overline{\mathscr{P}}}{S_{\text{incident}}} \equiv \sigma . \tag{8.50}$$

For example, according to Eq. (26), the angular distribution of the radiation scattered by a single dipole is rather broad; in particular, in the quasistatic case (43), within the Born approximation,

$$\frac{d\sigma}{d\Omega} = \left(\frac{\alpha k^2}{4\pi\varepsilon_0} \right)^2 \sin^2 \Theta . \tag{8.51}$$

If the wave is scattered by a small dielectric body, with a characteristic size $a << \lambda$ (i.e., $ka << 1$), then all its parts re-radiate the incident wave coherently. Hence, we can calculate it similarly, just replacing the molecular dipole moment (43) with the total dipole moment of the object – see Eq. (3.45):

$$\mathbf{p} = \mathbf{P}V = (\kappa - 1)\varepsilon_0 \mathbf{E} V , \tag{8.52}$$

where $V \sim a^3$ is the body's volume. As a result, the differential cross-section may be obtained from Eq. (51) with the replacement $\alpha_{\text{mol}} \to (\kappa - 1)\varepsilon_0 V$:

$$\frac{d\sigma}{d\Omega} = \left(\frac{k^2 V}{4\pi} \right)^2 (\kappa - 1)^2 \sin^2 \Theta , \tag{8.53}$$

i.e. follows the same $\sin^2\Theta$ law.

The situation for extended objects, with at least one dimension of the order of (or larger than) the wavelength, is different: here we have to take into account the phase shifts between the wave's re-radiation by various parts of the body. Let us analyze this issue first for an arbitrary collection of similar point scatterers located at points $\mathbf{r}_j$. If the wave vector of the incident plane wave is $\mathbf{k}_0$, the wave's field has the phase factor $\exp\{i\mathbf{k}_0 \cdot \mathbf{r}\}$ – see Eq. (7.79). At the location $\mathbf{r}_j$ of the j^{th} scattering center, this factor equals $\exp\{i\mathbf{k}_0 \cdot \mathbf{r}_j\}$, defining the time dependence of the dipole vector $\mathbf{p}$, and hence of the scattered wave.

[23] Just as in the case of the total cross-section, this definition is also similar to that accepted at the *particle* scattering – see, e.g., CM Sec. 3.5 and QM Sec. 3.3.

According to Eq. (17), the scattered wave with a wave vector $\mathbf{k}$ (with $k = k_0$) acquires, on its way from the source point $\mathbf{r}_j$ to the observation point $\mathbf{r}$, an additional phase factor $\exp\{i\mathbf{k}\cdot(\mathbf{r} - \mathbf{r}_j)\}$, so the scattered wave field is proportional to

$$\exp\left\{i\mathbf{k}_0\cdot\mathbf{r}_j + i\mathbf{k}(\mathbf{r}-\mathbf{r}_j)\right\} \equiv e^{i\mathbf{k}\cdot\mathbf{r}}\exp\left\{-i(\mathbf{k}-\mathbf{k}_0)\cdot\mathbf{r}_j\right\}. \tag{8.54}$$

Since the first factor in the last expression does not depend on $\mathbf{r}_j$, to calculate the total scattering wave, it is sufficient to sum up the last phase factors, $\exp\{-i\mathbf{q}\cdot\mathbf{r}_j\}$, where the vector

$$\mathbf{q} \equiv \mathbf{k} - \mathbf{k}_0 \tag{8.55}$$

has the physical sense of the wave vector change at scattering.[24] It may look like the phase factor depends on our choice of the reference frame. However, according to Eq. (7.42), the average *intensity* of the scattered wave is proportional to $E_\omega E_\omega^*$, i.e. to the following real scalar function of the vector $\mathbf{q}$:

Scattering function

$$F(\mathbf{q}) = \left(\sum_j \exp\{-i\mathbf{q}\cdot\mathbf{r}_j\}\right)\left(\sum_{j'} \exp\{-i\mathbf{q}\cdot\mathbf{r}_{j'}\}\right)^* \equiv \sum_{j,j'} \exp\{i\mathbf{q}\cdot(\mathbf{r}_j - \mathbf{r}_{j'})\} = \left|I(\mathbf{q})\right|^2, \tag{8.56}$$

where the complex function

Phase sum

$$I(\mathbf{q}) \equiv \sum_j \exp\{-i\mathbf{q}\cdot\mathbf{r}_j\} \tag{8.57}$$

is called the *phase sum*, and may be calculated in any reference frame without affecting the final result given by Eq. (56).

So, besides the $\sin^2\Theta$ factor, the differential cross-section (49) of scattering by an extended object is also proportional to the scattering function (56). Its double-sum form is convenient to notice that for a system of *many* ($N >> 1$) similar but randomly located scatterers, only the terms with $j = j'$ accumulate at summation, so $F(\mathbf{q})$, and hence $d\sigma/d\Omega$, scale as N, rather than N^2 – thus justifying again our treatment of the Rayleigh scattering problem in the previous section.

Now let us apply Eq. (56) to a simple problem of just *two* similar small scatterers, separated by a fixed distance a:

$$F(\mathbf{q}) = \sum_{j,j'=1}^{2} \exp\{i\mathbf{q}\cdot(\mathbf{r}_j - \mathbf{r}_{j'})\} = 2 + \exp\{-iq_a a\} + \exp\{iq_a a\} = 2\left(1+\cos q_a a\right) = 4\cos^2\frac{q_a a}{2}, \tag{8.58}$$

where $q_a \equiv \mathbf{q}\cdot\mathbf{a}/a$ is the component of the vector $\mathbf{q}$ along the vector $\mathbf{a}$ connecting the scatterers. The apparent simplicity of this result may be a bit misleading because the mutual plane of the vectors $\mathbf{k}$ and $\mathbf{k}_0$ (and hence of the vector $\mathbf{q}$) does not necessarily coincide with the mutual plane of the vectors $\mathbf{k}_0$ and $\mathbf{E}_\omega$, so the *scattering angle* θ between $\mathbf{k}$ and $\mathbf{k}_0$ is generally different from $(\pi/2 - \Theta)$ – see Fig. 5. Moreover, the angle between the vectors $\mathbf{q}$ and $\mathbf{a}$ (within their common plane) is one more parameter independent of both θ and Θ. As a result, the angular dependence of the scattered wave's intensity (and hence $d\sigma/d\Omega$), which depends on all three angles, may be rather involved, but some of its details are irrelevant for the basic physics of interference/diffraction.

[24] In quantum mechanics, $\hbar\mathbf{q}$ has a very clear sense of the momentum transferred from the scattering object to the scattered particle (for example, a photon), and this terminology is sometimes smuggled even into classical electrodynamics texts.

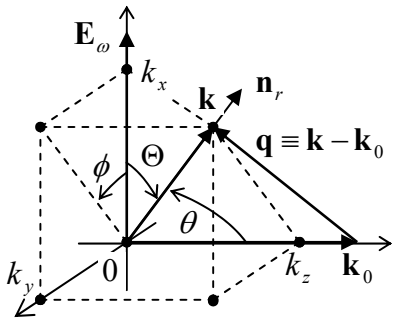


Fig. 8.5. The angles important for the general scattering problem.

This is why let me consider in detail only the simple cases when the vectors **k**, $\mathbf{k}_0$, and **a** all reside in the same plane, with $\mathbf{k}_0$ normal to **a** – see Fig. 6a.

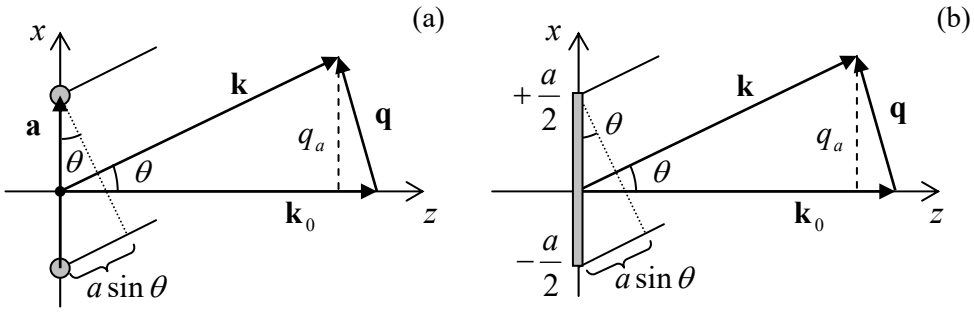


Fig. 8.6. The simplest cases of (a) interference and (b) diffraction.

In this case, $q_a = k\sin\theta$, and Eq. (58) is reduced to

$$F(\mathbf{q}) = 4\cos^2\frac{ka\sin\theta}{2}. \tag{8.59}$$

This function always has two maxima, at $\theta = 0$ and $\theta = \pi$, and, if the product ka is large enough, other maxima[25] at the special angles θ_n that satisfy the simple *Bragg condition*

$$ka\sin\theta_n = 2\pi m, \qquad \text{i.e. } a\sin\theta_n = n\lambda. \tag{8.60}$$

Bragg condition

As Fig. 6a shows, this condition may be readily understood as that of the in-phase addition (the *constructive interference*) of two coherent waves scattered from the two points, when the difference between their paths toward the observer, $a\sin\theta$, is equal to an integer number of wavelengths. At each such maximum, $F = 4$, due to the doubling of the wave amplitude and hence quadrupling its power.

If the distance between the point scatterers is large ($ka >> 1$), the first maxima (60) correspond to small scattering angles, $\theta << 1$. For this region, Eq. (59) is reduced to a simple periodic dependence of function F on the angle θ. Moreover, within the range of small θ, the wave polarization factor $\sin^2\Theta$ is virtually constant, so the angular dependence of the scattered wave's intensity, and hence of the differential cross-section, is also very simple:

$$\frac{d\sigma}{d\Omega} \propto F(\mathbf{q}) = 4\cos^2\frac{ka\theta}{2}. \tag{8.61}$$

Young's interference pattern

[25] In optics especially, such intensity maxima/minima patterns are called *interference fringes*.

This simple *interference pattern* is well known from Young's two-slit experiment.[26] (As will be discussed in the next section, the theoretical description of the two-slit experiment is more complex than that of the Born scattering, but is preferable experimentally because, at such scattering, the wave of intensity (61) has to be observed on the backdrop of a much stronger incident wave that propagates in almost the same direction, $\theta = 0$.)

A very similar analysis of scattering from $N > 2$ similar, equidistant scatterers, located along the same straight line shows that the positions (60) of the constructive interference maxima do not change (because the derivation of this condition is still applicable to each pair of adjacent scatterers), but the increase of N makes these peaks sharper and sharper. Leaving the quantitative analysis of this system for the reader's exercise, let me jump immediately to the limit $N \to 0$, in which we may ignore the scatterers' discreteness. The resulting pattern is similar to that at scattering by a continuous thin rod (see Fig. 6b), so let us first spell out the Born scattering formula for an arbitrary extended, continuous, uniform dielectric body. Transferring Eq. (56) from the sum to an integral, for the differential cross-section we get

$$\frac{d\sigma}{d\Omega} = \left(\frac{k^2}{4\pi}\right)^2 (\kappa - 1)^2 F(\mathbf{q}) \sin^2\Theta \equiv \left(\frac{k^2}{4\pi}\right)^2 (\kappa - 1)^2 \left|I(\mathbf{q})\right|^2 \sin^2\Theta, \tag{8.62}$$

where $I(\mathbf{q})$ now becomes the *phase integral*,[27]

Phase integral

$$I(\mathbf{q}) = \int_V \exp\{-i\mathbf{q}\cdot\mathbf{r}'\} d^3r', \tag{8.63}$$

with the dimensionality of volume.

Now we may return to the particular case of a thin rod (with both dimensions of the cross-section's area A much smaller than λ, but an arbitrary length a), otherwise keeping the same simple geometry as for two point scatterers – see Fig. 6b. In this case, the phase integral is just

Fraunhofer diffraction integral

$$I(\mathbf{q}) = A\int_{-a/2}^{+a/2} \exp\{-iq_a x')dx' = A\frac{\exp\{-iq_a a/2\} - \exp\{iq_a a/2\}}{-iq} \equiv V\frac{\sin\xi}{\xi}, \tag{8.64}$$

where $V = Aa$ is the volume of the rod, and ξ is the dimensionless argument defined as

$$\xi \equiv \frac{q_a a}{2} \equiv \frac{ka\sin\theta}{2}. \tag{8.65}$$

The fraction participating in the last form of Eq. (64) is met in physics so frequently that it has deserved the special name of the *sinc* (not "sync", please!) *function* (see Fig. 7):

[26] This experiment was described in 1803 by Thomas Young – one more universal genius of science, who also introduced the Young modulus in the elasticity theory (see, e.g., CM Chapter 7), besides numerous other achievements – including deciphering Egyptian hieroglyphs! It is fascinating that the first clear observation of wave interference was made as early as 1666 by another genius, Sir Isaac Newton, in the form of so-called *Newton's rings*. Unbelievably, Newton failed to give the most natural explanation of his observations – perhaps because he was vehemently opposed to the very idea of light as a wave, which was promoted in his times by others, notably by Christian Huygens. Due to Newton's enormous authority, only Young's two-slit experiments more than a century later have firmly established the wave picture of light – to be replaced by the dualistic wave/photon picture formalized by quantum electrodynamics (see, e.g., QM Ch. 9), in one more century.

[27] Since the observation point's position $\mathbf{r}$ does not participate in this formula explicitly, the prime sign in $\mathbf{r}'$ could be dropped, but I keep it as a reminder that the integral is taken over points $\mathbf{r}'$ of the *scattering object*.

$$\operatorname{sinc}\xi \equiv \frac{\sin\xi}{\xi}. \tag{8.66}$$

Sinc function

It vanishes at all points $\xi_n = \pi n$ with integer n, besides such point with $n = 0$: $\operatorname{sinc}\xi_0 \equiv \operatorname{sinc} 0 = 1$.

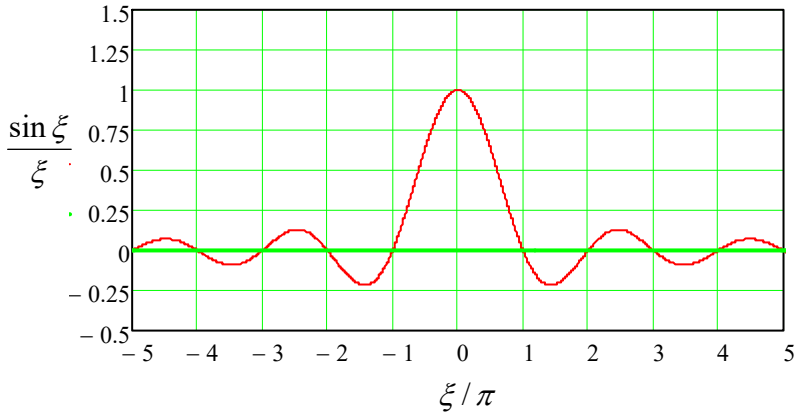


Fig. 8.7. The sinc function.

The function $F(\mathbf{q}) = V^2\operatorname{sinc}^2\xi$, given by Eq. (64) and plotted with the red line in Fig. 8, is called the *Fraunhofer diffraction pattern*.[28]

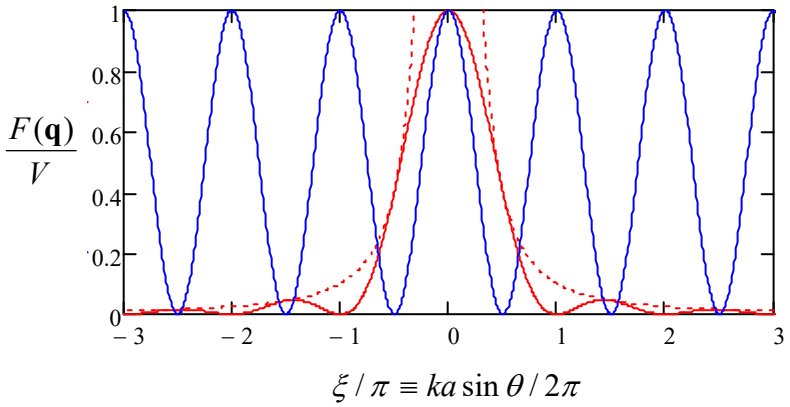


Fig. 8.8. The Fraunhofer diffraction pattern (solid red line) and its envelope $1/\xi^2$ (dashed red line). For comparison, the blue line shows the basic interference pattern $\cos^2\xi$ – cf. Eq. (61).

Note that it oscillates with the same argument's period $\Delta(ka\sin\theta) = 2\pi/ka$ as the interference pattern (61) from two-point scatterers (shown with the blue line in Fig. 8). However, at the interference, the scattered wave intensity vanishes at angles θ_n' that satisfy the condition

$$\frac{ka\sin\theta_n'}{2\pi} = n + \frac{1}{2}, \tag{8.67}$$

i.e. when the optical path difference $a\sin\alpha$ is equal to a semi-integer number of wavelengths $\lambda/2 = \pi/k$, and hence the two waves from the scatterers reach the observer in anti-phase – the so-called *destructive interference*. On the other hand, for the diffraction on a continuous rod the minima occur at a different set of scattering angles:

$$\frac{ka\sin\theta_n}{2\pi} = n, \tag{8.68}$$

[28] It is named after Joseph von Fraunhofer (1787-1826) – who invented the spectroscope, developed the diffraction grating (see below), and also discovered the dark *Fraunhofer lines* in the Sun's spectrum.

i.e. exactly where the two-point interference pattern has its maxima – please have a look at Fig. 8 again. The reason for this relation is that the wave diffraction on the rod may be considered as a simultaneous interference of waves from all its elementary fragments, and exactly at the observation angles when the rod edges give waves with phases shifted by $2\pi n$, the interior points of the rod give waves with all phases within this difference, with their algebraic sum equal to zero. As even more visible in Fig. 8, at the diffraction, the intensity oscillations are limited by a rapidly decreasing envelope function $1/\xi^2$ – while at the two-point interference, the oscillations retain the same amplitude. The reason for this fast decrease is that with each Fraunhofer diffraction period, a smaller and smaller fraction of the road gives an unbalanced contribution to the scattered wave.

If the rod's length is small ($ka << 1$, i.e. $a << \lambda$), then the sinc function's argument ξ is small at all scattering angles θ, so $I(\mathbf{q}) \approx V$, and Eq. (62) is reduced to Eq. (53). In the opposite limit, $a >> \lambda$, the first zeros of the function $I(\mathbf{q})$ correspond to very small angles θ, for which $\sin\Theta \approx 1$, so the differential cross-section is just

$$\frac{d\sigma}{d\Omega} = \left(\frac{k^2}{4\pi}\right)^2 (\kappa - 1)^2 \operatorname{sinc}^2 \frac{ka\theta}{2}, \tag{8.69}$$

i.e. Fig. 8 shows the scattering intensity as a function of the direction toward the observation point – if this point is within the plane containing the rod.

Finally, let us discuss a problem of large importance for applications: calculate the positions of the maxima of the interference pattern arising at the incidence of a plane wave on a very large 3D periodic system of point scatterers. For that, first of all, let us quantify the notion of 3D periodicity. The periodicity in one dimension is simple: the system we are considering (say, the positions of point scatterers) should be invariant with respect to the linear translation by some period a, and hence by any multiple sa of this period, where s is any integer. Anticipating the 3D generalization, we may require any of the possible *translation vectors* $\mathbf{R}$ to that the system is invariant, to be equal $s\mathbf{a}$, where the *primitive vector* $\mathbf{a}$ is directed along the (so far, the only) axis of the 1D system.

Now we are ready for the common definition of the 3D periodicity – as the invariance of the system with respect to the translation by any vector of the following set:

Bravais lattice

$$\mathbf{R} = \sum_{l=1}^{3} s_l \mathbf{a}_l, \tag{8.70}$$

where s_l are three independent integers, and $\{\mathbf{a}_l\}$ is a set of three linearly independent *primitive vectors*. The set of geometric points described by Eq. (70) is called the *Bravais lattice* (first analyzed in detail, circa 1850, by Auguste Bravais). Perhaps the most nontrivial feature of this relation is that the vectors $\mathbf{a}_l$ should not necessarily be orthogonal to each other. (That requirement would severely restrict the set of possible lattices and make it unsuitable for the description, for example, of many solid-state crystals.) For the scattering problem we are considering, we will assume that the position $\mathbf{r}_j$ of each point scatterer coincides with one of the points $\mathbf{R}$ of some Bravais lattice, with a given set of primitive vectors $\mathbf{a}_l$, so in the basic Eq. (57), the index j is coding the set of three integers $\{s_1, s_2, s_3\}$.

Now let us consider a similarly defined Bravais lattice, but in the reciprocal (wave-number) space, numbered by three independent integers $\{t_1, t_2, t_3\}$:

Reciprocal lattice

$$\mathbf{Q} = \sum_{m=1}^{3} t_m \mathbf{b}_m, \qquad \text{with } \mathbf{b}_m = 2\pi \frac{\mathbf{a}_{m''} \times \mathbf{a}_{m'}}{\mathbf{a}_m \cdot (\mathbf{a}_{m''} \times \mathbf{a}_{m'})}, \tag{8.71}$$

where in the last expression, the indices *m*, *m'*, and *m''* are all different. This is the so-called *reciprocal lattice*, which plays an important role in all physics of periodic structures, in particular in the quantum energy-band theory.[29] To reveal its most important property, and thus justify the above introduction of the primitive vectors $\mathbf{b}_m$, let us calculate the following scalar product:

$$\mathbf{R}\cdot\mathbf{Q} \equiv \sum_{l,m=1}^{3} s_l t_m \mathbf{a}_l \cdot \mathbf{b}_m \equiv 2\pi \sum_{l,m=1}^{3} s_l t_m \mathbf{a}_l \cdot \frac{\mathbf{a}_{m''}\times\mathbf{a}_{m'}}{\mathbf{a}_m\cdot\left(\mathbf{a}_{m''}\times\mathbf{a}_{m'}\right)} \equiv 2\pi \sum_{l,m=1}^{3} s_l t_k \frac{\mathbf{a}_l\cdot\left(\mathbf{a}_{m''}\times\mathbf{a}_{m'}\right)}{\mathbf{a}_m\cdot\left(\mathbf{a}_{m''}\times\mathbf{a}_{m'}\right)}. \quad (8.72)$$

Applying to the numerator of the last fraction the *operand rotation rule* of vector algebra,[30] we see that it is equal to zero if $l \neq m$, while for $l = m$ the whole fraction is evidently equal to 1. Thus the double sum (72) is reduced to a single sum:

$$\mathbf{R}\cdot\mathbf{Q} = 2\pi\sum_{l=1}^{3} s_l t_l = 2\pi\sum_{l=1}^{3} n_l\,, \quad (8.73)$$

where each of the products $n_l \equiv s_l t_l$ is an integer, and hence their sum,

$$n \equiv \sum_{l=1}^{3} n_l \equiv s_1 t_1 + s_2 t_2 + s_3 t_3, \quad (8.74)$$

is an integer as well, so the main property of the direct/reciprocal lattice couple is very simple:

$$\mathbf{R}\cdot\mathbf{Q} = 2\pi n, \qquad \text{and hence } \exp\{-i\mathbf{R}\cdot\mathbf{Q}\} = 1\,. \quad (8.75)$$

Now returning to the scattering function (56) for a Bravais lattice of point scatters, we see that if the vector $\mathbf{q} \equiv \mathbf{k} - \mathbf{k}_0$ coincides with *any* vector $\mathbf{Q}$ of the reciprocal lattice, then all terms of the phase sum (57) take their largest possible values (equal to 1), and hence the sum as the whole is largest as well, giving a constructive interference maximum. This equality, $\mathbf{q} = \mathbf{Q}$, where $\mathbf{Q}$ is given by Eq. (71), is called the *von Laue condition* (named after Max von Laue) of the constructive interference; it is, in particular, the basis of the whole field of the X-ray crystallography of solids and polymers – the main tool for revealing their atomic/molecular structure.[31]

In order to recast the von Laue condition in a more vivid geometric form, let us consider one of the vectors $\mathbf{Q}$ of the reciprocal lattice, corresponding to a certain integer *n* in Eq. (75), and notice that if that relation is satisfied for one point $\mathbf{R}$ of the direct Bravais lattice (70), i.e. for one set of the integers $\{s_1, s_2, s_3\}$, it is also satisfied for a 2D system of other integer sets, which may be parameterized, for example, by two integers S_1 and S_2:

$$s_1' = s_1 + S_1 t_3, \quad s_2' = s_2 + S_2 t_3, \quad s_3' = s_3 - S_1 t_1 - S_2 t_2\,. \quad (8.76)$$

Indeed, each of these sets has the same value of the integer *n*, defined by Eq. (74), as the original one:

$$n' \equiv s_1' t_1 + s_2' t_2 + s_3' t_3 \equiv \left(s_1 + S_1 t_3\right)t_1 + \left(s_2 + S_2 t_3\right)t_2 + \left(s_3 - S_1 t_1 - S_2 t_2\right)t_3 = n\,. \quad (8.77)$$

Since, according to Eq. (75), the vector of the distance between any pair of the corresponding points of the direct Bravais lattice (70),

[29] See, e.g., QM Sec. 3.4, where several particular Bravais lattices $\mathbf{R}$, and their reciprocals $\mathbf{Q}$, are considered.

[30] See, e.g., MA Eq. (7.6).

[31] For more reading on this important topic, I can recommend, for example, the classical monograph by B. Cullity, *Elements of X-Ray Diffraction*, 2nd ed., Addison-Wesley, 1978. (Note that its title uses the alternative name of the field, once again illustrating how blurry the boundary between the interference and diffraction is.)

$$\Delta\mathbf{R} = \Delta S_1 t_3 \mathbf{a}_1 + \Delta S_2 t_3 \mathbf{a}_2 - (\Delta S_1 t_1 + \Delta S_2 t_2)\mathbf{a}_3, \tag{8.78}$$

satisfies the condition $\Delta\mathbf{R}\cdot\mathbf{Q} = 2\pi\Delta n = 0$, this vector is normal to the (fixed) vector $\mathbf{Q}$. Hence, all the points corresponding to the 2D set (76) with arbitrary integers S_1 and S_2, are located on one geometric plane, called the *crystal* (or "lattice") *plane*. In a 3D system of $N >> 1$ scatterers (such as $N \sim 10^{20}$ atoms in a ~1-mm^3 solid crystal), with all linear dimensions comparable, such a plane contains $\sim N^{2/3} >> 1$ points. As a result, the constructive interference peaks are very sharp.

Now rewriting Eq. (75) as a relation for the vector $\mathbf{R}$'s component along the vector $\mathbf{Q}$,

$$R_Q = \frac{2\pi}{Q} n, \qquad \text{where } R_Q \equiv \mathbf{R}\cdot\mathbf{n}_Q \equiv \mathbf{R}\cdot\frac{\mathbf{Q}}{Q}, \quad \text{and } Q \equiv |\mathbf{Q}|, \tag{8.79}$$

we see that the parallel crystal planes corresponding to different numbers n (but the same $\mathbf{Q}$) are located in space periodically, with the smallest distance

$$d = \frac{2\pi}{Q}, \tag{8.80}$$

so the von Laue condition $\mathbf{q} = \mathbf{Q}$ may be rewritten as the following rule for the possible magnitudes of the scattering vector $\mathbf{q} \equiv \mathbf{k} - \mathbf{k}_0$:

$$q = \frac{2\pi n}{d}. \tag{8.81}$$

Figure 9a shows the diagram of the three wave vectors $\mathbf{k}$, $\mathbf{k}_0$, and $\mathbf{q}$, taking into account the elastic scattering condition $|\mathbf{k}| = |\mathbf{k}_0| = k \equiv 2\pi/\lambda$. From the diagram, we immediately get the famous *Bragg rule*[32] for the (equal) angles $\alpha \equiv \theta/2$ between the crystal plane and each of the vectors $\mathbf{k}$ and $\mathbf{k}_0$:

Bragg rule

$$k\sin\alpha = \frac{q}{2} = \frac{\pi n}{d}, \qquad \text{i.e. } 2d\sin\alpha = n\lambda. \tag{8.82}$$

The physical sense of this relation is very simple – see Fig. 9b drawn in the "direct" space of the radius-vectors $\mathbf{r}$, rather than in the reciprocal space of the wave vectors, as Fig. 9a. It shows that if the Bragg condition (82) is satisfied, the total difference $2d\sin\alpha$ of the optical paths of two waves, partly reflected from the adjacent crystal planes, is equal to an integer number of wavelengths, so these waves interfere constructively.

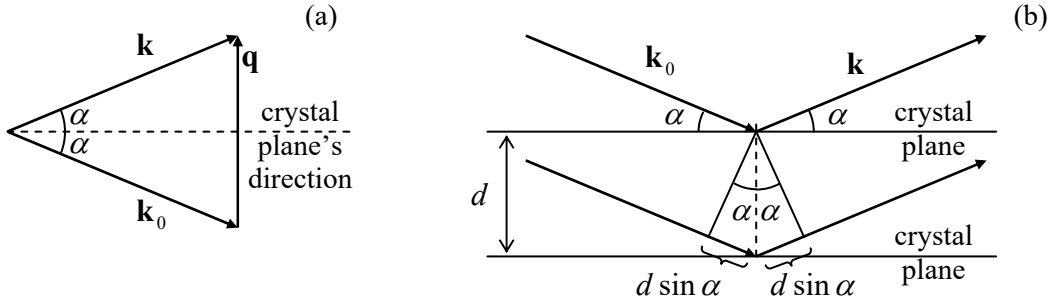


Fig. 8.9. Deriving the Bragg rule: (a) from the von Laue condition (in the reciprocal space), and (b) from a direct-space diagram. Note that the scattering angle θ equals 2α.

[32] Named after Sir William Bragg and his son, Sir William Lawrence Bragg, who were the first to demonstrate (in 1912) the X-ray diffraction by atoms in crystals. The Braggs' experiments have made the existence of atoms (before that, a somewhat hypothetical notion ignored by many physicists) indisputable.

Finally, note that the von Laue and Bragg rules, as well as the similar condition (60) for the 1D system of scatterers, are valid not only in the Born approximation but also follow from any adequate theory of scattering, because the phase sum (57) does not depend on the magnitude of the wave propagating from each elementary scatterer, provided that they are all equal.

8.5. The Huygens principle

As the reader could see, the Born approximation is very convenient for tracing the basic features of (and the difference between) the phenomena of interference and diffraction. Unfortunately, this approximation, based on the relative weakness of the scattered wave, cannot be used to describe more typical experimental implementations of these phenomena, for example, Young's two-slit experiment, or diffraction on a single slit or orifice – see, e.g. Fig. 10. Indeed, at such experiments, the orifice size a is typically much larger than the light's wavelength λ, and as a result, no clear decomposition of the fields to the "incident" and "scattered" waves is possible inside it.[33]

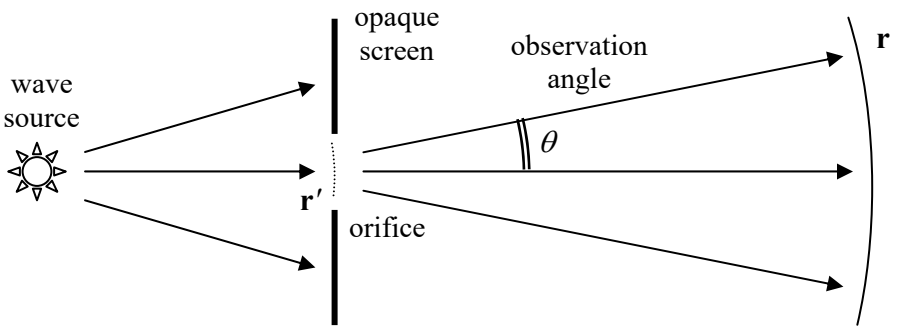


Fig. 8.10. Deriving the Huygens principle.

However, another approximation called the *Huygens* (or "Huygens-Fresnel") *principle*,[34] is very instrumental in the description of such situations. In this approach, the wave beyond the screen is represented as a linear superposition of spherical waves of the type (17), as if they were emitted by every point of the incident wave's front that has arrived at the orifice. This approximation is valid if the following strong conditions are satisfied:

$$\lambda << a << r\,, \tag{8.83}$$

where r is the distance of the observation point from the orifice. In addition, as we have seen in the last section, at small λ/a the diffraction phenomena are confined to angles $\theta \sim 1/ka \sim \lambda/a << 1$. For observation at such small angles, the mathematical expression of the Huygens principle, for the complex amplitude $f_\omega(\mathbf{r})$ of a monochromatic wave $f(\mathbf{r}, t) = \mathrm{Re}[f_\omega e^{-i\omega t}]$, is given by the following simple formula

$$f_\omega(\mathbf{r}) = C \int_{\text{orifice}} f_\omega(\mathbf{r}') \frac{e^{ikR}}{R} d^2 r'\,. \tag{8.84}$$

[33] Another complaint against the Born approximation is that it does not satisfy the so-called *optical* (or "forward scattering") *theorem* relating σ to scattering with $\mathbf{k} = \mathbf{k}_0$. This relation is especially important for the quantum-mechanical description of particle scattering, and in this series, will be discussed in its QM part (Sec. 3.3).

[34] Named after Christian Huygens (1629-1695) who had conjectured the wave nature of light (which remained controversial for more than a century, until T. Young's experiments), and Augustin-Jean Fresnel (1788-1827) who developed a quantitative theory of diffraction, and in particular gave a mathematical formulation of the Huygens principle. (Note that Eq. (91), sufficient for the purposes of this course, is not its most general form.)

Here f is any transverse component of any of the wave's fields (either **E** or **H**),[35] R is the distance between point **r**' at the orifice and the observation point **r** (i.e. the magnitude of vector **R** ≡ **r** – **r**'), and C is a complex constant.

Before describing the proof of Eq. (84), let me carry out its sanity check – which also will give us the constant C. Let us see what the Huygens principle gives for the case when the field under the integral is a plane wave with the complex amplitude $f_\omega(z)$, propagating along axis z, with an unlimited x-y front, (i.e. when there is no opaque screen at all), so in Eq. (84) we should take the whole [x, y] plane, say with $z' = 0$, as the integration area– see Fig. 11.

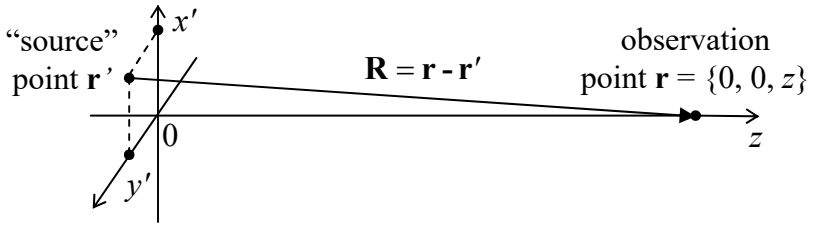


Fig. 8.11. Applying the Huygens principle to a plane incident wave.

Then, for the observation point with coordinates $x = 0$, $y = 0$, and $z > 0$, Eq. (84) yields

$$f_\omega(z) = Cf_\omega(0)\int dx'\int dy'\frac{\exp\left\{ik\left(x'^2+y'^2+z^2\right)^{1/2}\right\}}{\left(x'^2+y'^2+z^2\right)^{1/2}}. \tag{8.85}$$

Before specifying the integration limits, let us consider the range $|x'|, |y'| << z$. In this range, the square root participating in Eq. (85) twice, may be approximated as

$$\left(x'^2+y'^2+z^2\right)^{1/2} \equiv z\left(1+\frac{x'^2+y'^2}{z^2}\right)^{1/2} \approx z\left(1+\frac{x'^2+y'^2}{2z^2}\right) \equiv z+\frac{x'^2+y'^2}{2z}. \tag{8.86}$$

At $z >> \lambda$, the denominator of Eq. (85) is a much slower function of x' and y' than the exponent in the numerator, and in the former case, it is sufficient (as we will check *a posteriori*) to keep just the main, first term of expansion (86). With that, Eq. (85) becomes

$$f_\omega(z) = Cf_\omega(0)\frac{e^{ikz}}{z}\int dx'\int dy'\exp\frac{ik\left(x'^2+y'^2\right)}{2z} = Cf_\omega(0)\frac{e^{ikz}}{z}I_xI_y, \tag{8.87}$$

where I_x and I_y are two similar integrals; for example,

$$I_x = \int \exp\frac{ikx'^2}{2z}dx' = \left(\frac{2z}{k}\right)^{1/2}\int \exp\left\{i\xi^2\right\}d\xi \equiv \left(\frac{2z}{k}\right)^{1/2}\left[\int \cos\left(\xi^2\right)d\xi + i\int \sin\left(\xi^2\right)d\xi\right], \tag{8.88}$$

where $\xi \equiv (k/2z)^{1/2}x'$. These are the so-called *Fresnel integrals*. I will discuss them in more detail in the next section (see, in particular, Fig. 13), and for now, only one property[36] of these integrals is important

[35] The fact that the Huygens principle is valid for any field component should not too surprising. In the limit $a >> \lambda$, the real boundary conditions at the orifice edges are not important; it is only important for the screen that limits the orifice, to be opaque. Because of this, the Huygens principle (84) is a part of the so-called *scalar theory of diffraction*. (I will not have time to discuss the *vector theory* of these effects, which is more accurate at smaller a – see, e.g., Chapter 11 of the monograph by M. Born and E. Wolf, cited at the end of Sec. 7.1.)

for us: if taken in symmetric limits $[-\xi_0, +\xi_0]$, both of them rapidly converge to the same value, $(\pi/2)^{1/2}$, as soon as ξ_0 becomes much larger than 1. This means that even if we do not impose any exact limits on the integration area in Eq. (85), this integral converges to the value

$$f_\omega(z) = Cf_\omega(0)\frac{e^{ikz}}{z}\left\{\left(\frac{2z}{k}\right)^{1/2}\left[\left(\frac{\pi}{2}\right)^{1/2}+i\left(\frac{\pi}{2}\right)^{1/2}\right]\right\}^2 \equiv \left(C\frac{2\pi i}{k}\right)f_\omega(0)e^{ikz}, \tag{8.89}$$

due to contributions from the central area with a linear size corresponding to $\Delta\xi \sim 1$, i.e. to

$$\Delta x \sim \Delta y \sim \left(\frac{z}{k}\right)^{1/2} \sim (\lambda z)^{1/2}, \qquad . \tag{8.90}$$

so the net contribution from the front points $\mathbf{r}'$ well beyond the range (90) is negligible.[37] (Within our assumptions (83), which in particular require λ to be much less than z, the *diffraction angle* $\Delta x/z \sim \Delta y/z \sim (\lambda/z)^{1/2}$, corresponding to the important area of the front, is small.) According to Eq. (89), to sustain the unperturbed plane wave propagation, $f_\omega(z) = f_\omega(0)e^{ikz}$, the constant C has to be taken equal to $k/2\pi i$. Thus, the Huygens principle's prediction (84), in its final form, reads

$$f_\omega(\mathbf{r}) = \frac{k}{2\pi i}\int_{\text{orifice}} f_\omega(\mathbf{r}')\frac{e^{ikR}}{R}d^2r', \tag{8.91}$$

Huygens principle

and describes, in particular, the straight propagation of the plane wave (in a uniform medium).

Let me pause to emphasize how nontrivial this result is. It would be a natural corollary of Eqs. (25) (and the linear superposition principle) if all points of the orifice were filled with point scatterers that re-emit all the incident waves as spherical waves. However, as it follows from the above example, the Huygens principle also works if there is nothing in the orifice but the free space!

This is why let us discuss a proof of this principle,[38] based on Green's theorem (2.207). Let us apply it to the function $f = f_\omega$ where f_ω is the complex amplitude of a scalar component of one of the wave's fields, which satisfies the Helmholtz equation (7.204),

$$\left(\nabla^2 + k^2\right)f_\omega(\mathbf{r}) = 0, \tag{8.92}$$

and the function $g = G_\omega$ is the temporal Fourier image of the corresponding Green's function. The latter function may be defined, as usual, as the solution of the same equation with the added delta-functional right-hand side with an arbitrary coefficient, for example,

$$\left(\nabla^2 + k^2\right)G_\omega(\mathbf{r},\mathbf{r}') = -4\pi\delta(\mathbf{r}-\mathbf{r}'). \tag{8.93}$$

Using Eqs. (92) and (93) to express the Laplace operators of the functions f_ω and G_ω, we may rewrite Eq. (2.207) as

---

[36] See, e.g., MA Eq. (6.10).

[37] This result very is natural, because the function $\exp\{ikR\}$ oscillates fast with the change of $\mathbf{r}'$, so the contributions from various front points are averaged out. Indeed, the only reason why the central part of the plane $[x', y']$ gives a non-zero contribution (89) to $f_\omega(z)$ is that the phase exponents stop oscillating as $(x'^2 + y'^2)$ is reduced below $\sim z/k$ – see Eq. (86).

[38] This proof was given in 1882 by the same G. Kirchhoff whose circuit rules were discussed in Sec. 4.1 and 6.6.

$$\int_V \left\{ f_\omega \left[-k^2 G_\omega(\mathbf{r},\mathbf{r}') - 4\pi\delta(\mathbf{r}-\mathbf{r}') \right] - G_\omega(\mathbf{r},\mathbf{r}') \left[-k^2 f_\omega \right] \right\} d^3 r = \oint_S \left[f_\omega \frac{\partial G_\omega(\mathbf{r},\mathbf{r}')}{\partial n} - G_\omega(\mathbf{r},\mathbf{r}') \frac{\partial f_\omega}{\partial n} \right] d^2 r , \quad (8.94)$$

where **n** is the outward normal to the surface *S* limiting the integration volume *V*. Two terms on the left-hand side of this relation cancel, so after swapping the arguments **r** and **r**', we get

$$-4\pi f_\omega(\mathbf{r}) = \oint_S \left[f_\omega(\mathbf{r}') \frac{\partial G_\omega(\mathbf{r}',\mathbf{r})}{\partial n'} - g_\omega(\mathbf{r}',\mathbf{r}) \frac{\partial G_\omega(\mathbf{r}')}{\partial n'} \right] d^2 r' . \quad (8.95)$$

This relation is only correct if the selected volume *V* includes the point **r** (otherwise we would not get its left-hand side from the integration of the delta function), but does not include the genuine source of the wave – otherwise, Eq. (92) would have a non-zero right-hand side. Now let **r** be the field observation point, *V* be all the source-free half-space (for example, the space right of the screen in Fig. 10), so *S* is the surface of the screen, including the orifice. Then the right-hand side of Eq. (95) describes the field (at the observation point **r**) induced by the wave passing through the orifice points **r**'. Since no waves are emitted by the opaque parts of the screen, we can limit the integration by the orifice area.[39] Assuming also that the opaque parts of the screen do not re-emit the waves "radiated" by the orifice, we can take the solution of Eq. (93) to be the retarded potential for the free space:[40]

$$G_\omega(\mathbf{r},\mathbf{r}') = \frac{e^{ikR}}{R} . \quad (8.96)$$

Plugging this expression into Eq. (82), we get

Kirchhoff integral

$$-4\pi f_\omega(\mathbf{r}) = \oint_{\text{orifice}} \left[f_\omega(\mathbf{r}') \frac{\partial}{\partial n'} \left(\frac{e^{ikR}}{R} \right) - \left(\frac{e^{ikR}}{R} \right) \frac{\partial f_\omega(\mathbf{r}')}{\partial n'} \right] d^2 r' . \quad (8.97)$$

This is the so-called *Kirchhoff* (or "Fresnel-Kirchhoff") *integral*. (Again, with the integration extended over *all* boundaries of the volume *V*, this would be an exact mathematical result.) Now, let us make two additional approximations. The first of them stems from Eq. (83): at $ka >> 1$, the wave's spatial dependence in the orifice area may be represented as

$$f_\omega(\mathbf{r}') = (\text{a slow function of } \mathbf{r}') \times \exp\{i\mathbf{k}_0 \cdot \mathbf{r}'\} , \quad (8.98)$$

where "slow" means a function that changes on the scale of *a* rather than λ. If, also, $kR >> 1$, then the differentiation in Eq. (97) may be, in both instances, limited to the rapidly changing exponents, giving

$$-4\pi f_\omega(\mathbf{r}) = \oint_{\text{orifice}} i(\mathbf{k}+\mathbf{k}_0)\cdot\mathbf{n}' \frac{e^{ikR}}{R} f(\mathbf{r}') d^2 r' . \quad (8.99)$$

Second, if all observation angles are small, we may take $\mathbf{k}\cdot\mathbf{n}' \approx \mathbf{k}_0\cdot\mathbf{n}' \approx -k$. With that, Eq. (99) is reduced to the Huygens principle in its form (91).

---

[39] Actually, this is a nontrivial point of the proof. Indeed, it may be shown that the exact solution of Eq. (94) identically is equal to zero if $f(\mathbf{r}')$ and $\partial f(\mathbf{r}')/\partial n'$ vanish together at any *part* of the boundary, of a non-zero area. A more careful analysis of this issue (it is the task of the formal vector theory of diffraction, which I will not have time to pursue) confirms the validity of the described intuition-based approach at $a >> \lambda$.

[40] It follows, e.g., from Eq. (16) with a monochromatic source $q(t) = q_\omega \exp\{-i\omega t\}$, with the amplitude $q_\omega = 4\pi\varepsilon$ that fits the right-hand side of Eq. (93).

It is clear that the principle immediately gives a very simple description of the interference of waves passing through two small holes in the screen. Indeed, if the holes' sizes are negligible in comparison with the distance a between them (though still are much larger than the wavelength!), Eq. (91) yield

$$f_\omega(\mathbf{r}) = c_1 e^{ikR_1} + c_2 e^{ikR_2}, \quad \text{with } c_{1,2} \equiv k f_{1,2} A_{1,2} / 2\pi i R_{1,2}, \tag{8.100}$$

where $R_{1,2}$ are the distances between the holes and the observation point, and $A_{1,2}$ are the hole areas. For the wave intensity, Eq. (100) gives

$$\overline{S} \propto f_\omega f_\omega^* = |c_1|^2 + |c_2|^2 + 2|c_1||c_2|\cos\left[k(R_1 - R_2) + \varphi\right], \quad \text{where } \varphi \equiv \arg c_1 - \arg c_2. \tag{8.101}$$

The first two terms in the last expression clearly represent the intensities of the partial waves passed through each hole, while the last one is the result of their interference. The interference pattern's *contrast ratio*

$$\frac{\overline{S}_{\max}}{\overline{S}_{\min}} = \left(\frac{|c_1| + |c_2|}{|c_1| - |c_2|}\right)^2, \tag{8.102}$$

is the largest (infinite) when both waves have equal amplitudes.

The analysis of the interference pattern is simple if the line connecting the holes is perpendicular to wave vector $\mathbf{k} \approx \mathbf{k}_0$ – see Fig. 6a. Selecting the coordinate axes as shown in that figure, and using for the distances $R_{1,2}$ the same expansion as in Eq. (86), for the interference term in Eq. (101) we get

$$\cos\left[k(R_1 - R_2) + \varphi\right] \approx \cos\left(\frac{kxa}{z} + \varphi\right). \tag{8.103}$$

This means that the term does not depend on y, i.e. the interference pattern in the plane of constant z is a set of straight, parallel strips, perpendicular to the vector **a**, with the period given by Eq. (60), i.e. by the Bragg law.[41] This result is strictly valid only at $y^2 << z^2$; it is straightforward to use the next term in the Taylor expansion (73) to show that farther on from the interference plane $y = 0$, the strips start to diverge.

8.6. Fresnel and Fraunhofer diffraction patterns

Now let us use the Huygens principle to analyze a slightly more complex problem: plane wave's diffraction on a long, straight slit of a constant width a (Fig. 12). According to Eq. (83), to use the Huygens principle for the problem's analysis we need to have $\lambda << a << z$. Moreover, the simple version (91) of the principle is only valid for small observation angles, $|x| << z$. Note, however, that the relation between two dimensionless parameters of the problem, z/a and a/λ, which are both much less than 1, is so far arbitrary; as we will see in a minute, this relation determines the type of the observed diffraction pattern.

[41] The phase shift φ vanishes at the normal incidence of a plane wave on the holes. Note, however, that the spatial shift of the interference pattern following from Eq. (103), $\Delta x = -(z/ka)\varphi$, is extremely convenient for the experimental measurement of the phase shift between two waves, especially if it is induced by some factor (such as insertion of a transparent object into one of the interferometer's arms) that may be turned on/off at will.

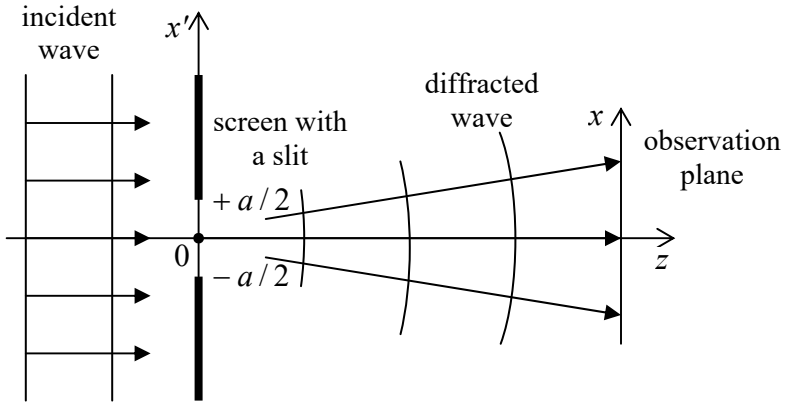


Fig. 8.12. Diffraction on a slit.

Let us apply Eq. (91) to our current problem (Fig. 12), for the sake of simplicity assuming the normal wave incidence, and taking $z' = 0$ at the screen plane:

$$f_\omega(x,z) = f_0 \frac{k}{2\pi i}\int_{-a}^{+a} dx' \int_{-\infty}^{+\infty} dy' \frac{\exp\left\{ik\left[(x-x')^2 + y'^2 + z^2\right]^{1/2}\right\}}{\left[(x-x')^2 + y'^2 + z^2\right]^{1/2}}, \tag{8.104}$$

where $f_0 \equiv f_\omega(x', 0) = \text{const}$ is the incident wave's amplitude. This is the same integral as in Eq. (85), except for the finite limits for the integration variable x', and may be simplified similarly, using the small-angle condition $(x-x')^2 + y'^2 << z^2$:

$$f_\omega(x,z) \approx f_0 \frac{k}{2\pi i}\frac{e^{ikz}}{z}\int_{-a/2}^{+a/2} dx' \int_{-\infty}^{+\infty} dy' \exp\frac{ik\left[(x-x')^2 + y'^2\right]}{2z} \equiv f_0 \frac{k}{2\pi i}\frac{e^{ikz}}{z} I_x I_y . \tag{8.105}$$

The integral over y' is the same as in the last section:

$$I_y \equiv \int_{-\infty}^{+\infty} \exp\frac{iky'^2}{2z} dy' = \left(\frac{2\pi i z}{k}\right)^{1/2}, \tag{8.106}$$

but the integral over x' is more general, because of its finite limits:

$$I_x \equiv \int_{-a/2}^{+a/2} \exp\frac{ik(x-x')^2}{2z} dx' . \tag{8.107}$$

It may be simplified in the following two (opposite) limits.

(i) *Fraunhofer diffraction* takes place when $z/a >> a/\lambda$ – the relation which may be rewritten either as $a << (z\lambda)^{1/2}$, or as $ka^2 << z$. In this limit, the ratio kx'^2/z is negligibly small for all values of x' under the integral, and we can approximate it as

$$\begin{aligned} I_x &= \int_{-a/2}^{+a/2} \exp\frac{ik\left(x^2 - 2xx' + x'^2\right)}{2z} dx' \approx \int_{-a/2}^{+a/2} \exp\frac{ik\left(x^2 - 2xx'\right)}{2z} dx' \\ &\equiv \exp\frac{ikx^2}{2z}\int_{-a/2}^{+a/2} \exp\left\{-\frac{ikxx'}{z}\right\} dx' = \frac{2z}{kx}\exp\left\{\frac{ikx^2}{2z}\right\}\sin\frac{kxa}{2z}, \end{aligned} \tag{8.108}$$

so Eq. (105) yields

$$f_\omega(x,z) \approx f_0 \frac{k}{2\pi i}\frac{e^{ikz}}{z}\frac{2z}{kx}\left(\frac{2\pi i z}{k}\right)^{1/2}\exp\left\{\frac{ikx^2}{2z}\right\}\sin\frac{kxa}{2z}, \tag{8.109}$$

and hence the relative wave intensity is

$$\frac{\bar{S}(x,z)}{S_0} = \left|\frac{f_\omega(x,z)}{f_0}\right|^2 = \frac{8z}{\pi k x^2}\sin^2\frac{kxa}{2z} \equiv \frac{2}{\pi}\frac{ka^2}{z}\operatorname{sinc}^2\left(\frac{ka\theta}{2}\right), \tag{8.110}$$

Fraunhofer diffraction pattern

where S_0 is the intensity of the incident wave, and $\theta \equiv x/z << 1$ is the observation angle. Comparing this expression with Eq. (69), we see that this diffraction pattern is exactly the same as that for a similar (uniform, 1D) object in the Born approximation – see the red line in Fig. 8. Note again that the angular width $\delta\theta$ of the Fraunhofer pattern is of the order of $1/ka$, so its linear width $\delta x = z\delta\theta$ is of the order of $z/ka \sim z\lambda/a$.[42] Hence the condition of the Fraunhofer approximation's validity may be also represented as $a << \delta x$.

(ii) *Fresnel diffraction*. In the opposite limit of a relatively wide slit, with $a >> \delta x = z\delta\theta \sim z/ka \sim z\lambda/a$, i.e. $ka^2 >> z$, the diffraction patterns at two edges of the slit are well separated. Hence, near each edge (for example, near $x' = -a/2$) we may simplify Eq. (107) as

$$I_x(x) \approx \int_{-a/2}^{+\infty} \exp\frac{ik(x-x')^2}{2z}dx' \equiv \left(\frac{2z}{k}\right)^{1/2}\int_{(k/2z)^{1/2}(x+a/2)}^{+\infty}\exp\{i\zeta^2\}d\zeta, \tag{8.111}$$

and express it via the special functions called the Fresnel integrals:[43]

$$\mathscr{C}(\xi) \equiv \left(\frac{2}{\pi}\right)^{1/2}\int_0^\xi \cos(\zeta^2)d\zeta, \qquad \mathscr{S}(\xi) \equiv \left(\frac{2}{\pi}\right)^{1/2}\int_0^\xi \sin(\zeta^2)d\zeta, \tag{8.112}$$

Fresnel integrals

whose plots are shown in Fig. 13a. As was mentioned above, at large values of their argument (ξ), both functions tend to ½.

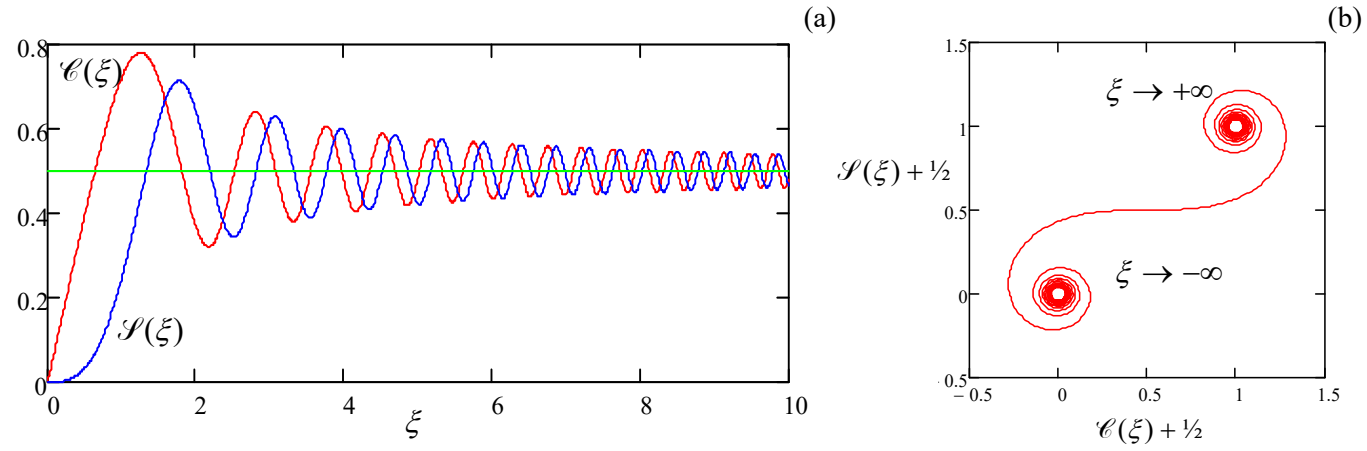


Fig. 8. 13. (a) The Fresnel integrals and (b) their parametric representation.

[42] Note also that since in this limit $ka^2 << z$, Eq. (97) shows that even the maximum value $S(0, z)$ of the diffracted wave's intensity is much lower than that (S_0) of the incident wave. This is natural because the incident power $S_0 a$ per unit length of the slit is now distributed over a much larger width $\delta x >> a$, so $S(0, z) \sim S_0 (a/\delta x) << S_0$.

[43] Slightly different definitions of these functions, affecting the constant factors, may also be met in literature.

Plugging this expression into Eqs. (105) and (111), for the diffracted wave intensity, in the Fresnel limit (i.e. at $|x + a/2| << a$), we get

Fresnel diffraction pattern

$$\frac{\overline{S}(x,z)}{S_0} = \frac{1}{2}\left\{\left[\mathscr{C}\left(\left(\frac{k}{2z}\right)^{1/2}\left(x+\frac{a}{2}\right)\right)+\frac{1}{2}\right]^2 + \left[\mathscr{S}\left(\left(\frac{k}{2z}\right)^{1/2}\left(x+\frac{a}{2}\right)\right)+\frac{1}{2}\right]^2\right\}. \tag{8.113}$$

A plot of this function (Fig. 14) shows that the diffraction pattern is very peculiar: while in the "dark" region $x < -a/2$ the wave intensity fades monotonically, the transition to the "bright" region within the gap ($x > -a/2$) is accompanied by intensity oscillations, just as at the Fraunhofer diffraction – cf. Fig. 8.

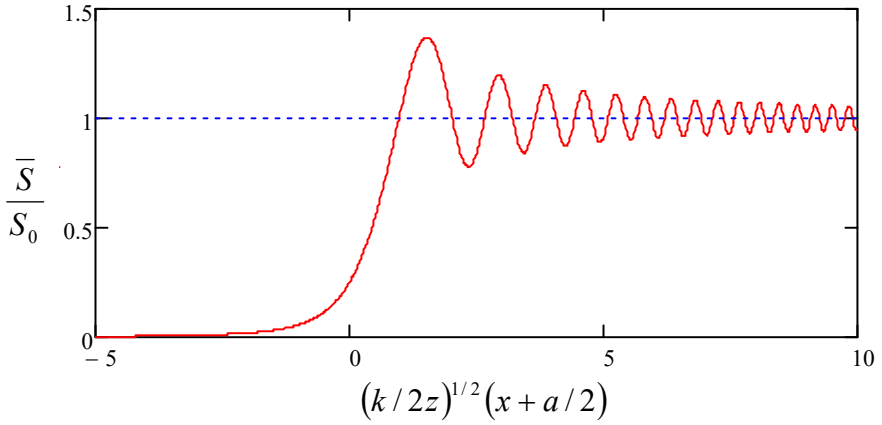


Fig. 8.14. The Fresnel diffraction pattern.

This behavior, which is described by the following asymptotes,

$$\frac{\overline{S}}{S_0} \to \begin{cases} 1 + \dfrac{1}{\sqrt{\pi}}\dfrac{\sin(\xi^2 - \pi/4)}{\xi}, & \text{for } \xi \equiv \left(\dfrac{k}{2z}\right)^{1/2}\left(x+\dfrac{a}{2}\right) \to +\infty, \\ \dfrac{1}{4\pi\xi^2}, & \text{for } \xi \to -\infty, \end{cases} \tag{8.114}$$

is essentially an artifact of "observing" just the wave intensity (i.e. its real amplitude) rather than its phase as well. Indeed, as may be seen even more clearly from the parametric representation of the Fresnel integrals, shown in Fig. 13b, these functions oscillate similarly at large positive and negative values of their argument. (This famous pattern is called either the *Euler spiral* or the *Cornu spiral*.) Physically, this means that the wave diffraction at the slit edge leads to similar oscillations of its phase at $x < -a/2$ and $x > -a/2$; however, in the latter region (i.e. inside the slit) the diffracted wave overlaps the incident wave passing through the slit directly, and their interference reveals the phase oscillations, making them visible in the observed intensity as well.

Note that according to Eq. (113), the linear scale δx of the Fresnel diffraction pattern is of the order of $(2z/k)^{1/2}$, i.e. complies with the estimate given by Eq. (90). If the slit is gradually narrowed so that its width a becomes comparable to δx,[44] the Fresnel diffraction patterns from both edges start to "collide" (interfere). The resulting wave, fully described by Eq. (107), is just a sum of two contributions of the type (111) from both edges of the slit. The resulting interference pattern is somewhat complicated, and only when a becomes substantially less than δx, it is reduced to the simple Fraunhofer pattern (110).

[44] Note that this condition may be also rewritten as $a \sim \delta x$, i.e. $z/a \sim a/\lambda$.

Of course, this crossover from the Fresnel to Fraunhofer diffraction may be also observed, at fixed wavelength λ and slit width a, by increasing z, i.e. by measuring the diffraction pattern farther and farther away from the slit.

Note also that the Fraunhofer limit is always valid if the diffraction is measured as a function of the diffraction angle θ alone. This may be done, for example, by collecting the diffracted wave with a "positive" (converging) lens and observing the diffraction pattern in its focal plane.

8.7. Geometrical optics placeholder

I would not like the reader to miss, behind all these details, the main feature of the Fresnel diffraction, which has an overwhelming practical significance. Namely, besides narrow diffraction "cones" (actually, parabolic-shaped regions) with the lateral scale $\delta x \sim (\lambda z)^{1/2}$, the wave far behind a slit of width $a >> \lambda, \delta x$, reproduces the field just behind the slit, i.e. reproduces the unperturbed incident wave inside it, and has a negligible intensity in the shade regions outside it. An evident generalization of this fact is that when a plane wave (in particular an electromagnetic wave) passes any opaque object of a large size $a >> \lambda$, it propagates around it, by distances z up to $\sim a^2/\lambda$, along straight lines, with virtually negligible diffraction effects. This fact gives the strict foundation for the notion of the *wave ray* (or *beam*), as the line perpendicular to the local front of a quasi-plane wave. In a uniform media such a ray follows a straight line,[45] but it refracts in accordance with the Snell law at the interface of two media with different values of the wave speed v, i.e. different values of the refraction index. The concept of rays enables the whole vast field of *geometric optics*, devoted mostly to ray tracing in various (sometimes very sophisticated) optical systems.

This is why, at this point, an E&M course that followed the scientific logic more faithfully than this one, would give an extended discussion of the geometric and quasi-geometric optics, including (as a minimum[46]) such vital topics as

- the so-called *lensmaker's equation* expressing the focus length f of a lens via the curvature radii of its spherical surfaces and the refraction index of the lens material,
- the *thin lens formula* relating the image distance from the lens via f and the source distance,
- the concepts of basic optical instruments such as *glasses*, *telescopes*, and *microscopes*,
- the concepts of the spherical, angular, and chromatic *aberrations* (optical image distortions).

However, since I have made a (possibly, wrong) decision to follow the common tradition in selecting the main topics for this course, I do not have time/space left for such discussion. Still, I am using this "placeholder" pseudo-section to relay my deep conviction that any educated physicist has to know the geometric optics basics. If the reader has not been exposed to this subject during their undergraduate studies, I highly recommend at least browsing one of the available textbooks.[47]

---

[45] In application to optical waves, this notion may be traced back to at least the work by Hero (a.k.a. Heron) of Alexandria (circa 170 AD). Curiously, he correctly described light reflection from one or several plane mirrors, starting from the completely wrong idea of light propagation *from* the eye of the observer *to* the observed object.

[46] Admittedly, even this list leaves aside several spectacular effects, including such a beauty as *conical refraction* in biaxial crystals – see, e.g., Chapter 15 of the textbook by M. Born and E. Wolf, cited in the end of Sec. 7.1.

[47] My top recommendation for that purpose would be Chapters 3-6 and Sec. 8.6 in Born and Wolf. A simpler alternative is Chapter 10 in G. Fowles, *Introduction to Modern Optics*, 2nd ed., Dover, 1989. Note also that the venerable field of optical microscopy is currently revitalized by holographic/tomographic methods, using the

8.8. Fraunhofer diffraction from more complex scatterers

So far, our quantitative analysis of diffraction has been limited to a very simple geometry – a single slit in an otherwise opaque screen (Fig. 12). However, in the most important Fraunhofer limit, $z >> ka^2$, it is easy to get a very simple expression for the plane wave diffraction/interference by a plane orifice (with a linear size scale a) of arbitrary shape. Indeed, the evident 2D generalization of the approximation (106)-(107) is

$$\begin{aligned} I_x I_y &= \int_{\text{orifice}} \exp\frac{ik\left[(x-x')^2+(y-y')^2\right]}{2z}dx'dy' \\ &\approx \exp\left\{\frac{ik(x^2+y^2)}{2z}\right\}\int_{\text{orifice}} \exp\left\{-i\frac{kxx'}{z}-i\frac{kyy'}{z}\right\}dx'dy', \end{aligned} \tag{8.115}$$

so besides the inconsequential total phase factor, Eq. (105) is reduced to

General Fraunhofer diffraction pattern

$$f(\boldsymbol{\rho}) \propto f_0 \int_{\text{orifice}} \exp\{-i\boldsymbol{\kappa}\cdot\boldsymbol{\rho}'\}d^2\rho' \equiv f_0 \int_{\text{all screen}} T(\boldsymbol{\rho}')\exp\{-i\boldsymbol{\kappa}\cdot\boldsymbol{\rho}'\}d^2\rho' . \tag{8.116}$$

Here the 2D vector $\boldsymbol{\kappa}$ (not to be confused with wave vector $\mathbf{k}$, which is virtually perpendicular to $\boldsymbol{\kappa}$!) is defined as

$$\boldsymbol{\kappa} \equiv k\frac{\boldsymbol{\rho}}{z} \approx \mathbf{q} \equiv \mathbf{k} - \mathbf{k}_0 , \tag{8.117}$$

and $\boldsymbol{\rho} = \{x, y\}$ and $\boldsymbol{\rho}' = \{x', y'\}$ are 2D radius vectors in the, respectively, observation and orifice planes – both nearly normal to the vectors $\mathbf{k}$ and $\mathbf{k}_0$.[48] In the last form of Eq. (116), the function $T(\boldsymbol{\rho}')$ describes the screen's transparency at point $\boldsymbol{\rho}'$, and the integral is over the whole screen plane $z' = 0$. (Though the two forms of Eq. (116) are strictly equivalent only if $T(\boldsymbol{\rho}')$ is equal to either 1 or 0, its last form may be readily obtained from Eq. (91) with $f(\mathbf{r}') = T(\boldsymbol{\rho}')f_0$ for any transparency profile, provided that $T(\boldsymbol{\rho}')$ is any function that changes substantially only at distances much larger than $\lambda \equiv 2\pi/k$.)

From the mathematical point of view, the last form of Eq. (116) is just the 2D spatial Fourier transform of the function $T(\boldsymbol{\rho}')$, with the variable $\boldsymbol{\kappa}$ defined by the observation point's position: $\boldsymbol{\rho} \equiv (z/k)\,\boldsymbol{\kappa} \equiv (z\lambda/2\pi)\,\boldsymbol{\kappa}$. This interpretation is useful because of the experience we all have with the Fourier transform, if only in the context of its time/frequency applications. For example, if the orifice is a single small hole, $T(\boldsymbol{\rho}')$ may be approximated by a delta function, so Eq. (116) yields $|f(\boldsymbol{\rho})| \approx$ const. This result corresponds (at least for the small diffraction angles $\theta \equiv \rho/z$, for which the Huygens approximation is valid) to a spherical wave spreading from the point-like orifice. Next, for two small holes, Eq. (116) immediately gives the interference pattern (103). Let me now use Eq. (116) to analyze other simplest (and most important) 1D transparency profiles, leaving a few 2D cases for the reader's exercise.

(i) A single slit of width a (Fig. 12) may be described by transparency

---

scattered wave's phase information. These methods are especially productive in biology and medicine – see, e.g., M. Brezinski, *Optical Coherence Tomography*, Academic Press, 2006, and G. Popescu, *Quantitative Phase Imaging of Cells and Tissues*, McGraw-Hill (2011).

[48] Note that for a thin uniform plate of the same shape as the orifice we are discussing now, the Born phase integral (63) with $q << k$ gives a result functionally similar to Eq. (116).

$$T(\boldsymbol{\rho}') = \begin{cases} 1, & \text{for } |x'| < a/2, \\ 0, & \text{otherwise.} \end{cases} \tag{8.118}$$

Its substitution into Eq. (116) yields

$$f(\boldsymbol{\rho}) \propto f_0 \int_{-a/2}^{+a/2} \exp\{-i\kappa_x x'\}\, dx' = f_0 \frac{\exp\{-i\kappa_x a/2\} - \exp\{i\kappa_x a/2\}}{-i\kappa_x} \propto \operatorname{sinc}\left(\frac{\kappa_x a}{2}\right) = \operatorname{sinc}\left(\frac{kxa}{2z}\right), \tag{8.119}$$

naturally returning us to Eqs. (64) and (110), and hence to the red lines in Fig. 8 for the wave intensity. (Please note again that Eq. (116) describes only the Fraunhofer, but not the Fresnel diffraction!)

(ii) Two infinitely narrow, similar, parallel slits with a larger distance a between them (i.e. the simplest model of Young's two-slit experiment) may be described by taking

$$T(\boldsymbol{\rho}') \propto \delta\left(x' - \frac{a}{2}\right) + \delta\left(x' + \frac{a}{2}\right), \tag{8.120}$$

so Eq. (116) yields the generic 1D interference pattern,

$$f(\boldsymbol{\rho}) \propto f_0 \left[\exp\left\{-\frac{i\kappa_x a}{2}\right\} + \exp\left\{\frac{i\kappa_x a}{2}\right\}\right] \propto \cos\frac{\kappa_x a}{2} = \cos\frac{kxa}{2z}, \tag{8.121}$$

whose intensity is shown with the blue line in Fig. 8.

(iii) In a more realistic model of Young's experiment, each slit has a width (say, w) that is much larger than the light wavelength λ, but still much smaller than the slit spacing a. This situation may be described by the following transparency function

$$T(\boldsymbol{\rho}') = \sum_{\pm} \begin{cases} 1, & \text{for } |x' \pm a/2| < w/2, \\ 0, & \text{otherwise,} \end{cases} \tag{8.122}$$

for which Eq. (116) yields a natural combination of the results (119) (with a replaced with w) and (121):

$$f(\mathbf{r}) \propto \operatorname{sinc}\left(\frac{kxw}{2z}\right)\cos\left(\frac{kxa}{2z}\right). \tag{8.123}$$

This is the usual interference pattern, but modulated with a Fraunhofer-diffraction envelope – shown in Fig. 15 with the dashed blue line. Since the function $\operatorname{sinc}^2\xi$ decreases very fast beyond its first zeros at $\xi = \pm\pi$, the practical number of observable interference fringes is close to $2a/w$.

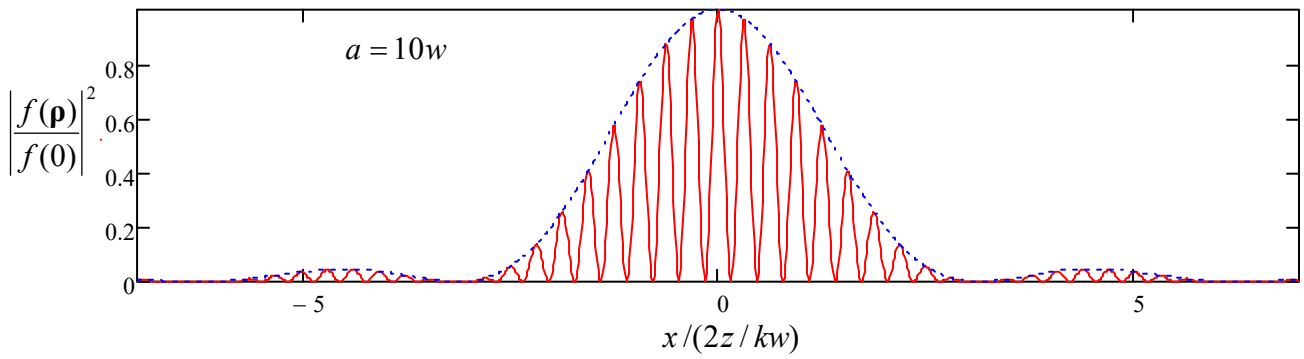


Fig. 8.15. Young's double-slit interference pattern for a finite-width slit.

(iv) A structure very useful for experimental and engineering practice is a set of many parallel, similar slits, called the *diffraction grating*.[49] If the slit's width is much smaller than period d of the grating, its transparency function may be approximated as

$$T(\boldsymbol{\rho}') \propto \sum_{n=-\infty}^{+\infty} \delta(x' - nd), \tag{8.124}$$

and Eq. (116) yields

$$f(\boldsymbol{\rho}) \propto \sum_{n=-\infty}^{n=+\infty} \exp\{-in\kappa_x d\} = \sum_{n=-\infty}^{n=+\infty} \exp\left\{-i\frac{nkxd}{z}\right\}. \tag{8.125}$$

This sum vanishes for all values of $\kappa_x d$ that are not multiples of 2π, so the result describes sharp intensity peaks at the following diffraction angles:

$$\theta_m \equiv \left(\frac{x}{z}\right)_m = \left(\frac{\kappa_x}{k}\right)_m = \frac{2\pi}{kd}m = \frac{\lambda}{d}m. \tag{8.126}$$

Taking into account that this result is only valid for small angles $|\theta_m| \ll 1$, it may be interpreted exactly as Eq. (59) – see Fig. 6a. However, in contrast with the interference (121) from two slits, the destructive interference from many slits kills the net wave as soon as the angle is even slightly different from each value (60). This is very convenient for spectroscopic purposes because the diffraction lines produced by multi-frequency waves do not overlap even if the frequencies of their adjacent components are very close.

Two unavoidable features of practical diffraction gratings make their properties different from this simple, ideal picture. First, the finite number N of slits, which may be described by limiting the sum (125) to the interval $n = [-N/2, +N/2]$, results in a non-zero spread, $\delta\theta/\theta \sim 1/N$, of each diffraction peak, and hence in the reduction of the grating's spectral resolution. (Unintentional variations of the inter-slit distance d have a similar effect, so before the advent of high-resolution photolithography, special high-precision mechanical tools had been used for grating fabrication.)

Second, a finite slit width w leads to the diffraction peak pattern modulation by the $\mathrm{sinc}^2(kw\theta/2)$ envelope, similar to the pattern shown in Fig. 15. Actually, for spectroscopic purposes, such modulation is sometimes a plus, because only one diffraction peak (say, with $m = \pm 1$) is practically used, and if the frequency spectrum of the analyzed wave is very broad (covers more than one octave), the higher peaks produce undesirable hindrance. Because of this reason, w is frequently selected to be equal exactly to $d/2$, thus suppressing each other diffraction maximum. Moreover, sometimes semi-transparent films are used to make the transparency function $T(\mathbf{r}')$ continuous and close to a sinusoidal one:

$$T(\boldsymbol{\rho}') \approx T_0 + T_1 \cos\frac{2\pi x'}{d} \equiv T_0 + \frac{T_1}{2}\left(\exp\left\{i\frac{2\pi x'}{d}\right\} + \exp\left\{-i\frac{2\pi x'}{d}\right\}\right). \tag{8.127}$$

Plugging the last expression into Eq. (116) and integrating, we see that the output wave consists of just 3 components: the direct-passing wave (proportional to T_0) and two diffracted waves (proportional to T_1) propagating in the directions of the two lowest Bragg angles, $\theta_{\pm 1} = \pm\lambda/d$.[50]

[49] The rudimentary diffraction grating effect, produced by the parallel fibers of a bird's feather, was discovered as early as 1673 by James Gregory (who also invented the "Gregorian" telescope – one of the basic designs for reflecting telescopes).

The same Eq. (116) may be also used to obtain one more general (and rather curious) result, called the *Babinet principle*.[51] Consider two experiments with the diffraction of similar plane waves on two "complementary" screens – such that together they would cover the whole plane, without a hole or an overlap. (Think, for example, about an opaque disk of radius R and a large opaque screen with a round orifice of the same radius.) Then, according to the Babinet principle, the diffracted wave patterns produced by these two screens in all directions with $\theta \neq 0$ are *identical*.

The proof of this principle is straightforward: since the transparency functions produced by the screens are complementary:

$$T(\boldsymbol{\rho}') \equiv T_1(\boldsymbol{\rho}') + T_2(\boldsymbol{\rho}') = 1, \tag{8.128}$$

and the diffracted wave is (in the Fraunhofer approximation only!) a Fourier transform of $T(\boldsymbol{\rho}')$, which is a linear operation, we get

$$f_1(\boldsymbol{\rho}) + f_2(\boldsymbol{\rho}) = f_0(\boldsymbol{\rho}), \tag{8.129}$$

where f_0 is the wave "scattered" by the composite screen with $T_0(\boldsymbol{\rho}') \equiv 1$, i.e. the unperturbed initial wave propagating in the initial direction ($\theta = 0$). In all other directions, $f_1 = -f_2$, i.e. the diffracted waves are indeed similar besides the difference in sign – which is equivalent to a phase shift by $\pm\pi$. However, it is important to remember that the Babinet principle notwithstanding, in real experiments, with screens at finite distances, the diffracted waves may interfere with the unperturbed plane wave $f_0(\boldsymbol{\rho})$, leading to different diffraction patterns in cases 1 and 2 – see, e.g., Fig. 14 and its discussion.

8.9. Magnetic dipole and electric quadrupole radiation

Throughout this chapter, we have seen how many important results may be obtained from Eq. (26) for the electric dipole radiation by a small-size source (Fig. 1). Only in rare cases when this radiation is absent, for example, if the dipole moment $\mathbf{p}$ of the source equals zero (or does not change in time – either at all or at the frequency of our interest), higher-order effects may be important. I will now discuss the main two of them, *quadrupole electric radiation* and *dipole magnetic radiation*.

In Sec. 2 above, the electric dipole radiation was calculated by plugging the expansion (19) into the exact formula (17b) for the retarded vector potential $\mathbf{A}(\mathbf{r}, t)$. Let us make a more exact calculation, by keeping the second term of that expansion as well:

$$\mathbf{j}\left(\mathbf{r}', t - \frac{R}{v}\right) \approx \mathbf{j}\left(\mathbf{r}', t - \frac{r}{v} + \frac{\mathbf{r}'\cdot\mathbf{n}}{v}\right) \equiv \mathbf{j}\left(\mathbf{r}', t' + \frac{\mathbf{r}'\cdot\mathbf{n}}{v}\right), \qquad \text{where } t' \equiv t - \frac{r}{v}. \tag{8.130}$$

Since the expansion is only valid if the last term in the time argument of $\mathbf{j}$ is relatively small, in the Taylor expansion of $\mathbf{j}$ with respect to that argument we may keep just two leading terms:

$$\mathbf{j}\left(\mathbf{r}', t' + \frac{\mathbf{r}'\cdot\mathbf{n}}{v}\right) \approx \mathbf{j}(\mathbf{r}', t') + \frac{\partial \mathbf{j}(\mathbf{r}', t')}{\partial t'}\frac{(\mathbf{r}'\cdot\mathbf{n})}{v}, \tag{8.131}$$

[50] Similar tricks are used in the so-called *phased-array antennas*, broadly used in radar systems and radioastronomy, in which electronically controlled mutual phase shifts of microwave signals feeding many similar component antennas are used to steer the direction of the resulting narrow beam. For more on this important technology, see, e.g. T. Milligan, *Modern Antenna Design,* 2nd ed., Wiley (2005).

[51] Named after Jacques Babinet (1784-1874) who made several important contributions to optics.

so Eq. (17b) yields $\mathbf{A} = \mathbf{A}_d + \mathbf{A}'$, where $\mathbf{A}_d$ is the electric dipole contribution as given by Eq. (23), and $\mathbf{A}'$ is the new term of the next order in the small parameter $r' << r$:

$$\mathbf{A}'(\mathbf{r},t) = \frac{\mu}{4\pi r v}\frac{\partial}{\partial t'}\int \mathbf{j}(\mathbf{r}',t')(\mathbf{r}'\cdot\mathbf{n})d^3r'. \tag{8.132}$$

Just as it was done in Sec. 2, let us evaluate this term for a system of non-relativistic particles with electric charges q_k and radius vectors $\mathbf{r}_k(t)$:

$$\mathbf{A}'(\mathbf{r},t) = \frac{\mu}{4\pi r v}\left[\frac{d}{dt}\sum_k q_k \dot{\mathbf{r}}_k(\mathbf{r}_k\cdot\mathbf{n})\right]_{t=t'}. \tag{8.133}$$

Using the "bac minus cab" identity of the vector algebra again,[52] the vector operand of Eq. (133) may be rewritten as

$$\begin{aligned}\dot{\mathbf{r}}_k(\mathbf{r}_k\cdot\mathbf{n}) &\equiv \frac{1}{2}\dot{\mathbf{r}}_k(\mathbf{r}_k\cdot\mathbf{n}) + \frac{1}{2}\dot{\mathbf{r}}_k(\mathbf{n}\cdot\mathbf{r}_k) = \frac{1}{2}(\mathbf{r}_k\times\dot{\mathbf{r}}_k)\times\mathbf{n} + \frac{1}{2}\mathbf{r}_k(\mathbf{n}\cdot\dot{\mathbf{r}}_k) + \frac{1}{2}\dot{\mathbf{r}}_k(\mathbf{n}\cdot\mathbf{r}_k)\\ &\equiv \frac{1}{2}(\mathbf{r}_k\times\dot{\mathbf{r}}_k)\times\mathbf{n} + \frac{1}{2}\frac{d}{dt}[\mathbf{r}_k(\mathbf{n}\cdot\mathbf{r}_k)],\end{aligned} \tag{8.134}$$

so the right-hand side of Eq. (133) may be represented as a sum of two terms, $\mathbf{A}' = \mathbf{A}_m + \mathbf{A}_q$, where

$$\mathbf{A}_m(\mathbf{r},t) = \frac{\mu}{4\pi r v}\dot{\mathbf{m}}(t')\times\mathbf{n} \equiv \frac{\mu}{4\pi r v}\dot{\mathbf{m}}\left(t-\frac{r}{v}\right)\times\mathbf{n}, \qquad \text{with } \mathbf{m}(t) \equiv \frac{1}{2}\sum_k \mathbf{r}_k(t)\times q_k\dot{\mathbf{r}}_k(t); \tag{8.135}$$

$$\mathbf{A}_q(\mathbf{r},t) = \frac{\mu}{8\pi r v}\left[\frac{d^2}{dt^2}\sum_k q_k\mathbf{r}_k(\mathbf{n}\cdot\mathbf{r}_k)\right]_{t=t'}. \tag{8.136}$$

Comparing the second of Eqs. (135) with Eq. (5.91), we see that $\mathbf{m}$ is just the total magnetic moment of the source. On the other hand, the first of Eqs. (135) is absolutely similar in structure to Eq. (23), with $\mathbf{p}$ replaced with $(\mathbf{m}\times\mathbf{n})/v$, so for the corresponding component of the magnetic field it gives (in the same approximation $r >> \lambda$) a result similar to Eq. (24):

Magnetic dipole radiation: field

$$\mathbf{B}_m(\mathbf{r},t) = \frac{\mu}{4\pi r v}\nabla\times\left[\dot{\mathbf{m}}\left(t-\frac{r}{v}\right)\times\mathbf{n}\right] = -\frac{\mu}{4\pi r v^2}\mathbf{n}\times\left[\ddot{\mathbf{m}}\left(t-\frac{r}{v}\right)\times\mathbf{n}\right]. \tag{8.137}$$

According to this expression, just as at the electric dipole radiation, the vector $\mathbf{B}$ is perpendicular to the vector $\mathbf{n}$, and its magnitude is also proportional to $\sin\Theta$, where Θ is now the angle between the direction toward the observation point and the second time derivative of the vector $\mathbf{m}$ – rather than $\mathbf{p}$:

$$B_m = \frac{\mu}{4\pi r v^2}\ddot{m}\left(t-\frac{r}{v}\right)\sin\Theta. \tag{8.138}$$

As a result, the intensity of this *magnetic dipole radiation* has a similar angular distribution:

Magnetic dipole radiation: power

$$S_r = ZH^2 = \frac{Z}{(4\pi v^2 r)^2}\left[\ddot{m}\left(t-\frac{r}{v}\right)\right]^2\sin^2\Theta \tag{8.139}$$

[52] If you still need it, see MA Eq. (7.5).

- cf. Eq. (26), besides the (generally) different meaning of the angle Θ.

Note, however, that this radiation is usually much weaker than its electric-dipole counterpart. For example, for a non-relativistic particle with electric charge q, moving on a trajectory of linear size ~a, the electric dipole moment is of the order of qa, while its magnetic moment scales as $qa^2\omega$, where ω is the motion frequency. As a result, the ratio of the magnetic and electric dipole radiation intensities is of the order of $(a\omega/v)^2$, i.e. the squared ratio of the particle's speed to the speed of the emitted waves – that has to be much smaller than 1 for our non-relativistic calculation to be valid.

The angular distribution of the *electric quadrupole radiation* described by Eq. (136) is more involved. To show this, let us add to $\mathbf{A}_q$ a vector parallel to $\mathbf{n}$ (i.e. directed along the wave's propagation), getting

$$\mathbf{A}_q(\mathbf{r},t) \to \frac{\mu}{24\pi r v}\ddot{\mathscr{Q}}\left(t-\frac{r}{v}\right), \quad \text{where } \mathscr{Q} \equiv \sum_k q_k\left\{3\mathbf{r}_k(\mathbf{n}\cdot\mathbf{r}_k)-\mathbf{n}r_k^2\right\}, \tag{8.140}$$

since this addition does not contribute to the transverse components of the electric and magnetic fields, i.e. to the radiated wave. According to the above definition of the vector $\mathscr{Q}$, its Cartesian components may be represented as

$$\mathscr{Q}_j = \sum_{j'=1}^{3}\mathscr{Q}_{jj'}n_{j'}, \tag{8.141}$$

where $\mathscr{Q}_{jj'}$ are the elements of the electric quadrupole tensor of the system – see the last of Eqs. (3.4):[53]

$$\mathscr{Q}_{jj'} = \sum_k q_k\left(3r_j r_{j'}-r^2\delta_{jj'}\right)_k . \tag{8.142}$$

Now taking the curl of the first of Eqs. (140) at $r >> \lambda$, we get

$$\mathbf{B}_q(\mathbf{r},t) = -\frac{\mu}{24\pi r v^2}\mathbf{n}\times\dddot{\mathscr{Q}}\left(t-\frac{r}{v}\right). \tag{8.143}$$

Electric quadrupole radiation: field

This expression is similar to Eqs. (24) and (137), but according to Eqs. (140) and (142), components of the vector $\mathscr{Q}$ do depend on the direction of the vector $\mathbf{n}$, leading to a different angular dependence of S_r.

As the simplest example, let us consider the system of two equal point electric charges moving symmetrically, at equal distances $d(t) << \lambda$ from a stationary center – see Fig. 16.

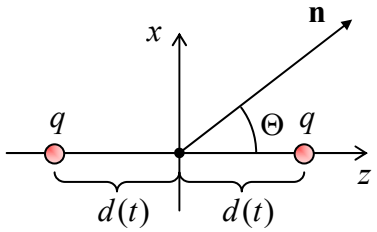


Fig. 8.16. The simplest system emitting electric quadrupole radiation.

Due to the symmetry of the system, its dipole moments $\mathbf{p}$ and $\mathbf{m}$ (and hence its electric and magnetic dipole radiation) vanish, but the quadrupole tensor (142) still has non-zero elements. With the coordinate choice shown in Fig. 16, these elements are diagonal:

[53] Let me hope that the reader has already acquired some experience in the calculation of this tensor's elements – e.g., for the simple systems specified in Problems 3.2-3.4.

$$\mathcal{Q}_{xx} = \mathcal{Q}_{yy} = -2qd^2, \qquad \mathcal{Q}_{zz} = 4qd^2. \tag{8.144}$$

With the x-axis selected within the common plane of the z-axis and the direction $\mathbf{n}$ toward the observation point (Fig. 16), so $n_x = \sin\Theta$, $n_y = 0$, and $n_z = \cos\Theta$, Eq. (141) yields

$$\mathcal{Q}_x = -2qd^2 \sin\Theta, \qquad \mathcal{Q}_y = 0, \qquad \mathcal{Q}_z = 4qd^2 \cos\Theta, \tag{8.145}$$

and the vector product in Eq. (143) has only one non-vanishing Cartesian component:

$$\left(\mathbf{n} \times \dddot{\mathcal{Q}}\right)_y = n_z \dddot{\mathcal{Q}}_x - n_x \dddot{\mathcal{Q}}_z = -6q \sin\Theta \cos\Theta \frac{d^3}{dt^3}\left[d^2(t)\right]. \tag{8.146}$$

As a result, the quadrupole radiation intensity, $S \propto B_q^2$, is proportional to $\sin^2\Theta\cos^2\Theta$, i.e. vanishes not only along the symmetry axis of the system (as the electric-dipole and the magnetic-dipole radiations would), but also in all directions perpendicular to this axis, reaching its maxima at $\Theta = \pm\pi/4$.

For more complex systems, the angular distribution of the electric quadrupole radiation may be different, but it may be proved that its total (instant) power always obeys the following simple formula:

Electric quadrupole radiation: power

$$\mathcal{P}_q = \frac{Z}{720\pi v^4} \sum_{j,j'=1}^{3} \left(\dddot{\mathcal{Q}}_{jj'}\right)^2. \tag{8.147}$$

Let me finish this section by giving, also without proof, one more fact important for some applications: due to their different spatial structure, the magnetic-dipole and electric-quadrupole radiation fields do not interfere, i.e. the total power of radiation (neglecting the electric-dipole and higher multipole terms) may be found as the sum of these components, calculated independently. On the contrary, the electric-dipole and magnetic-dipole radiations of the same system typically interfere coherently, so their radiation fields (rather than powers) should be summed up.

8.10. Exercise problems

8.1. Equation (8) obviously has standing-wave solutions $\chi(r, t) = \mathrm{Re}\,[C\sin kr \exp\{-i\omega t\}]$, turning the scalar potential $\phi = \chi/r$ into a finite constant at $r = 0$ and into zero at $kr = \pi n$, with $n = 0, 1, 2, \ldots$ This fact seems to imply that a cavity of radius R, carved inside a good conductor, has resonant modes with a purely radial electric field $\mathbf{E}(\mathbf{r}) = \mathbf{n}_r E(r)$ and that the lowest nonvanishing of them, with $k = \pi/R$, gives the lowest (fundamental) frequency $\omega \equiv vk = \pi(v/R)$ of the cavity. Is this conclusion correct?

8.2. Simplify the Lorentz reciprocity theorem (6.121) for space-localized field sources. Then find out what it says about the fields of two compact, well-separated sources of electric-dipole radiation.

8.3. In the electric-dipole approximation, calculate the angular distribution and the total power of electromagnetic radiation by the hydrogen atom within the following classical model: an electron rotates, at a constant distance R, about a much heavier proton. Use this result to calculate the law of a gradual reduction of R in time. Finally, evaluate the classical lifetime of the atom by borrowing the initial value of R from quantum mechanics: $R(0) = r_B \approx 0.53\times10^{-10}$ m.

8.4. A non-relativistic particle of mass m, with electric charge q, is placed into a time-independent uniform magnetic field $\mathbf{B}$. Derive the law of decrease of the particle's kinetic energy due to

its electromagnetic radiation at the *cyclotron frequency* $\omega_c = qB/m$. Evaluate the rate of such *radiation cooling* of electrons in a magnetic field of 1 T, and estimate the energy interval in which this result is quantitatively correct.

Hint: The cyclotron motion will be discussed in detail (for arbitrary particle velocities) in Sec. 9.6 below, but I hope that the reader already knows that in the non-relativistic case ($v << c$), the above formula for ω_c may be readily obtained by combining the 2nd Newton law $mv_\perp^2/R = qv_\perp B$ for the particle's circular rotation under the effect of the magnetic component of the Lorentz force (5.10), and the geometric relation $v_\perp = R\omega_c$. (Here $\mathbf{v}_\perp$ is the particle's velocity in the plane normal to the vector **B**.)

8.5. A particle with mass m, electric charge q, and an initial kinetic energy $T << mc^2$ collides head-on with a much more massive particle of charge $\mathcal{F}q$, in free space. Calculate the total energy of electromagnetic radiation during this collision, assuming it to be much lower than T.

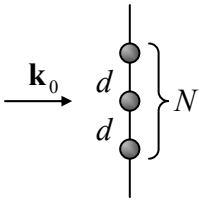
8.6. Solve the dipole antenna radiation problem discussed in Sec. 2 (see Fig. 3) for the optimal value $l = \lambda/2$ of its length, assuming that the current distribution in each of its arms is sinusoidal: $I(z, t) = I_0\cos(\pi z/l)\cos\omega t$. [54]

8.7. A plane wave is scattered by a localized object in free space. Relate the differential cross-section of the wave's scattering to the average force it exerts on the object. Use this general relation to calculate the force exerted by a plane monochromatic wave on a free non-relativistic particle and compare the result with those obtained in Problems 7.4 and 7.5.

8.8. Use the Lorentz oscillator model of a bound charge, given by Eq. (7.30), to explore the transition between the two scattering limits discussed in Sec. 3 and, in particular, the *resonant scattering* taking place at $\omega \approx \omega_0$. In the last context, discuss the contribution of scattering to the oscillator's damping.

8.9.* A sphere of radius R, made of a material with a uniform permanent electric polarization $\mathbf{P}_0$ and a constant mass density ρ, is free to rotate about its center. Calculate its average total cross-section for scattering of a linearly polarized plane electromagnetic wave of frequency $\omega << R/c$, incident from free space, in the weak-wave limit, assuming that the initial orientation of the polarization vector $\mathbf{P}_0$ is random.

8.10. Use Eq. (56) to analyze the interference/diffraction pattern produced by a plane wave's scattering on a set of N similar equidistant small objects on a straight line normal to the direction of the incident wave's propagation – see the figure on the right. Discuss the trend(s) of the pattern in the limit $N \to \infty$.



8.11. Use the Born approximation to calculate the differential cross-section of a plane wave's scattering by a uniform dielectric sphere of an arbitrary radius R. In the limits $kR << 1$ and $1 << kR$ (where k is the wave number), analyze the angular dependence of the differential cross-section and calculate the total cross-section of scattering.

[54] As was emphasized in Sec. 2, this is a reasonable guess rather than a controllable approximation. The exact (rather involved!) theory shows that this assumption gives errors ~5%, depending on the wire's diameter.

8.12. A sphere of radius R is made of a uniform dielectric material, with an arbitrary dielectric constant. Calculate its total cross-section of scattering a linearly-polarized low-frequency ($k << 1/R$) wave and compare the result with the solution of the previous problem.

8.13. Use the Born approximation to calculate the differential cross-section of a plane wave's scattering on a right circular cylinder of length l and radius R, for an arbitrary angle of incidence.

8.14. Formulate the quantitative condition of the Born approximation's validity for a uniform dielectric scatterer, with all linear dimensions of the order of the same scale a.

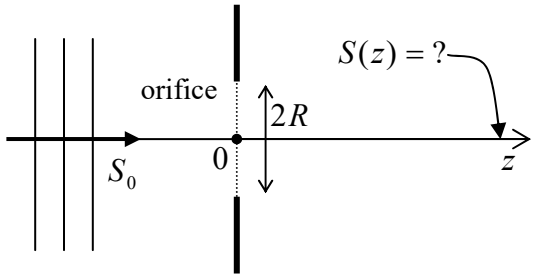
8.15. If a scatterer absorbs some part of the incident wave's power, it may be characterized by an *absorption cross-section* σ_a defined similarly to Eq. (39) for the scattering cross-section:

$$\sigma_a \equiv \frac{\overline{\mathscr{P}_a}}{|E_\omega|^2 / 2Z_0},$$

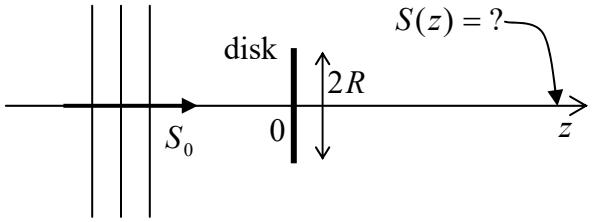
where the numerator is the time-averaged absorbed power. Use two different approaches to calculate σ_a of a very small sphere of radius $R << k^{-1}, \delta_s$, made of a nonmagnetic material with an Ohmic conductivity σ and the high-frequency permittivity $\varepsilon_{opt} = \varepsilon_0$. Can σ_a of such a sphere be larger than its geometric cross-section πR^2?

8.16. Use the Huygens principle to calculate the wave's intensity on the symmetry plane of the slit diffraction experiment (i.e. at $x = 0$ in Fig. 12), for arbitrary ratio z/ka^2.

8.17. A plane wave with wavelength λ is normally incident on an opaque planar screen with a round orifice of radius $R >> \lambda$. Use the Huygens principle to calculate the passing wave's intensity on the system's symmetry axis, at distances $z >> R$ from the screen (see the figure on the right), and analyze the result.



8.18. A plane monochromatic wave is now normally incident on an opaque circular disk of radius $R >> \lambda$. Use the Huygens principle to calculate the wave's intensity at a distance $z >> R$ behind the disk's center – see the figure on the right. Discuss the result.

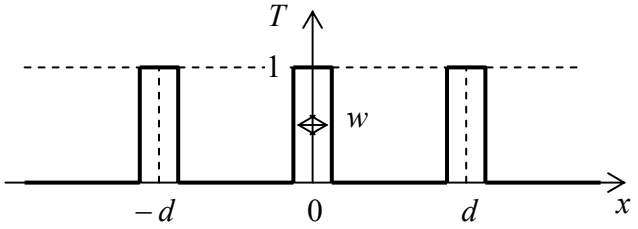


8.19. Use the Huygens principle to analyze the Fraunhofer diffraction of a plane wave normally incident on a square-shaped hole, of size $a \times a$, in an opaque screen. Sketch the diffraction pattern you would observe at a sufficiently large distance, and quantify the expression "sufficiently large" for this case.

8.20. Use the Huygens principle to analyze the propagation of a monochromatic Gaussian beam described by Eq. (7.181), with the initial characteristic width $a_0 >> \lambda$, in a uniform isotropic medium. Use the result for a semi-quantitative derivation of the so-called *Abbe limit* for the spatial resolution of

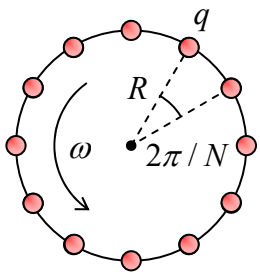
an optical system: $w_{min} = \lambda/2\sin\theta$, where θ is the half-angle of the wave cone propagating from the object and captured by the system.

8.21. Within the Fraunhofer approximation, analyze the pattern produced by a diffraction grating with the 1D-periodic transparency profile shown in the figure on the right, for the normal incidence of a monochromatic plane wave.



8.22. N equal point charges are attached, at equal intervals, to a circle rotating with a constant angular velocity about its center – see the figure on the right. For what values of N does the system emit:

(i) the electric dipole radiation?
(ii) the magnetic dipole radiation?
(iii) the electric quadrupole radiation?



8.23. What general statements can you make about:

(i) the electric dipole radiation, and
(ii) the magnetic dipole radiation,

due to a collision of an arbitrary number of similar non-relativistic classical particles?

8.24. Calculate the angular distribution and the total power radiated by a small planar loop antenna of radius R, fed with ac current with frequency ω and amplitude I_0, into free space.

8.25. The orientation of a magnetic dipole, with a constant magnitude m of its moment, is rotating about a certain axis with an angular velocity ω, with the angle α between them staying constant. Calculate the angular distribution and the average power of its radiation into the free space.

8.26. Solve Problem 12 (also in the low-frequency limit $kR << 1$), for the case when the sphere's material has a frequency-independent Ohmic conductivity σ, and $\varepsilon_{opt} = \varepsilon_0$, in two limits:

(i) of a very large skin depth ($\delta_s >> R$), and
(ii) of a very small skin depth ($\delta_s << R$).

8.27. Complete the solution of the problem started in Sec. 9, by calculating the full power of radiation of the system of two charges oscillating in antiphase along the same straight line – see Fig. 16. Also, calculate the average radiation power for the case of harmonic oscillations, $d(t) = a\cos\omega t$, compare it with the case of a single charge performing similar oscillations, and interpret the difference.

8.28. The system of four alternating charges located at the angles of a square, considered in Problem 3.3(i), is now being rotated around the axis normal to their plane and passing through the square's center, with a constant angular frequency $\omega << v/a$. Calculate the time-averaged angular distribution and the total power of the resulting radiation.

Chapter 9. Special Relativity

This chapter starts with a review of special relativity's basics, including its very convenient 4-vector formalism. This background is then used for the analysis of the relation between the electromagnetic field's values measured in different inertial reference frames moving relative to each other. The results enable us to discuss relativistic particle dynamics in the electric and magnetic fields, and the analytical mechanics of the particles – and of the electromagnetic field as such.

9.1. Einstein postulates and the Lorentz transform

As was emphasized at the derivation of expressions for the dipole and quadrupole radiation in the last chapter, they are only valid for systems of non-relativistic particles moving with velocities **u** much lower than *c*. In order to generalize these results to particles moving with arbitrary **u**, we need help from the relativity theory. Moreover, an analysis of the motion of charged relativistic particles in electric and magnetic fields is also a natural part of electrodynamics. This is why I will follow the tradition of using this course for a (by necessity, brief) introduction to the special relativity theory. This theory is based on the fundamental idea that measurements of physical variables (including the spatial and even temporal intervals between two events) may give different results in different reference frames, in particular in two inertial frames moving relative to each other translationally (i.e. without rotation), with a certain constant velocity **v** (Fig. 1).

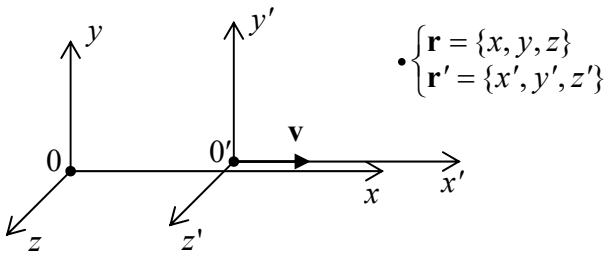


Fig. 9.1. Translational mutual motion of two reference frames.

In the non-relativistic (Newtonian) mechanics the problem of transfer between such reference frames has a simple solution at least in the limit $v << c$, because the basic equation of particle dynamics (the 2nd Newton law) [1]

$$m_k \ddot{\mathbf{r}}_k = -\nabla_k \sum_{k'} U(\mathbf{r}_k - \mathbf{r}_{k'}), \tag{9.1}$$

where U is the potential energy of inter-particle interactions, is invariant with respect to the so-called *Galilean transformation* (or just "transform" for short).[2] Choosing the coordinates in both frames so that their axes x and x' are parallel to the vector **v** (as in Fig. 1), the transform may be represented as

[1] Let me hope that the reader does not need a reminder that for Eq. (1) to be valid, the reference frames 0 and 0′ have to be inertial – see, e.g., CM Sec. 1.2.

[2] It had been first formulated by Galileo Galilei, if only rather informally, as early as 1638 – four years before Isaac Newton was *born*! Note also the very unfortunate term "boost" used sometimes to describe such translational transformations. (It is especially unnatural in the special relativity, not describing accelerations.) In my course, this term is avoided, with the equivalent "transform" used instead.

$$x = x' + vt', \qquad y = y', \qquad z = z', \qquad t = t', \tag{9.2a}$$

Galilean transform

and plugging Eq. (2a) into Eq. (1), we get an absolutely similarly looking equation of motion in the "moving" reference frame 0*'*. Since the reciprocal transform,

$$x' = x - vt, \qquad y = y', \qquad z' = z, \qquad t' = t, \tag{9.2b}$$

is similar to the direct one, with the replacement of $(+v)$ with $(-v)$, we may say that the Galilean invariance means that there is no "master" (*absolute*) spatial reference frame in classical mechanics, although the spatial and temporal intervals between different instant events are absolute, i.e. reference-frame invariant: $\Delta x = \Delta x'$,…, $\Delta t = \Delta t'$.

However, it is straightforward to use Eq. (2) to check that the form of the wave equation

$$\left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2} - \frac{1}{c^2} \frac{\partial^2}{\partial t^2} \right) f = 0, \tag{9.3}$$

describing, in particular, the electromagnetic wave propagation in free space,[3] is *not* Galilean-invariant.[4] For the "usual" (say, elastic) waves, which obey a similar equation albeit with a different speed,[5] this lack of Galilean invariance is natural and is compatible with the invariance of Eq. (1), from which the wave equation originates. This is because the elastic waves are essentially the oscillations of interacting particles of a certain medium (e.g., an elastic solid), making the reference frame connected to this medium, special. So, if the electromagnetic waves were oscillations of a certain special medium (which was first called the "luminiferous aether"[6] and later *aether* – or just "ether"), similar arguments might be applicable to reconcile Eqs. (2) and (3).

The detection of such a medium was the goal of the measurements carried out between 1881 and 1887 (with better and better precision) by Albert Abraham Michelson and Edward Williams Morley, which are sometimes called "the most famous failed experiments in physics". Figure 2 shows a crude scheme of these experiments.

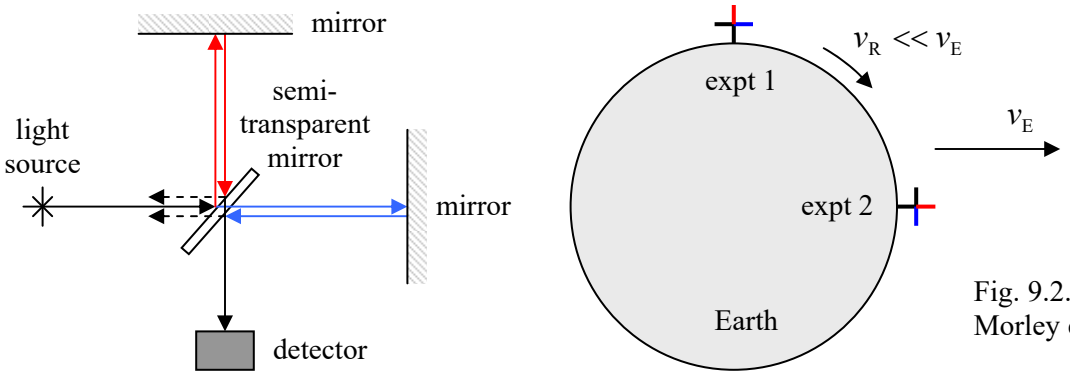


Fig. 9.2. The Michelson-Morley experiment.

[3] The discussions in this chapter and most of the next chapter will be restricted to the free-space (and hence dispersion-free) case; some media effects on the radiation by relativistic particles will be discussed in Sec.10.4.

[4] It is interesting that the usual (non-relativistic) Schrödinger equation, whose fundamental solution for a free particle is a similar monochromatic wave (albeit with a different dispersion law), *is* Galilean-invariant, with a certain change of the wavefunction's phase – see, e.g., QM Chapter 1.

[5] See, e.g., CM Secs. 6.5 and 7.7.

[6] In ancient Greek mythology, aether is the clean air breathed by the gods residing on Mount Olympus.

A nearly monochromatic wave from a light source is split into two parts (optimally, of equal intensity), using a semi-transparent mirror tilted by the angle $\pi/4$ to the incident wave direction. These two partial waves are reflected back by two fully-reflecting mirrors and arrive at the same semi-transparent mirror again. Here half of each wave is directed toward the light source (they vanish there without affecting the source), but another half is passed toward an intensity detector, forming, with its counterpart, an interference pattern similar to that in the Young experiment. Thus each of the interfering waves has traveled twice (back and forth) each of two mutually perpendicular "arms" of the interferometer. Assuming that the aether, in which light propagates with speed c, moves with speed $v < c$ along one of the arms, of length l_l, it is straightforward (and hence left for the reader's exercise :-) to get the following expression for the difference between the light roundtrip times:

$$\Delta t = \frac{2}{c}\left[\frac{l_t}{\left(1-v^2/c^2\right)^{1/2}} - \frac{l_l}{1-v^2/c^2}\right] \approx \frac{l}{c}\left(\frac{v}{c}\right)^2, \tag{9.4}$$

where l_t is the length of the second, "transverse" arm of the interferometer (perpendicular to $\mathbf{v}$), and the last, approximate expression is valid at $l_t \approx l_l \equiv l$ and $v << c$.

Since the Earth moves around the Sun with a speed $v_E \approx 30$ km/s $\approx 10^{-4}\ c$, the arm positions relative to this motion alternate, due to the Earth's rotation about its axis, every 6 hours – see the right panel of Fig. 2. Hence if we assume that the aether rests in the Sun's reference frame, then Δt (and the corresponding shift of the interference fringes), has to change its sign with this half-period as well. The same alternation may be achieved, at a smaller time scale, by a deliberate rotation of the instrument by $\pi/2$. In the most precise version of the Michelson-Morley experiment (circa 1887), this shift was expected to be close to 0.4 of the interference pattern period. The results of the search for such a shift were negative, with the error bar about 0.01 of the period.[7]

The most prominent immediate explanation for this zero result[8] was suggested in 1889 by George Francis FitzGerald and (independently and more qualitatively) by H. Lorentz in 1892: as evident from Eq. (4), if the longitudinal arm of the interferometer itself experiences the so-called *length contraction*:

$$l_l(v) = l_l(0)\left(1-\frac{v^2}{c^2}\right)^{1/2}, \tag{9.5}$$

while the transverse arm's length is not affected by its motion through the aether, this effect kills the shift Δt. This radical idea received strong support from the proof, in 1887-1905, that the Maxwell equations, and hence the wave equation (3), are form-invariant under the so-called *Lorentz transform*,[9] which in particular describes Eq. (5). For the choice of coordinates shown in Fig. 1, the transform reads

[7] Through the 20th century, the Michelson-Morley-type experiments were repeated using more and more refined experimental techniques, always with zero results for the apparent aether motion speed. For example, recent experiments using cryogenically cooled optical resonators have reduced the upper limit for such speed to just $3\times10^{-15}\ c$ –see H. Müller *et al*., *Phys. Rev. Lett*. **91**, 020401 (2003).

[8] The zero result of a slightly later experiment, namely a precise measurement of the torque that should be exerted by the moving aether on a charged capacitor, carried out in 1903 by F. Trouton and H. Noble (following G. FitzGerald's suggestion), seconded the Michelson and Morley's conclusions.

[9] The theoretical work toward this result included important contributions by Woldemart Voigt (in 1887), Hendrik Lorentz (in 1892-1904), Joseph Larmor (in 1897 and 1900), and Henri Poincaré (in 1900 and 1905).

$$x = \frac{x' + vt'}{\left(1 - v^2/c^2\right)^{1/2}}, \quad y = y', \quad z = z', \quad t = \frac{t' + (v/c^2)x'}{\left(1 - v^2/c^2\right)^{1/2}}. \tag{9.6a}$$

Lorentz transform

It is elementary to solve these equations for the primed coordinates to get the reciprocal transform

$$x' = \frac{x - vt}{\left(1 - v^2/c^2\right)^{1/2}}, \quad y' = y, \quad z' = z, \quad t' = \frac{t - (v/c^2)x}{\left(1 - v^2/c^2\right)^{1/2}}. \tag{9.6b}$$

(I will soon represent Eqs. (6) in a more elegant form – see Eqs. (19) below.)

The Lorentz transform relations (6) are evidently reduced to the Galilean transform formulas (2) at $v^2 << c^2$. However, all attempts to give a reasonable interpretation of these equalities while keeping the notion of the aether have failed, in particular because of the restrictions imposed by results of earlier experiments carried out in 1851 and 1853 by Hippolyte Fizeau – which were repeated with higher accuracy by the same Michelson and Morley in 1886. These experiments have shown that if one sticks to the aether concept, this hypothetical medium has to be partially "dragged" by any moving dielectric material with a speed proportional to $(\kappa - 1)$. Such local drag would be irreconcilable with the assumed continuity of the aether.

In his famous 1905 paper, Albert Einstein suggested a bold resolution of this contradiction, essentially removing the concept of the aether altogether.[10] Moreover, he argued that the Lorentz transform is the general property of time and space, rather than of the electromagnetic field alone. He started with two postulates, the first one essentially repeating the relativity principle formulated a bit earlier (in 1904) by H. Poincaré in the following form:

> "*…the laws of physical phenomena should be the same, whether for an observer fixed or for an observer carried along in a uniform movement of translation; so that we have not and could not have any means of discerning whether or not we are carried along in such a motion.*"[11]

The second Einstein postulate was that the speed of light c, in free space, should be constant in all reference frames. (This is essentially a denial of the aether's existence.)

Then, Einstein showed that the Lorenz transform relations (6) naturally follow from his postulates, with a few (very natural) additional assumptions. Let a point source emit a short flash of light, at the moment $t = t' = 0$ when the origins of the reference frames shown in Fig. 1 coincide. Then, according to the second of Einstein's postulates, in each of the frames, the spherical wave propagates with the same speed c, i.e. the coordinates of points of its front, measured in the two frames, have to obey the following equalities:

$$\begin{aligned} (ct)^2 - (x^2 + y^2 + z^2) &= 0, \\ (ct')^2 - (x'^2 + y'^2 + z'^2) &= 0. \end{aligned} \tag{9.7}$$

[10] In hindsight, this was much relief, because the aether had been a very awkward construct to start with. In particular, according to the basic theory of elasticity (see, e.g., CM Ch. 7), in order to carry such transverse waves as the electromagnetic ones, this medium would need to have a non-zero shear modulus, i.e. behave as an elastic solid – rather than as a rarified gas hypothesized initially by C. Huygens.

[11] Note that though the relativity principle excludes the notion of the special ("absolute") spatial reference frame, its quoted verbal formulation still leaves the possibility of the Galilean "absolute time" $t = t'$ open. The quantitative relativity theory kills this option – see Eqs. (6) and their discussion below.

What may be the general relation between the combinations in the left-hand side of these equations – not for this particular wave's front, but in general? A very natural (essentially, the only justifiable) choice is

$$\left[(ct)^2-(x^2+y^2+z^2)\right]=f(v^2)\left[(ct')^2-(x'^2+y'^2+z'^2)\right]. \tag{9.8}$$

Now, according to the first postulate, the same relation should be valid if we swap the reference frames ($x \leftrightarrow x'$, etc.) and replace v with ($-v$). This is only possible if $f^2 = 1$, so excluding the option $f = -1$ (which is incompatible with the Galilean transform in the limit $v/c \to 0$), we are left with $f = +1$, i.e.

$$(ct)^2-(x^2+y^2+z^2)=(ct')^2-(x'^2+y'^2+z'^2). \tag{9.9}$$

For the line with $y = y' = 0$ and $z = z' = 0$, Eq. (9) is reduced to

$$(ct)^2-x^2=(ct')^2-x'^2. \tag{9.10}$$

It is very illuminating to interpret this relation as the one resulting from a mutual rotation of the reference frames (that now have to include clocks to measure time) on the plane of the coordinate x and the so-called *imaginary time* $\tau \equiv ict$ – see Fig. 3.

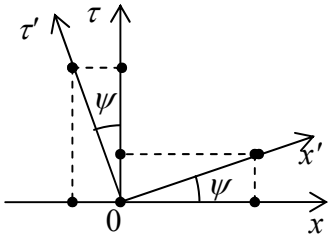


Fig. 9.3. The Lorentz transform as a mutual rotation of two reference frames on the [x, τ] plane.

Indeed, rewriting Eq. (10) as

$$\tau^2+x^2=\tau'^2+x'^2, \tag{9.11}$$

we may consider it as the invariance of the squared radius at the rotation shown in Fig. 3 and described by the following geometric relations:

$$\begin{aligned} x &= x'\cos\psi-\tau'\sin\psi, \\ \tau &= x'\sin\psi+\tau'\cos\psi, \end{aligned} \tag{9.12a}$$

with the reciprocal relations

$$\begin{aligned} x' &= x\cos\psi+\tau\sin\psi, \\ \tau' &= -x\sin\psi+\tau\cos\psi. \end{aligned} \tag{9.12b}$$

So far, the angle ψ has been arbitrary. In the spirit of Eq. (8), a natural choice is $\psi = \psi(v)$, with the requirement $\psi(0) = 0$. To find this function, let us write the definition of the velocity v of frame 0', as measured in frame 0 (which was implied above): for $x' = 0$, $x = vt$. In the variables x and τ, this means

$$\frac{x}{\tau}\Big|_{x'=0} \equiv \frac{x}{ict}\Big|_{x'=0} = \frac{v}{ic}. \tag{9.13}$$

On the other hand, for the same point $x' = 0$, Eqs. (12a) yield

$$\frac{x}{\tau}\Big|_{x'=0} = -\tan\psi\,. \tag{9.14}$$

These two expressions are compatible only if

$$\tan\psi = \frac{iv}{c}, \tag{9.15}$$

so

$$\sin\psi \equiv \frac{\tan\psi}{\left(1+\tan^2\psi\right)^{1/2}} = \frac{iv/c}{\left(1-v^2/c^2\right)^{1/2}} \equiv i\beta\gamma, \qquad \cos\psi \equiv \frac{1}{\left(1+\tan^2\psi\right)^{1/2}} = \frac{1}{\left(1-v^2/c^2\right)^{1/2}} \equiv \gamma, \tag{9.16}$$

where β and γ are two very convenient and commonly used dimensionless parameters defined as

$$\boldsymbol{\beta} \equiv \frac{\mathbf{v}}{c}, \qquad \gamma \equiv \frac{1}{\left(1-v^2/c^2\right)^{1/2}} \equiv \frac{1}{\left(1-\beta^2\right)^{1/2}}. \tag{9.17}$$

Relativistic parameters $\boldsymbol{\beta}$ and γ

(The vector $\boldsymbol{\beta}$ is called the *normalized velocity*, while the scalar γ is the *Lorentz factor*.)[12]

Using the above relations for ψ, Eqs. (12) become

$$x = \gamma\left(x' - i\beta\tau'\right), \qquad \tau = \gamma\left(i\beta x' + \tau'\right), \tag{9.18a}$$

$$x' = \gamma\left(x + i\beta\tau\right), \qquad \tau' = \gamma\left(-i\beta x + \tau\right). \tag{9.18b}$$

Now returning to the real variables [x, ct], we get the Lorentz transform relations (6), in a more compact form:

$$x = \gamma\left(x' + \beta\, ct'\right), \quad y = y', \quad z = z', \quad ct = \gamma\left(ct' + \beta\, x'\right), \tag{9.19a}$$

$$x' = \gamma\left(x - \beta\, ct\right), \quad y' = y, \quad z' = z, \quad ct' = \gamma\left(ct - \beta\, x\right). \tag{9.19b}$$

Lorentz transform – again

An immediate corollary of Eqs. (19) is that for γ to stay real, we need $v^2 \le c^2$, i.e. that the speed of any physical body (to which we could connect a meaningful reference frame) cannot exceed the speed of light, as measured in *any* other meaningful reference frame.[13]

9.2. Relativistic kinematic effects

Before proceeding to other corollaries of Eqs. (19), let us spend a few minutes discussing what these relations actually mean. Evidently, they are trying to tell us that the spatial and temporal intervals are not absolute (as they are in the Newtonian space), but do depend on the reference frame they are measured in. So, we have to understand very clearly what exactly may be measured – and thus may be discussed in a meaningful physics theory. Recognizing this necessity, A. Einstein introduced the notion of numerous imaginary *observers* that may be distributed all over each reference frame. Each observer

[12] Note the following identities: $\gamma^2 \equiv 1/(1-\beta^2)$ and $(\gamma^2 - 1) \equiv \beta^2/(1-\beta^2) \equiv \gamma^2\beta^2$, which are frequently handy in relativity-related algebra. One more function of β, the *rapidity* $\varphi \equiv \tanh^{-1}\beta$ (so that $\psi = i\varphi$), is also useful for some calculations.

[13] All attempts to rationally conjecture particles moving with $v > c$ (called *tachyons*) have failed – so far, at least. Possibly the strongest objection against their existence is the fact that the tachyons could be used to communicate back in time, thus violating the causality principle – see, e.g., G. Benford *et al.*, *Phys. Rev. D* **2**, 263 (1970).

has a clock and may use it to measure the instants of *local* events, taking place at the observer's location. He also conjectured, very reasonably, that:

(i) all observers within the same reference frame may agree on a common length measure ("a scale"), i.e. on their relative positions in that frame, and synchronize their clocks,[14] and

(ii) the observers belonging to different reference frames may agree on the nomenclature of *world events* (e.g., short flashes of light) to which their respective measurements refer.

Actually, these additional postulates have been already implied in our "derivation" of the Lorentz transform in Sec. 1. For example, by the set $\{x, y, z, t\}$ we mean the results of space and time measurements of a certain world event, about that all observers belonging to frame 0 agree. Similarly, all observers of frame 0' have to agree about the results $\{x', y', z', t'\}$. Finally, when the origin of frame 0' passes by some sequential points x_k of frame 0, the observers in the latter frame may measure its passage times t_k without a fundamental error, and know that all these times belong to $x' = 0$.

Now we can analyze the major corollaries of the Lorentz transform, which are rather striking from the point of view of our everyday (rather non-relativistic) experience.

(i) Length contraction. Let us consider a thin rigid rod oriented along the x-axis, with its length $l \equiv x_2 - x_1$, where $x_{1,2}$ are the coordinates of the rod's ends, as measured in its rest frame 0, at any instant t (Fig. 4). What would be the rod's length l' measured by the Einstein observers in the moving frame 0'?

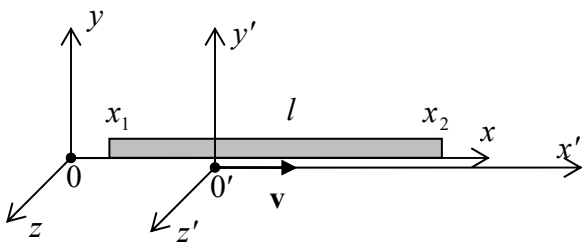


Fig. 9.4. The relativistic length contraction.

At a time instant t' agreed upon in advance, the observers who find themselves exactly at the rod's ends, may register that fact, and then subtract their coordinates $x'_{1,2}$ to calculate the apparent rod length $l' \equiv x_2' - x_1'$ in the moving frame. According to Eq. (19a), l may be expressed via this l' as

$$l \equiv x_2 - x_1 = \gamma(x_2' + \beta c t') - \gamma(x_1' + \beta c t') = \gamma(x_2' - x_1') \equiv \gamma l' . \tag{9.20a}$$

Hence, the rod's length, as measured in the *moving* reference frame is

Length contraction

$$l' = \frac{l}{\gamma} = l\left(1 - \frac{v^2}{c^2}\right)^{1/2} \le l , \tag{9.20b}$$

in accordance with the FitzGerald-Lorentz hypothesis (5). This is the *relativistic length contraction* effect: an object is always the longest (has the so-called *proper length* l) if measured in its *rest frame*.

[14] A posteriori, the Lorenz transform may be used to show that consensus-creating procedures (such as clock synchronization) are indeed possible. The basic idea of the proof is that since at $v << c$, the relativistic corrections to space and time intervals are of the order of $(v/c)^2$, they have negligible effects on clocks being brought together into the same point for synchronization slowly, with a speed $u << c$. The reader interested in a detailed discussion of this and other fine points of special relativity may be referred to, e.g., either H. Arzeliès, *Relativistic Kinematics*, Pergamon, 1966, or W. Rindler, *Introduction to Special Relativity*, 2nd ed., Oxford U. Press, 1991.

Note that according to Eqs. (19), the length contraction takes place only in the direction of the relative motion of two reference frames. As was noted in Sec. 1, this result immediately explains the zero result of the Michelson-Morley-type experiments, so they give very convincing evidence (if not irrefutable proof) of Eqs. (18)-(19).

(ii) Time dilation. Now let us use Eqs. (19a) to find the time interval Δt, as measured in some reference frame 0, between two world events – say, two ticks of a clock moving with another frame 0' (Fig. 5), i.e. having fixed values of x', y', and z'.

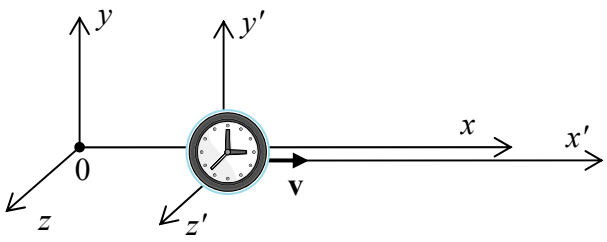


Fig. 9.5. The relativistic time dilation.

Let the time interval between these two events, measured in the clock's rest frame 0', be $\Delta t' \equiv t_2' - t_1'$. At these two moments, the clock would fly by two Einstein's observers at rest in frame 0, so they can record the corresponding moments $t_{1,2}$ shown by their clocks, and then calculate Δt as their difference. According to the last of Eqs. (19a),

$$c\Delta t \equiv ct_2 - ct_1 = \gamma\left[(ct_2' + \beta x') - (ct_1' + \beta x')\right] \equiv \gamma c\Delta t', \tag{9.21a}$$

so, finally,

$$\Delta t = \gamma \Delta t' \equiv \frac{\Delta t'}{\left(1 - v^2/c^2\right)^{1/2}} \geq \Delta t'. \tag{9.21b}$$

Time dilation

This is the famous *relativistic time dilation* (or "dilatation") effect: a time interval is *longer* if measured in a frame (in our case, frame 0) *moving relative to the clock*, while that in the clock's rest frame is the shortest possible – the so-called *proper time interval*.

This rather counter-intuitive effect is the everyday reality in experiments with high-energy elementary particles. For example, in a typical (and by no means record-breaking) experiment carried out in Fermilab, a beam of charged 200 GeV pions with $\gamma \approx 1{,}400$ traveled a distance of $l = 300$ m with the measured loss of only 3% of the initial beam intensity due to the pion decay (mostly, into muon-neutrino pairs) with the proper lifetime $t_0 \approx 2.56\times10^{-8}$ s. Without the time dilation, only an $\exp\{-l/ct_0\} \sim 10^{-17}$ fraction of the initial pions would survive, while the relativity-corrected number, $\exp\{-l/ct\} = \exp\{-l/c\gamma t_0\} \approx 0.97$, was in full accordance with experimental measurements.

As another example, the global positioning systems (say, the GPS) are designed with the account of the time dilation due to the velocity of their satellites (and also some gravity-induced, i.e. general-relativity corrections, which I would not have time to discuss) and would give large errors without such corrections. So, there is no doubt that time dilation (21) is a reality, though the precision of its experimental tests I am aware of[15] has been limited to a few percent, because of the almost unavoidable involvement of less controllable gravity effects – which provide a time interval change of the opposite sign in most experiments near the Earth's surface.

[15] See, e.g., J. Hafele and R. Keating, *Science* **177**, 166 (1972).

Before the first reliable observation of time dilation (by B. Rossi and D. Hall in 1940), there had been serious doubts about the reality of this effect, the most famous being the *twin paradox* first posed (together with an immediate suggestion of its resolution) by P. Langevin in 1911. Let us send one of two twins on a long space roundtrip with the maximum speed approaching c. Upon his return to Earth, who of the twins would be older? The naïve approach is to say that due to the relativity principle, not one can be (and hence there is no time dilation) because each twin could claim that their counterpart rather than them, was moving, with the same speed but in the opposite direction. The resolution of the paradox is that one of the twins had to be accelerated to be brought back, and hence the reference frames have to be dissimilar: only one of them may stay inertial all the time. As a result, the twin who had been accelerated ("actually traveling") would be younger than their sibling when they finally came together. Constructive proof of this conclusion for the particular case of straight-line travel with a piecewise-constant acceleration, is simple and hence left for the reader's exercise.

(iii) Velocity transformation. Now let us calculate the velocity $\mathbf{u}$ of a moving point, as observed in reference frame 0, provided that its velocity, as measured in frame 0*'*, is $\mathbf{u}'$ (Fig. 6).

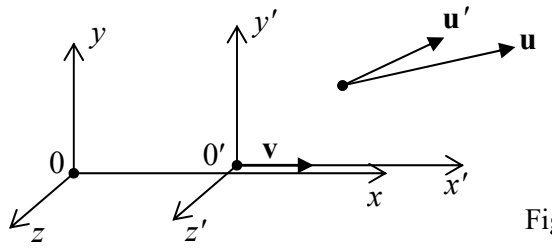


Fig. 9.6. The relativistic velocity addition.

Keeping the usual definition of velocity, but with due attention to the relativity of not only spatial but also temporal intervals, we may write

$$\mathbf{u} \equiv \frac{d\mathbf{r}}{dt}, \qquad \mathbf{u}' \equiv \frac{d\mathbf{r}'}{dt'}. \tag{9.22}$$

Plugging in the differentials of the Lorentz transform relations (6a) into these definitions, we get

$$u_x \equiv \frac{dx}{dt} = \frac{dx' + v dt'}{dt' + v dx'/c^2} \equiv \frac{u'_x + v}{1 + u'_x v/c^2}, \qquad u_y \equiv \frac{dy}{dt} = \frac{1}{\gamma}\frac{dy'}{dt' + v dx'/c^2} \equiv \frac{1}{\gamma}\frac{u'_y}{1 + u'_x v/c^2}, \tag{9.23}$$

with a similar formula for u_z. In the classical limit $v/c \to 0$, these relations are reduced to

$$u_x = u'_x + v, \qquad u_y = u'_y, \qquad u_z = u'_z, \tag{9.24a}$$

and may be merged into the familiar Galilean form

$$\mathbf{u} = \mathbf{u}' + \mathbf{v}, \qquad \text{for } v << c. \tag{9.24b}$$

In order to see how unusual the full relativistic rules (23) are at $u \sim c$, let us first consider a purely longitudinal motion, $u_y = u_z = 0$; then[16]

[16] With an account of the identity $\tanh(a + b) = (\tanh a + \tanh b)/(1 + \tanh a \tanh b)$, which readily follows from MA Eq. (3.5), Eq. (25) shows that rapidities $\varphi \equiv \tanh^{-1}\beta$ add up exactly as longitudinal velocities at non-relativistic motion, making that notion very convenient for the analysis of transfer between several frames.

$$u = \frac{u' + v}{1 + u'v/c^2}, \tag{9.25}$$

Longitudinal velocity addition

where $u \equiv u_x$ and $u' \equiv u'_x$. Figure 7 shows this u as the function of u', for several values of the reference frames' relative velocity v.

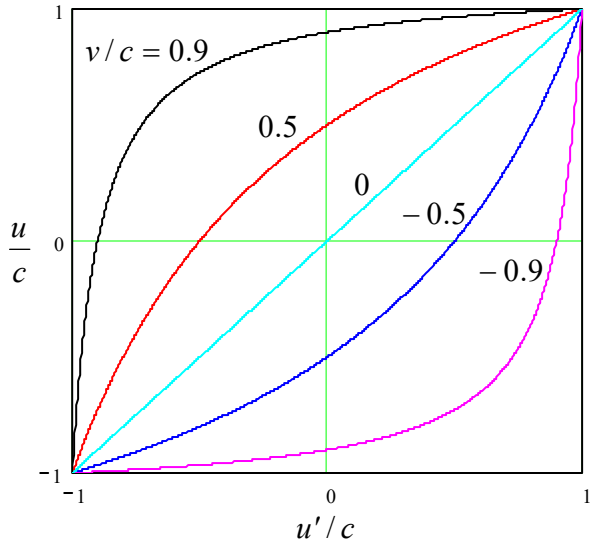


Fig. 9.7. The addition of longitudinal velocities.

The first sanity check is that if $v = 0$, i.e. if the reference frames are at rest relative to each other, then $u = u'$, as it should be – see the diagonal straight line in Fig. 7. Next, if magnitudes of u' and v are both below c, so is the magnitude of u. (Also good, because otherwise, ordinary particles in one frame would be tachyons in the other one, and the theory would be in big trouble.) Now strange things begin: even as u' and v are both approaching c, then u is also close to c, but does not exceed it. As an example, if we fired forward a bullet with the relative speed of $0.9c$, from a spaceship moving from the Earth also at $0.9c$, Eq. (25) predicts the speed of the bullet relative to the Earth to be just $[(0.9 + 0.9)/(1 + 0.9\times0.9)]c \approx 0.994c < c$, rather than $(0.9 + 0.9)\,c = 1.8\,c > c$ as in the Galilean kinematics. Actually, we could expect this strangeness, because it is necessary to fulfill the 2nd Einstein's postulate: the independence of the speed of light in any reference frame. Indeed, for $u' = \pm c$, Eq. (25) yields $u = \pm c$, regardless of v.

In the opposite case of a purely transverse motion, when a point moves across the relative motion of the frames (for example, at our choice of coordinates, $u'_x = u'_z = 0$), Eqs. (23) yield a much less spectacular result

$$u_y = \frac{1}{\gamma} u'_y \le u'_y. \tag{9.26}$$

This effect comes purely from the time dilation because the transverse spatial intervals are Lorentz-invariant.

In the case when both u_x' and u_y' are substantial (but u_z' is still zero), we may divide Eqs. (23) by each other to relate the angles θ of the point's propagation, as observed in the two reference frames:

$$\tan\theta \equiv \frac{u_y}{u_x} = \frac{u'_y}{\gamma(u'_x + v)} = \frac{\sin\theta'}{\gamma(\cos\theta' + v/u')}. \tag{9.27}$$

Stellar aberration effect

This expression describes, in particular, the so-called *stellar aberration* effect: the dependence of the observed direction θ toward a star on the speed v of the telescope's motion relative to the star – see Fig. 8. (The effect is readily observable experimentally as the *annual aberration* due to the periodic change of speed v by $2v_E \approx 60$ km/s because of the Earth's rotation about the Sun. Since the aberration's main part is of the first order in $v_E/c \sim 10^{-4}$, this effect is very significant and has been known since the early 1700s.)

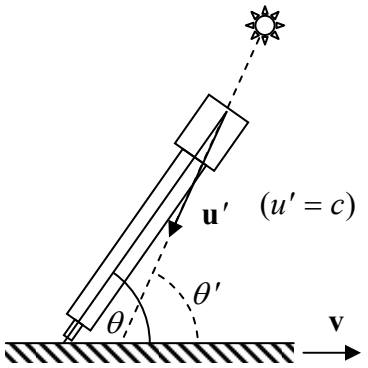


Fig. 9.8. The stellar aberration.

For the analysis of this effect, it is sufficient to take, in Eq. (27), $u' = c$, i.e. $v/u' = \beta$, and interpret θ' as the "proper" direction to the star, which would be measured at $v = 0$.[17] At $\beta << 1$, both Eq. (27) and the Galilean result (which the reader is invited to derive directly from Fig. 8),

$$\tan\theta = \frac{\sin\theta'}{\cos\theta' + \beta}, \tag{9.28}$$

may be well approximated by the first-order term

$$\Delta\theta \equiv \theta - \theta' \approx -\beta\sin\theta'. \tag{9.29}$$

Unfortunately, it is not easy to use the difference between Eqs. (27) and (28), of the second order in β, for special relativity's confirmation, because other components of the Earth's motion, such as its rotation, nutation, and torque-induced precession,[18] give masking first-order contributions to the aberration.

Finally, for a completely arbitrary direction of the vector $\mathbf{u}'$, Eqs. (22) may be readily used to calculate the velocity's magnitude. The most popular form of the resulting expression is the following expression for the square of the relative velocity (or rather the reduced relative velocity $\boldsymbol{\beta}$) of two points,

$$\beta^2 = \frac{(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2)^2 - |\boldsymbol{\beta}_1 \cdot \boldsymbol{\beta}_2|}{(1 - \boldsymbol{\beta}_1 \cdot \boldsymbol{\beta}_2)^2} \le 1. \tag{9.30}$$

where $\boldsymbol{\beta}_{1,2} \equiv \mathbf{v}_{1,2}/c$ are their normalized velocities as measured in the same reference frame.

[17] Strictly speaking, to reconcile the geometries shown in Fig. 1 (for which all our formulas, including Eq. (27), are valid) and Fig. 8 (giving the traditional scheme of the stellar aberration), it is necessary to invert the signs of $\mathbf{u}$ (and hence of $\sin\theta'$ and $\cos\theta'$) and $\mathbf{v}$, but as it is evident from Eq. (27), all the minus signs cancel, and the formula is valid "as is".

[18] See, e.g., CM Secs. 4.4-4.5.

(iv) The Doppler effect. Let us consider a monochromatic plane wave of some physical nature, traveling along the x-axis:

$$f = \mathrm{Re}\left[f_\omega \exp\{i(kx - \omega t\}\right] \equiv \left|f_\omega\right| \cos\left(kx - \omega t + \arg f_\omega\right) \equiv \left|f_\omega\right| \cos \Psi . \tag{9.31}$$

Its total phase, $\Psi \equiv kx - \omega t + \arg f_\omega$ (in contrast to its amplitude $\left|f_\omega\right|$ – see Sec. 5 below) cannot depend on the observer's reference frame, because the variable f vanishes completely at $\Psi = \pi(n + ½)$ (for all integer n), and such "world events" should be observable in all reference frames. The only way to keep $\Psi = \Psi'$ at all times is to have[19]

$$kx - \omega t = k'x' - \omega' t' . \tag{9.32}$$

First, let us use this general relation to consider the Doppler effect in the usual non-relativistic mechanical waves, e.g., oscillations of particles of a certain medium. Using the Galilean transform (2), we may rewrite Eq. (32) as

$$k(x' + vt) - \omega t = k'x' - \omega' t . \tag{9.33}$$

Since this transform leaves all space intervals (including the wavelength $\lambda = 2\pi/k$) intact, we can take $k = k'$, so Eq. (33) yields

$$\omega' = \omega - kv . \tag{9.34}$$

For a dispersion-free medium, the wave number k is the ratio of its frequency ω, as measured in the reference frame bound to the medium, and the wave velocity v_w. In particular, if the wave source rests in the medium, we may bind the reference frame 0 to the medium as well, and frame 0' to the wave's receiver (i.e. $v = v_r$), so

$$k = \frac{\omega}{v_w} , \tag{9.35}$$

and for the frequency perceived by the receiver, Eq. (34) yields

$$\omega' = \omega \frac{v_w - v_r}{v_w} . \tag{9.36}$$

On the other hand, if the receiver and the medium are at rest in the reference frame 0', while the wave source is bound to the frame 0 (so $v = -v_s$), Eq. (35) should be replaced with

$$k = k' = \frac{\omega'}{v_w} , \tag{9.37}$$

and Eq. (34) yields a different result:

$$\omega' = \omega \frac{v_w}{v_w - v_s} , \tag{9.38}$$

Finally, if both the source and detector are moving, it is straightforward to combine these two results to get the general relation

$$\omega' = \omega \frac{v_w - v_r}{v_w - v_s} . \tag{9.39}$$

[19] Strictly speaking, Eq. (32) is valid to an additive constant, but for notation simplicity, it may be always made equal to zero by selecting (as has already been done in all relations of Sec. 1) the reference frame origins and/or clock turn-on times so that at $t = 0$ and $x = 0$, $t' = 0$ and $x' = 0$ as well.

At low speeds of both the source and the receiver, this result simplifies,

$$\omega' \approx \omega(1-\beta), \qquad \text{with } \beta \equiv \frac{v_r - v_s}{v_w}, \tag{9.40}$$

but at speeds comparable to v_w we have to use the more general Eq. (39). Thus, the usual Doppler effect is generally affected not only by the relative speed ($v_r - v_s$) of the wave's source and detector but also by their speeds relative to the medium in which the waves propagate.

Somewhat counter-intuitively, for the electromagnetic waves the calculations are *simpler* because for them the propagation medium (aether) does not exist, the wave velocity equals $\pm c$ in any reference frame, and there are no two separate cases: we can always take $k = \pm\omega/c$ and $k' = \pm\omega'/c$. Plugging these relations, together with the Lorentz transform (19a), into the phase-invariance condition (32), we get

$$\pm\frac{\omega}{c}\gamma(x' + \beta c t') - \omega\gamma\frac{ct' + \beta x'}{c} = \pm\frac{\omega'}{c}x' - \omega' t'. \tag{9.41}$$

This relation has to hold for any x' and t', so we may require that the net coefficients before these variables vanish. These two requirements yield the same equality:

$$\omega' = \omega\gamma(1 \mp \beta). \tag{9.42}$$

This result is already quite simple, but may be transformed further to be even more illuminating:

$$\omega' = \omega\frac{1 \mp \beta}{(1-\beta^2)^{1/2}} \equiv \omega\left[\frac{(1 \mp \beta)(1 \mp \beta)}{(1+\beta)(1-\beta)}\right]^{1/2}. \tag{9.43}$$

At any sign before β, one pair of parentheses cancels, so[20]

Longitudinal Doppler effect

$$\boxed{\omega' = \omega\left(\frac{1 \mp \beta}{1 \pm \beta}\right)^{1/2}}. \tag{9.44}$$

Thus the Doppler effect for electromagnetic waves depends only on the relative velocity $v = \beta c$ between the wave source and detector – as it should be, given the aether's absence. At velocities much lower than c, Eq. (44) may be approximated as

$$\omega' \approx \omega\frac{1 \mp \beta/2}{1 \pm \beta/2} \approx \omega(1 \mp \beta), \tag{9.45}$$

i.e. in the first approximation in $\beta \equiv v/c$, it tends to the corresponding limit (40) of the usual Doppler effect.

If the wave vector **k** is tilted by angle θ to the vector **v** (as measured in frame 0), then we have to repeat the calculations, with k replaced by k_x, and components k_y and k_z left intact at the Lorentz transform. As a result, Eq. (42) is generalized as

[20] It may look like the reciprocal expression of ω via ω' is different, violating the relativity principle. However, in this case, we have to change the sign of β, because the relative velocity of the system is opposite, so we return to Eq. (44) again.

$$\omega' = \omega\gamma\left(1 - \beta\cos\theta\right). \tag{9.46}$$

For the case $\cos\theta = \pm 1$, Eq. (46) reduces to our previous result (42). However, at $\theta = \pi/2$ (i.e. $\cos\theta = 0$), the relation is rather different:

$$\omega' = \gamma\omega \equiv \frac{\omega}{\left(1-\beta^2\right)^{1/2}}. \tag{9.47}$$

Transverse Doppler effect

This is the *transverse Doppler effect* – which is absent in non-relativistic physics. Its first experimental evidence was obtained using electron beams (as had been suggested in 1906 by J. Stark), by H. Ives and G. Stilwell in 1938 and 1941. Later, similar experiments were repeated several times, but the first unambiguous measurements were performed only in 1979 by D. Hasselkamp *et al.* who confirmed Eq. (47) with a relative accuracy of about 10%. This precision may not look too spectacular, but besides the special tests discussed above, the Lorentz transform formulas have been also confirmed, less directly, by a huge body of other experimental data, especially in high energy physics, agreeing with calculations incorporating this transform as their part. This is why, with due respect to the spirit of challenging authority, I should warn the reader: if you decide to challenge the relativity theory (called "theory" by tradition only), you would also need to explain all these data. Best luck with that! [21]

9.3. 4-vectors, momentum, mass, and energy

Before proceeding to the relativistic dynamics, let us discuss the mathematical formalism that makes all calculations more compact – and more beautiful. We have already seen that the three spatial coordinates $\{x, y, z\}$ and the product ct are Lorentz-transformed similarly – see Eqs. (18)-(19) again. So it is natural to consider them as components of a single four-component vector (or, for short, *4-vector*),

$$\{x_0, x_1, x_2, x_3\} \equiv \{ct, \mathbf{r}\}, \tag{9.48}$$

with components

$$x_0 \equiv ct, \quad x_1 \equiv x, \quad x_2 \equiv y, \quad x_3 \equiv z. \tag{9.49}$$

Space-time 4-vector

According to Eqs. (19), its components are Lorentz-transformed as

$$x_j = \sum_{j'=0}^{3} L_{jj'} x'_{j'}, \tag{9.50}$$

Lorentz transform: 4-form

where $L_{jj'}$ are the elements of the following 4×4 *Lorentz transform matrix*

$$\begin{pmatrix} \gamma & \beta\gamma & 0 & 0 \\ \beta\gamma & \gamma & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}. \tag{9.51}$$

Lorentz transform matrix

Since such 4-vectors are a new notion for this course and will be used for many more purposes than just the space-time transform, we need to discuss the general mathematical rules they obey. Indeed,

[21] The same fact, ignored by crackpots, is also valid for other favorite directions of their attacks, including the Universe expansion, quantum measurement uncertainty, and entropy growth in physics, and the evolution theory in biology.

as was already mentioned in Sec. 8.9, the usual (three-component) vector is not just any ordered set (*string*) of three scalars $\{A_x, A_y, A_z\}$; if we want it to represent a reference-frame-independent physical reality, the vector's components have to obey certain rules at the transfer from one reference frame to another. In particular, in the non-relativistic limit the vector's *norm* (its magnitude squared),

$$A^2 = A_x^2 + A_y^2 + A_z^2, \tag{9.52}$$

should be invariant with respect to the transfer between different reference frames. However, a naïve extension of this approach to 4-vectors would not work, because, according to the calculations of Sec. 1, the Lorentz transform keeps intact the combinations of the type (7), with one sign negative, rather than the sum of all components squared. Hence for the 4-vectors, all the rules of the game have to be reviewed and adjusted – or rather redefined from the very beginning, for example as follows.[22]

An arbitrary 4-vector is a string of 4 scalars,[23]

General 4-vector

$$\{A_0, A_1, A_2, A_3\}, \tag{9.53}$$

whose components A_j, as measured in the reference frames 0 and 0' shown in Fig. 1, obey the Lorentz transform relations similar to Eq. (50):

Lorentz transform: general 4-vector

$$A_j = \sum_{j'=0}^{3} L_{jj'} A'_{j'}. \tag{9.54}$$

As we have already seen in the example of the space-time 4-vector (48), this means in particular that

Lorentz invariance

$$A_0^2 - \sum_{j=1}^{3} A_j^2 = (A'_0)^2 - \sum_{j=1}^{3} (A'_j)^2. \tag{9.55}$$

This is the so-called *Lorentz invariance* condition for the *4-vector's norm*. (The difference between this relation and Eq. (52), pertaining to Euclidian geometry, is the reason why the Minkowski space is called *pseudo-Euclidian*.) It is also straightforward to use Eqs. (51) and (54) to check that the evident generalization of the norm, the *scalar product* of two arbitrary 4-vectors,

Scalar 4-product

$$A_0 B_0 - \sum_{j=1}^{3} A_j B_j, \tag{9.56}$$

is also Lorentz-invariant.

Now consider the 4-vector corresponding to a small *interval* between two close world events:

$$\{dx_0, dx_1, dx_2, dx_3\} = \{cdt, d\mathbf{r}\}; \tag{9.57}$$

its norm,

Interval

$$(ds)^2 \equiv dx_0^2 - \sum_{j=1}^{3} dx_j^2 = c^2 (dt)^2 - (dr)^2, \tag{9.58}$$

[22] The most prominent alternative, which has both advantages and drawbacks, is to use 4-vectors with one imaginary component – for example, the imaginary time *ict* instead of the real product *ct* in Eq. (48).

[23] Such vectors are said to reside in so-called 4D *Minkowski spaces* – called after Hermann Minkowski who was the first one to recast (in 1907) the special relativity relations in a form in which the spatial coordinates and time (or rather *ct*) are treated on an equal footing.

is of course also Lorentz-invariant. Since the speed of any particle (or signal) cannot be larger than c, for any pair of world events that are in a causal relation with each other, $(dr)^2$ cannot be larger than $(cdt)^2$, i.e. such *time-like* interval $(ds)^2$ cannot be negative. The 4D surface separating such intervals from *space-like* intervals $(ds)^2 < 0$ is called the light cone (Fig. 9).

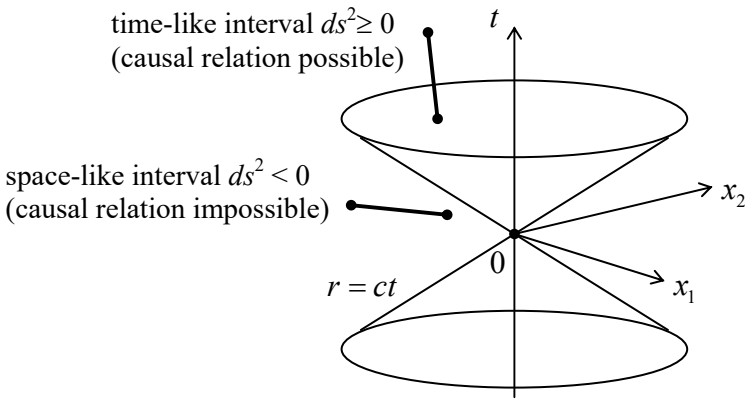


Fig. 9.9. A 2+1 dimensional image of the light cone – which is actually 3+1 dimensional.

Now let us consider two close world events that happen with the same point moving with velocity **u**. Then in the frame moving with the point (**v** = **u**), the last term on the right-hand side of Eq. (58) equals zero, while the involved time is the proper one, so

$$ds = cd\tau\,, \tag{9.59}$$

where $d\tau$ is the proper time interval. But according to Eq. (21), this means that we can write

$$d\tau = \frac{dt}{\gamma}\,, \tag{9.60}$$

where dt is the time interval in an *arbitrary* (besides being inertial) reference frame, while

$$\boldsymbol{\beta} \equiv \frac{\mathbf{u}}{c} \quad \text{and} \quad \gamma \equiv \frac{1}{\left(1-\beta^2\right)^{1/2}} = \frac{1}{\left(1-u^2/c^2\right)^{1/2}} \tag{9.61}$$

are the parameters (17) corresponding to the *point's* velocity (**u**) in that frame, so $ds = cdt/\gamma$.[24]

Let us use Eq. (60) to explore whether a 4-vector may be formed using the spatial Cartesian components of the point's velocity

$$\mathbf{u} = \left\{\frac{dx}{dt}, \frac{dy}{dt}, \frac{dz}{dt}\right\}. \tag{9.62}$$

Here we have a problem: per Eqs. (22), these components do not obey the Lorentz transform. However, let us use $d\tau \equiv dt/\gamma$, the proper time interval of the point, to form the following string:

$$\left\{\frac{dx_0}{d\tau}, \frac{dx_1}{d\tau}, \frac{dx_2}{d\tau}, \frac{dx_3}{d\tau}\right\} \equiv \gamma\left\{c, \frac{dx}{dt}, \frac{dy}{dt}, \frac{dz}{dt}\right\} \equiv \gamma\{c, \mathbf{u}\}. \tag{9.63}$$

4-velocity

[24] I have opted against using special indices (e.g., $\boldsymbol{\beta}_u$ and γ_u) to distinguish Eqs. (17) and (61) here and below, in a hope that the suitable velocity (of either a reference frame or a particle) will be always clear from the context.

As it follows from the comparison of the middle form of this expression with Eq. (48), since the time-space vector obeys the Lorentz transform, and τ is Lorentz-invariant, the string (63) is a legitimate 4-vector; it is called the *4-velocity* of a point – or of a point particle.

Now we are well equipped to proceed to relativistic dynamics. Let us start with such basic notions as the momentum **p** and the energy $\mathscr{E}$ – so far, for a free particle.[25] Perhaps the most elegant way to "derive" (or rather guess[26]) the expressions for **p** and $\mathscr{E}$ as functions of the particle's velocity **u**, is based on analytical mechanics. Due to the conservation of **v**, the trajectory of a free particle in the 4D Minkowski space $\{ct, \mathbf{r}\}$ is always a straight line. Hence, from the Hamilton principle,[27] we may expect its action $\mathscr{S}$, between points 1 and 2, to be a linear function of the space-time interval (59):

Free particle: action

$$\mathscr{S} = \alpha \int_1^2 ds \equiv \alpha c \int_1^2 d\tau \equiv \alpha c \int_{t_1}^{t_2} \frac{dt}{\gamma}, \tag{9.64}$$

where α is some constant. On the other hand, in analytical mechanics, the action is defined as

$$\mathscr{S} \equiv \int_{t_1}^{t_2} \mathscr{L} dt\,, \tag{9.65}$$

where $\mathscr{L}$ is the particle's Lagrangian function.[28] Comparing these two expressions, we get

$$\mathscr{L} = \frac{\alpha c}{\gamma} \equiv \alpha c \left(1 - \frac{u^2}{c^2}\right)^{1/2}. \tag{9.66}$$

In the non-relativistic limit ($u << c$), this function tends to

$$\mathscr{L} \approx \alpha c \left(1 - \frac{u^2}{2c^2}\right) \equiv \alpha c - \frac{\alpha u^2}{2c}. \tag{9.67}$$

In order to correspond to the Newtonian mechanics,[29] the last (velocity-dependent) term should equal $mu^2/2$. From here we find $\alpha = -mc$, so, finally,

Free particle: Lagrangian function

$$\mathscr{L} = -mc^2 \left(1 - \frac{u^2}{c^2}\right)^{1/2} \equiv -\frac{mc^2}{\gamma}. \tag{9.68}$$

Now we can find the Cartesian components p_j of the particle's momentum as the generalized momenta corresponding to the corresponding components r_j (j = 1, 2, 3) of the 3D radius-vector **r**:[30]

---

[25] I am sorry for using, just as in Sec. 6.3, the same traditional notation (**p**) for the particle's momentum as had been used earlier for the electric dipole moment. However, since the latter notion will be virtually unused in the balance of this course, this may hardly lead to confusion.

[26] Indeed, such a derivation uses additional assumptions, however natural (such as the Lorentz-invariance of $\mathscr{S}$), i.e. it can hardly be considered as a real proof of the final results, so they require experimental confirmation. Fortunately, such confirmations have been numerous – see below.

[27] See, e.g., CM Sec. 10.3.

[28] See, e.g., CM Sec. 2.1.

[29] See, e.g., CM Eq. (2.19b).

[30] See, e.g., CM Sec. 2.3, in particular Eq. (2.31).

$$p_j = \frac{\partial \mathscr{L}}{\partial \dot{r}_j} \equiv \frac{\partial \mathscr{L}}{\partial u_j} = -mc^2 \frac{\partial}{\partial u_j}\left(1 - \frac{u_1^2 + u_2^2 + u_3^2}{c^2}\right)^{1/2} = \frac{mu_j}{\left(1 - u^2/c^2\right)^{1/2}} \equiv m\gamma u_j . \quad (9.69)$$

Thus for the 3D vector of momentum, we can write the result in the same form as in non-relativistic mechanics,

$$\boxed{\mathbf{p} = m\gamma \mathbf{u} \equiv M\mathbf{u},} \quad (9.70)$$

Relativistic momentum

using the reference-frame-dependent scalar M (called the *relativistic mass*) defined as

$$\boxed{M \equiv m\gamma = \frac{m}{\left(1 - u^2/c^2\right)^{1/2}} \geq m,} \quad (9.71)$$

Relativistic mass

m being the non-relativistic mass of the particle. (More often, m is called the *rest mass*, because in the reference frame in which the particle rests, Eq. (71) yields $M = m$.)

Next, let us return to analytical mechanics to calculate the particle's energy $\mathscr{E}$ (which for a free particle coincides with its Hamiltonian function $\mathscr{H}$):[31]

$$\mathscr{E} = \mathscr{H} \equiv \sum_{j=1}^{3} p_j u_j - \mathscr{L} = \mathbf{p}\cdot\mathbf{u} - \mathscr{L} = \frac{mu^2}{\left(1 - u^2/c^2\right)^{1/2}} + mc^2\left(1 - \frac{u^2}{c^2}\right)^{1/2} \equiv \frac{mc^2}{\left(1 - u^2/c^2\right)^{1/2}} . \quad (9.72)$$

Thus, we have arrived at the most famous of Einstein's formulas – and probably of physics as a whole:

$$\boxed{\mathscr{E} = m\gamma c^2 \equiv Mc^2,} \quad (9.73)$$

$\mathscr{E} = Mc^2$

which expresses the relation between the free particle's mass and its energy.[32] In the non-relativistic limit, it reduces to

$$\mathscr{E} = \frac{mc^2}{\left(1 - u^2/c^2\right)^{1/2}} \approx mc^2\left(1 + \frac{u^2}{2c^2}\right) = mc^2 + \frac{mu^2}{2}, \quad (9.74)$$

the first term mc^2 being called the *rest energy* of a particle.

Now let us consider the following string of 4 scalars:

$$\boxed{\left\{\frac{\mathscr{E}}{c}, p_1, p_2, p_3\right\} \equiv \left\{\frac{\mathscr{E}}{c}, \mathbf{p}\right\}.} \quad (9.75)$$

4-vector of energy-momentum

Using Eqs. (70) and (73) to represent this expression as

$$\left\{\frac{\mathscr{E}}{c}, \mathbf{p}\right\} = m\gamma\{c, \mathbf{u}\}, \quad (9.76)$$

[31] See, e.g., CM Eq. (2.32).

[32] Let me hope that the reader understands that all the layman talk about the "mass to energy conversion" is only valid in a very limited sense of the word. While the Einstein relation (73) does allow the conversion of "massive" particles (with $m \neq 0$) into particles with $m = 0$, such as photons, each of the latter particles also has a non-zero relativistic mass M, and *simultaneously* the energy $\mathscr{E}$ related to this M by Eq. (73).

and comparing the result with Eq. (63), we immediately see that, since m is a Lorentz-invariant constant, this string is a legitimate 4-vector of *energy-momentum*. As a result, its norm,

$$\left(\frac{\mathscr{E}}{c}\right)^2 - p^2, \tag{9.77a}$$

is Lorentz-invariant, and in particular, has to be equal to the norm in the particle-bound frame. But in that frame, $p = 0$, and according to Eq. (73), $\mathscr{E} = mc^2$, and the norm is just

$$\left(\frac{\mathscr{E}}{c}\right)^2 = \left(\frac{mc^2}{c}\right)^2 \equiv (mc)^2, \tag{9.77b}$$

so in an arbitrary frame

$$\left(\frac{\mathscr{E}}{c}\right)^2 - p^2 = (mc)^2. \tag{9.78a}$$

This very important relation[33] between the relativistic energy and momentum (valid for free particles only!) is usually represented in the form[34]

Free particle: energy

$$\boxed{\mathscr{E}^2 = (mc^2)^2 + (pc)^2}. \tag{9.78b}$$

According to Eq. (70), in the so-called *ultra-relativistic limit* $u \to c$, p tends to infinity, while mc^2 stays constant, so $pc/mc^2 \to \infty$. As follows from Eq. (78), in this limit $\mathscr{E} \approx pc$. Though the above discussion was for particles with finite m, the 4-vector formalism allows us to consider compact objects with zero rest mass as ultra-relativistic particles for which the above energy-to-moment relation,

$$\mathscr{E} = pc, \qquad \text{for } m = 0, \tag{9.79}$$

is exact. Quantum electrodynamics[35] tells us that under certain conditions, the electromagnetic field quanta (photons) may be also considered as such *massless particles* with momentum $\mathbf{p} = \hbar\mathbf{k}$. Plugging (the modulus of) the last relation into Eq. (78), for the photon's energy we get $\mathscr{E} = pc = \hbar kc = \hbar\omega$. Please note again that according to Eq. (73), the relativistic mass of a photon is not equal to zero: $M = \mathscr{E}/c^2 = \hbar\omega/c^2$, so the term "massless particle" has a limited meaning: $m = 0$. For example, the relativistic mass of an optical phonon is of the order of 10^{-36} kg. On the human scale, this is not too much, but still, a noticeable (approximately one-millionth) part of the rest mass m_e of an electron.

The fundamental relations (70) and (73) have been repeatedly verified in numerous particle collision experiments, in which the total energy and momentum of a system of particles are conserved – at the same conditions as in non-relativistic dynamics. (For the momentum, this is the absence of external forces, and for the energy, the elasticity of particle interactions – in other words, the absence of alternative channels of energy escape.) Of course, generally only the total energy of the system is conserved, including the potential energy of particle interactions. However, at typical high-energy

[33] Please note one more simple and useful relation following from Eqs. (70) and (73): $\mathbf{p} = (\mathscr{E}/c^2)\mathbf{u}$.

[34] It may be tempting to interpret this relation as the perpendicular-vector-like addition of the rest energy mc^2 and the "kinetic energy" pc, but from the point of view of the total energy conservation (see below), a better definition of the kinetic energy is $T(u) \equiv \mathscr{E}(u) - \mathscr{E}(0)$.

[35] It is briefly reviewed in QM Chapter 9.

particle collisions, the potential energy vanishes so rapidly with the distance between them that we can use the momentum and energy conservation laws using Eq. (73).

As an example, let us calculate the minimum energy $\mathscr{E}_{\min}$ of a proton (p_a), necessary for the well-known high-energy reaction that generates a new proton-antiproton pair, $p_a + p_b \to p + p + p + \bar{p}$, provided that before the collision, proton p_b had been at rest in the lab frame. This minimum corresponds to the vanishing relative velocity of the reaction products, i.e. their motion with virtually the same velocity ($\mathbf{u}_{\text{fin}}$), as seen from the lab frame – see Fig. 10.

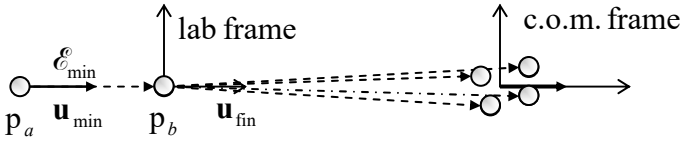


Fig. 9.10. A high-energy proton reaction at $\mathscr{E} \approx \mathscr{E}_{\min}$ – schematically.

Due to the momentum conservation, this velocity should have the same direction as the initial velocity ($\mathbf{u}_{\min}$) of proton p_a. This is why two scalar equations: for energy conservation,

$$\frac{mc^2}{\left(1-u_{\min}^2/c^2\right)^{1/2}} + mc^2 = \frac{4mc^2}{\left(1-u_{\text{fin}}^2/c^2\right)^{1/2}}, \tag{9.80a}$$

and for momentum conservation,

$$\frac{mu}{\left(1-u_{\min}^2/c^2\right)^{1/2}} + 0 = \frac{4mu_{\text{fin}}}{\left(1-u_{\text{fin}}^2/c^2\right)^{1/2}}, \tag{9.80b}$$

are sufficient to find both $u_{\min}$ and u_{fin}. After a rather tedious solution of this system of two nonlinear equations, we get

$$u_{\min} = \frac{4\sqrt{3}}{7}c \approx 0.990\,c, \qquad u_{\text{fin}} = \frac{\sqrt{3}}{2}c \approx 0.866\,c. \tag{9.81}$$

Finally, we can use Eq. (72) to calculate the required energy; the result is $\mathscr{E}_{\min} = 7\,mc^2$. (Note that at this threshold, only a minor $2mc^2$ part of the kinetic energy $T_{\min} = \mathscr{E}_{\min} - mc^2 = 6mc^2$ of the initially moving particle, goes into the "useful" proton-antiproton pair production.) The proton's rest mass, $m_p \approx 1.67\times10^{-27}$ kg, corresponds to $m_pc^2 \approx 1.502\times10^{-10}$ J ≈ 0.938 GeV, so $\mathscr{E}_{\min} \approx 6.57$ GeV.

The second, more intelligent way to solve the same problem is to use the center-of-mass (*c.o.m.*) reference frame that, in relativity, is defined as the frame in which the total momentum of the system vanishes.[36] In this frame, at $\mathscr{E} = \mathscr{E}_{\min}$, the velocity and momenta of all reaction products are vanishing, while the velocities of the protons p_a and p_b before the collision are equal and opposite, with an initially unknown magnitude u'. Hence the energy conservation law becomes

$$\frac{2mc^2}{\left(1-u'^2/c^2\right)^{1/2}} = 4mc^2, \tag{9.82}$$

[36] Note that according to this definition, the c.o.m.'s radius-vector is $\mathbf{R} = \sum_k M_k\mathbf{r}_k/\sum_k M_k \equiv \sum_k \gamma_k m_k\mathbf{r}_k/\sum_k \gamma_k m_k$, i.e. is generally different from the well-known non-relativistic expression $\mathbf{R} = \sum_k m_k\mathbf{r}_k/\sum_k m_k$.

readily giving $u'/c = \sqrt{3}/2$. (This is of course the same result as Eq. (81) gives for $u_{\rm fin}$.) Now we can use the fact that the velocity of the proton p_a in the c.o.m. frame is $(-u')$, to find its lab-frame speed, using the velocity transform (25):

$$u_{\min} = \frac{2u'}{1+u'^2/c^2}. \tag{9.83}$$

With the above result for u', this relation gives the same result as the first method, $u_{\min}/c = 4\sqrt{3}/7$, but in a simpler way.

9.4. More on 4-vectors and 4-tensors

This is a good moment to introduce a formalism that will allow us, in particular, to solve the same proton collision problem in one more (and arguably, the most elegant) way. Much more importantly, this formalism will be virtually necessary for the description of the Lorentz transform of the electromagnetic field, and its interaction with relativistic particles – otherwise the formulas would be too cumbersome.

Let us call the 4-vectors we have used before,

Contravariant 4-vectors

$$A^\alpha \equiv \{A_0, \mathbf{A}\}, \tag{9.84}$$

contravariant, and denote them with top indices, and introduce also *covariant* vectors,

Covariant 4-vectors

$$A_\alpha \equiv \{A_0, -\mathbf{A}\}, \tag{9.85}$$

marked by bottom indices. Now if we form a scalar product of these two vectors using the *standard* (3D-like) rule, just as a sum of the products of the corresponding components, we immediately get

$$A_\alpha A^\alpha \equiv A^\alpha A_\alpha \equiv A_0^2 - A^2. \tag{9.86}$$

Note that the first and the second expressions may be understood as sums over four components of the product, with the summation sign dropped.[37] The scalar product (86) is just the norm of the 4-vector in our former definition, and as we already know, is Lorentz-invariant. Moreover, the scalar product of two different vectors (also a Lorentz invariant), may be rewritten in any of two similar forms:[38]

Scalar product's forms

$$A_0 B_0 - \mathbf{A}\cdot\mathbf{B} \equiv A_\alpha B^\alpha = A^\alpha B_\alpha; \tag{9.87}$$

again, the only caveat is to take one vector in the covariant, and the other one in the contravariant form.

Now let us return to our sample problem (Fig. 10). Since all components ($\mathscr{E}/c$ and $\mathbf{p}$) of the total 4-momentum of our system are conserved at the collision, its norm is conserved as well:

$$(p_a + p_b)_\alpha (p_a + p_b)^\alpha = (4p)_\alpha (4p)^\alpha. \tag{9.88}$$

[37] This compact notation may take some time to be accustomed to, but is very convenient (compact) and can hardly lead to any confusion, due to the following rule: the summation is implied when, and only when the same index is repeated twice, once on the top and another at the bottom. (It is frequently called *dummy index*, because its notation may be replaced with any other letter not used in the same formula.) In this course, this shorthand notation will be used only for 4-vectors, but not for the usual (3D spatial) vectors.

[38] Note also that, by definition, for any two 4-vectors, $A_\alpha B^\alpha = B^\alpha A_\alpha$.

Since now the vector product is the usual math construct, we know that the parentheses on the left-hand side of this equation may be multiplied as usual. We may also swap the operands and move constant factors through products as convenient. As a result, we get

$$(p_a)_\alpha (p_a)^\alpha + (p_b)_\alpha (p_b)^\alpha + 2(p_a)_\alpha (p_b)^\alpha = 16 p_\alpha p^\alpha . \qquad (9.89)$$

Thanks to the Lorentz invariance of each of the terms, we may calculate it in the reference frame we like. For the first two terms on the left-hand side, as well as for the right-hand side term, it is beneficial to use the frames in which that particular proton is at rest; as a result, according to Eq. (77b), each of the two left-hand-side terms equals $(mc)^2$, while the right-hand side equals $16(mc)^2$. On the contrary, the last term on the left-hand side is more easily evaluated in the lab frame, because in it, the three spatial components of the 4-momentum p_b vanish, and the scalar product is just the product of the scalars $\mathscr{E}/c$ for protons a and b. For the latter proton, being at rest, this ratio is just mc so we get a simple equation,

$$(mc)^2 + (mc)^2 + 2\frac{\mathscr{E}_{\min}}{c} mc = 16(mc)^2 , \qquad (9.90)$$

immediately giving the final result $\mathscr{E}_{\min} = 7\, mc^2$, already obtained earlier in two more complex ways.

Let me hope that this example was a convincing demonstration of the convenience of representing 4-vectors in the contravariant (84) and covariant (85) forms,[39] with Lorentz-invariant norms (86). To be useful for more complex tasks, this formalism should be developed a little bit further. In particular, it is crucial to know how the 4-vectors change under the Lorentz transform. For contravariant vectors, we already know the answer (54); let us rewrite it in our new notation:

$$\boxed{A^\alpha = L^\alpha_\beta A'^\beta .} \qquad (9.91)$$

Lorentz transform: contravariant vectors

where L^α_β is the matrix (51), generally called the *mixed Lorentz tensor*:[40]

$$\boxed{L^\alpha_\beta = \begin{pmatrix} \gamma & \beta\gamma & 0 & 0 \\ \beta\gamma & \gamma & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix},} \qquad (9.92)$$

Mixed Lorentz tensor

Note that though the position of the indices α and β in the Lorentz tensor notation is not crucial, because this tensor is symmetric, it is convenient to place them using the general *index balance rule*: the difference of the numbers of the upper and lower indices should be the same in both parts of any 4-vector/tensor equality. (You may check that all the formulas above do satisfy this rule.)

[39] These forms are 4-vector extensions of the notions of contravariance and covariance, introduced in the 1850s by J. Sylvester (who also introduced the term "matrix" in its mathematical sense) for the description of the change of the usual 3-component spatial vectors at the transfer between different reference frames – e.g., resulting from the frame rotation. In this case, the contravariance or covariance of a vector is uniquely determined by its nature: if the Cartesian coordinates of a vector (such as the non-relativistic velocity $\mathbf{v} = d\mathbf{r}/dt$) are transformed similarly to the radius-vector $\mathbf{r}$, it is called contravariant, while the vectors (such as ∇f) that require the reciprocal transform, are called covariant. In the 4D Minkowski space, both forms may be used for any 4-vector.

[40] Just as the 4-vectors, 4-tensors with two top indices are called contravariant, and those with two bottom indices, are covariant. The tensors with one top and one bottom index are called *mixed*.

In order to rewrite Eq. (91) in a more general form that would not depend on the particular orientation of the coordinate axes (Fig. 1), let us use the contravariant and covariant forms of the 4-vector of the time-space interval (57),

$$dx^{\alpha} = \{cdt, d\mathbf{r}\}, \qquad dx_{\alpha} = \{cdt, -d\mathbf{r}\}; \tag{9.93}$$

then its norm (58) may be represented as[41]

$$(ds)^2 \equiv (cdt)^2 - (dr)^2 = dx^{\alpha} dx_{\alpha} = dx_{\alpha} dx^{\alpha}. \tag{9.94}$$

Applying Eq. (91) to the first, contravariant form of the 4-vector (93), we get

$$dx^{\alpha} = L^{\alpha}_{\beta} dx'^{\beta}. \tag{9.95}$$

But with our new shorthand notation, we can also write the usual rule of differentiation of each component x^{α}, considering it a function (in our case, linear) of four arguments x'^{β}, as follows:[42]

$$dx^{\alpha} = \frac{\partial x^{\alpha}}{\partial x'^{\beta}} dx'^{\beta}. \tag{9.96}$$

Comparing Eqs. (95) and (96), we can rewrite the general Lorentz transform rule (92) in a new form,

Lorentz transform: general form

$$\boxed{A^{\alpha} = \frac{\partial x^{\alpha}}{\partial x'^{\beta}} A'^{\beta}}. \tag{9.97a}$$

which does not depend on the coordinate axes' orientation.

It is straightforward to verify that the reciprocal transform may be represented as

Reciprocal Lorentz transform

$$\boxed{A'^{\alpha} = \frac{\partial x'^{\alpha}}{\partial x^{\beta}} A^{\beta}}. \tag{9.97b}$$

However, the reciprocal transform has to differ from the direct one only by the sign of the relative velocity of the frames, so for the coordinate choice shown in Fig. 1, its matrix is

[41] Another way to write this relation is $(ds)^2 = g_{\alpha\beta} dx^{\alpha} dx^{\beta} = g^{\alpha\beta} dx_{\alpha} dx_{\beta}$, where double summation over indices α and β is implied, and g is the so-called *metric tensor*,

$$g^{\alpha\beta} \equiv g_{\alpha\beta} \equiv \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 \end{pmatrix},$$

which may be used, in particular, to transfer a covariant vector into the corresponding contravariant one and back: $A^{\alpha} = g^{\alpha\beta} A_{\beta}$, $A_{\alpha} = g_{\alpha\beta} A^{\beta}$. The metric tensor plays a key role in general relativity, in which it is affected by gravity – "curved" by particles' masses.

[42] Note that in the index balance rule, the top index in the denominator of a fraction is counted as a bottom index in the numerator, and vice versa.

$$\frac{\partial x'^{\alpha}}{\partial x^{\beta}} = \begin{pmatrix} \gamma & -\beta\gamma & 0 & 0 \\ -\beta\gamma & \gamma & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \tag{9.98}$$

Since according to Eqs. (84)-(85), covariant 4-vectors differ from the contravariant ones by the sign of their spatial components, their direct transform is given by matrix (98). Hence their direct and reciprocal transforms may be represented, respectively, as

$$\boxed{A_{\alpha} = \frac{\partial x'^{\beta}}{\partial x^{\alpha}} A'_{\beta}, \qquad A'_{\alpha} = \frac{\partial x^{\beta}}{\partial x'^{\alpha}} A_{\beta},} \tag{9.99}$$

Lorentz transform: covariant vectors

evidently satisfying the index balance rule. (Note that primed quantities are now multiplied, rather than divided as in the contravariant case.) As a sanity check, let us apply this formalism to the scalar product $A_{\alpha}A^{\alpha}$. As Eq. (96) shows, the implicit-sum notation allows us to multiply and divide any equality by the same partial differential of a coordinate, so we can write:

$$A_{\alpha}A^{\alpha} = \frac{\partial x'^{\beta}}{\partial x^{\alpha}} \frac{\partial x^{\alpha}}{\partial x'^{\gamma}} A'_{\beta}A'^{\gamma} = \frac{\partial x'^{\beta}}{\partial x'^{\gamma}} A'_{\beta}A'^{\gamma} = \delta_{\beta\gamma} A'_{\beta}A'^{\gamma} = A'_{\gamma}A'^{\gamma}, \tag{9.100}$$

i.e. the scalar product $A_{\alpha}A^{\alpha}$ (as well as $A^{\alpha}A_{\alpha}$) is Lorentz-invariant, as it should be.

Now, let us consider the 4-vectors of derivatives. Here we should be very careful. Consider, for example, the following 4-vector operator

$$\frac{\partial}{\partial x^{\alpha}} \equiv \left\{ \frac{\partial}{\partial (ct)}, \nabla \right\}, \tag{9.101}$$

As was discussed above, the operator is not changed by its multiplication and division by another differential, e.g., $\partial x'^{\beta}$ (with the corresponding implied summation over all four values of β), so

$$\frac{\partial}{\partial x^{\alpha}} = \frac{\partial x'^{\beta}}{\partial x^{\alpha}} \frac{\partial}{\partial x'^{\beta}}. \tag{9.102}$$

But, according to the first of Eqs. (99), this is exactly how the covariant vectors are Lorentz-transformed! Hence, we have to consider the derivative over a *contravariant* space-time interval as a *covariant* 4-vector, and vice versa.[43] (This result might be also expected from the index balance rule.) In particular, this means that the scalar product

$$\frac{\partial}{\partial x^{\alpha}} A^{\alpha} \equiv \frac{\partial A_0}{\partial (ct)} + \nabla \cdot \mathbf{A} \tag{9.103}$$

should be Lorentz-invariant for any legitimate 4-vector. A convenient shorthand for the covariant derivative, which complies with the index balance rule, is

$$\frac{\partial}{\partial x^{\alpha}} \equiv \partial_{\alpha}, \tag{9.104}$$

[43] As was mentioned above, this is also a property of the reference-frame transform of the "usual" 3D vectors.

so the invariant scalar product may be written just as $\partial_\alpha A^\alpha$. A similar definition of the contravariant derivative,

$$\partial^\alpha \equiv \frac{\partial}{\partial x_\alpha} = \left\{ \frac{\partial}{\partial (ct)}, -\nabla \right\}, \tag{9.105}$$

allows us to write the Lorentz-invariant scalar product (103) in any of the following two forms:

$$\frac{\partial A_0}{\partial (ct)} + \nabla \cdot \mathbf{A} = \partial^\alpha A_\alpha = \partial_\alpha A^\alpha . \tag{9.106}$$

Finally, let us see how the general Lorentz transform changes 4-tensors. A second-rank 4×4 matrix is a legitimate 4-tensor if the 4-vectors it relates obey the Lorentz transform. For example, if two legitimate 4-vectors are related as

$$A^\alpha = T^{\alpha\beta} B_\beta , \tag{9.107}$$

we should require that

$$A'^\alpha = T'^{\alpha\beta} B'_\beta , \tag{9.108}$$

where A^α and A'^α are related by Eqs. (97), while B_β and B'_β, by Eqs. (99). This requirement immediately yields

Lorentz transform of 4-tensors

$$\boxed{T^{\alpha\beta} = \frac{\partial x^\alpha}{\partial x'^\gamma} \frac{\partial x^\beta}{\partial x'^\delta} T'^{\gamma\delta}, \qquad T'^{\alpha\beta} = \frac{\partial x'^\alpha}{\partial x^\gamma} \frac{\partial x'^\beta}{\partial x^\delta} T^{\gamma\delta},} \tag{9.109}$$

with the implied summation over two indices, γ and δ. The rules for the covariant and mixed tensors are similar.[44]

9.5. Maxwell equations in the 4-form

This 4-vector formalism background is sufficient to analyze the Lorentz transform of the electromagnetic field. Just to warm up, let us consider the continuity equation (4.5),

$$\frac{\partial \rho}{\partial t} + \nabla \cdot \mathbf{j} = 0 , \tag{9.110}$$

which expresses the electric charge conservation, and as we already know, is compatible with the Maxwell equations. If we now define the contravariant and covariant *4-vectors of electric current* as

4-vector of electric current

$$\boxed{j^\alpha \equiv \{\rho c, \mathbf{j}\}, \qquad j_\alpha \equiv \{\rho c, -\mathbf{j}\},} \tag{9.111}$$

then Eq. (110) may be represented in the form

Continuity equation: 4-form

$$\boxed{\partial^\alpha j_\alpha = \partial_\alpha j^\alpha = 0 ,} \tag{9.112}$$

showing that the continuity equation is *form-invariant*[45] with respect to the Lorentz transform.

---

[44] It is straightforward to check that transfer between the contravariant and covariant forms of the same tensor may be readily achieved using the metric tensor g: $T_{\alpha\beta} = g_{\alpha\gamma} T^{\gamma\delta} g_{\delta\beta}$, $T^{\alpha\beta} = g^{\alpha\gamma} T_{\gamma\delta} g^{\delta\beta}$.

[45] In some texts, the formulas preserving their form at a transform are called "covariant", creating a possibility for confusion with the covariant vectors and tensors. On the other hand, calling such *formulas* "invariant" would not distinguish them properly from invariant *quantities*, such as the scalar products of 4-vectors.

Of course, such a form-invariance of a relation does not mean that all component *values* of the 4-vectors participating in it are the same in both frames. For example, let us have some static charge density ρ in frame 0; then Eq. (97b), applied to the contravariant form of the 4-vector (111), reads

$$j'^{\alpha} = \frac{\partial x'^{\alpha}}{\partial x^{\beta}} j^{\beta}, \qquad \text{with } j^{\beta} = \{\rho c, 0, 0, 0\}. \tag{9.113}$$

Using the particular form (98) of the reciprocal Lorentz matrix for the coordinate choice shown in Fig. 1, we see that this relation yields

$$\rho' = \gamma\rho, \qquad j'_x = -\gamma\beta\rho c = -\gamma v\rho, \qquad j'_y = j'_z = 0. \tag{9.114}$$

Lorentz transforms of ρ and **j**

Since the charge velocity, as observed from frame 0', is (–**v**), the non-relativistic results would be $\rho' = \rho$, $\mathbf{j}' = -\mathbf{v}\rho$. The additional γ factor in the relativistic results is caused by the length contraction: $dx' = dx/\gamma$, so to keep the total charge $dQ = \rho d^3r = \rho dxdydz$ inside the elementary volume $d^3r = dxdydz$ intact, ρ (and hence j_x) should increase proportionally.

Next, at the end of Chapter 6 we have seen that Maxwell equations for the electromagnetic potentials ϕ and **A** may be represented in similar forms (6.118), under the Lorenz (again, not "Lorentz", please!) gauge condition (6.117). For free space, this condition takes the form

$$\nabla \cdot \mathbf{A} + \frac{1}{c^2}\frac{\partial \phi}{\partial t} = 0. \tag{9.115}$$

This expression gives us a hint of how to form the 4-vector of electromagnetic potentials:[46]

$$A^{\alpha} \equiv \left\{\frac{\phi}{c}, \mathbf{A}\right\}, \qquad A_{\alpha} \equiv \left\{\frac{\phi}{c}, -\mathbf{A}\right\}; \tag{9.116}$$

4-vector of potentials

indeed, this vector satisfies Eq. (115) in its 4-form:

$$\partial^{\alpha} A_{\alpha} = \partial_{\alpha} A^{\alpha} = 0. \tag{9.117}$$

Lorenz gauge: 4-form

Since this scalar product is Lorentz-invariant, and the derivatives (104)-(105) are legitimate 4-vectors, this implies that the 4-vector (116) is also legitimate, i.e. obeys the Lorentz transform formulas (97), (99). Even more convincing evidence of this fact may be obtained from the Maxwell equations (6.118) for the potentials. In free space, they may be rewritten as

$$\left[\frac{\partial^2}{\partial(ct)^2} - \nabla^2\right]\frac{\phi}{c} = \frac{\rho c}{\varepsilon_0 c^2} \equiv \mu_0(\rho c), \qquad \left[\frac{\partial^2}{\partial(ct)^2} - \nabla^2\right]\mathbf{A} = \mu_0 \mathbf{j}. \tag{9.118}$$

Using the definition (116), these equations may be merged to one:[47]

$$\Box A^{\alpha} = \mu_0 j^{\alpha}, \tag{9.119}$$

Maxwell equation for 4-potential

where $\Box$ is the *d'Alembert operator*,[48] which may be represented as either of two scalar products:

[46] In the Gaussian units, the scalar potential should not be divided by c in this relation.

[47] In the Gaussian units, the coefficient μ_0 in Eq. (119) should be replaced, as usual, with $4\pi/c$.

D'Alembert operator

$$\Box \equiv \frac{\partial^2}{\partial (ct)^2} - \nabla^2 = \partial^\beta \partial_\beta = \partial_\beta \partial^\beta, \tag{9.120}$$

and hence is Lorentz-invariant. Because of that, and the fact that the Lorentz transform changes both 4-vectors A^α and j^α in a similar way, Eq. (119) does not depend on the reference frame choice. Thus we have arrived at a key point of this chapter: we see that the Maxwell equations are indeed form-invariant with respect to the Lorentz transform. As a by-product, the 4-vector form (119) of these equations (for potentials) is extremely simple – and beautiful!

However, as we have seen in Chapter 7, for many applications the Maxwell equations for the field vectors are more convenient; so let us represent them in the 4-form as well. For that, we may express all Cartesian components of the usual (3D) field vector vectors (6.7),

$$\mathbf{E} = -\nabla\phi - \frac{\partial \mathbf{A}}{\partial t}, \qquad \mathbf{B} = \nabla \times \mathbf{A}, \tag{9.121}$$

via those of the potential 4-vector A^α. For example,

$$E_x = -\frac{\partial \phi}{\partial x} - \frac{\partial A_x}{\partial t} \equiv -c\left(\frac{\partial}{\partial x}\frac{\phi}{c} + \frac{\partial A_x}{\partial (ct)}\right) \equiv -c\left(\partial^0 A^1 - \partial^1 A^0\right), \tag{9.122}$$

$$B_x = \frac{\partial A_z}{\partial y} - \frac{\partial A_y}{\partial z} \equiv -\left(\partial^2 A^3 - \partial^3 A^2\right). \tag{9.123}$$

Completing similar calculations for other field components (or just generating them by appropriate index shifts), we find that the following antisymmetric, contravariant *field-strength tensor*,

$$F^{\alpha\beta} \equiv \partial^\alpha A^\beta - \partial^\beta A^\alpha, \tag{9.124}$$

may be expressed via the field components as follows:[49]

Field-strength tensors

$$F^{\alpha\beta} = \begin{pmatrix} 0 & -E_x/c & -E_y/c & -E_z/c \\ E_x/c & 0 & -B_z & B_y \\ E_y/c & B_z & 0 & -B_x \\ E_z/c & -B_y & B_x & 0 \end{pmatrix}, \tag{9.125a}$$

so the covariant form of the tensor is

$$F_{\alpha\beta} \equiv g_{\alpha\gamma} F^{\gamma\delta} g_{\delta\beta} = \begin{pmatrix} 0 & E_x/c & E_y/c & E_z/c \\ -E_x/c & 0 & -B_z & B_y \\ -E_y/c & B_z & 0 & -B_x \\ -E_z/c & -B_y & B_x & 0 \end{pmatrix}. \tag{9.125b}$$

[48] Named after Jean-Baptiste le Rond d'Alembert (1717-1783), who has made several pioneering contributions to the general theory of waves – see, e.g., CM Chapter 6. (Some older textbooks use notation $\Box^2$ for this operator.)

[49] In Gaussian units, this formula, as well as Eq. (131) for $G^{\alpha\beta}$, do not have the factor c in all the denominators.

If Eq. (124) looks a bit too bulky, please note that as a reward, the pair of *inhomogeneous* Maxwell equations, i.e. two equations of the system (6.99), which in free space ($\mathbf{D} = \varepsilon_0\mathbf{E}$, $\mathbf{B} = \mu_0\mathbf{H}$) may be rewritten as

$$\nabla\cdot\frac{\mathbf{E}}{c} = \mu_0 c\rho, \qquad \nabla\times\mathbf{B} - \frac{\partial}{\partial(ct)}\frac{\mathbf{E}}{c} = \mu_0\mathbf{j}, \tag{9.126}$$

may now be expressed in a very simple (and manifestly form-invariant) way,

$$\boxed{\partial_\alpha F^{\alpha\beta} = \mu_0 j^\beta}, \tag{9.127}$$

Maxwell equation for tensor *F*

which is comparable with Eq. (119) in its simplicity – and beauty. Somewhat counter-intuitively, the pair of *homogeneous* Maxwell equations of the system (6.99),

$$\nabla\times\mathbf{E} + \frac{\partial\mathbf{B}}{\partial t} = 0, \qquad \nabla\cdot\mathbf{B} = 0, \tag{9.128}$$

look, in the 4-vector notation, a bit more complicated:[50]

$$\boxed{\partial_\alpha F_{\beta\gamma} + \partial_\beta F_{\gamma\alpha} + \partial_\gamma F_{\alpha\beta} = 0}. \tag{9.129}$$

Note, however, that Eqs. (128) may be also represented in a much simpler 4-form,

$$\partial_\alpha G^{\alpha\beta} = 0, \tag{9.130}$$

using the so-called *dual tensor*

$$G^{\alpha\beta} = \begin{pmatrix} 0 & B_x & B_y & B_z \\ -B_x & 0 & -E_z/c & E_y/c \\ -B_y & E_z/c & 0 & -E_x/c \\ -B_z & -E_y/c & E_x/c & 0 \end{pmatrix}, \tag{9.131}$$

which may be obtained from $F^{\alpha\beta}$, given by Eq. (125a), by the following replacements:

$$\frac{\mathbf{E}}{c} \to -\mathbf{B}, \quad \mathbf{B} \to \frac{\mathbf{E}}{c}. \tag{9.132}$$

Besides the proof of the form-invariance of the Maxwell equations with respect to the Lorentz transform, the 4-vector formalism allows us to achieve our initial goal: to find out how the electric and magnetic field components change at the transfer between two (inertial!) reference frames. For that, let us apply to the tensor $F^{\alpha\beta}$ the reciprocal Lorentz transform described by the second of Eqs. (109). Generally, it gives, for each field component, a sum of 16 terms, but since (for our choice of coordinates, shown in Fig. 1) there are many zeros in the Lorentz transform matrix, and the diagonal components of $F^{\gamma\delta}$ equal zero as well, the calculations are rather doable. Let us calculate, for example, $E'_x \equiv -cF'^{01}$. The only non-zero terms on the right-hand side are

$$E'_x = -cF'^{01} = -c\left(\frac{\partial x'^0}{\partial x^1}\frac{\partial x'^1}{\partial x^0}F^{10} + \frac{\partial x'^0}{\partial x^0}\frac{\partial x'^1}{\partial x^1}F^{01}\right) \equiv -c\gamma^2\left(\beta^2 - 1\right)\frac{E_x}{c} \equiv E_x. \tag{9.133}$$

[50] To be fair, note that just as Eq. (127), Eq. (129) is also a set of four scalar equations – in the latter case with the indices α, β, and γ taking any three *different* values of the set {0, 1, 2, 3}.

Repeating the calculation for the other five components of the fields, we get very important relations

$$\begin{aligned} E'_x &= E_x, & B'_x &= B_x, \\ E'_y &= \gamma\left(E_y - vB_z\right), & B'_y &= \gamma\left(B_y + vE_z/c^2\right), \\ E'_z &= \gamma\left(E_z + vB_y\right), & B'_z &= \gamma\left(B_z - vE_y/c^2\right), \end{aligned} \tag{9.134}$$

whose more compact "semi-vector" form is

Lorentz transform of field components

$$\boxed{\begin{aligned} E'_{||} &= E_{||}, & B'_{||} &= B_{||}, \\ \mathbf{E}'_{\perp} &= \gamma\left(\mathbf{E} + \mathbf{v}\times\mathbf{B}\right)_{\perp}, & \mathbf{B}'_{\perp} &= \gamma\left(\mathbf{B} - \mathbf{v}\times\mathbf{E}/c^2\right)_{\perp}, \end{aligned}} \tag{9.135}$$

where the indices $||$ and $\perp$ stand, respectively, for the field components parallel and normal to the relative velocity $\mathbf{v}$ of the two reference frames. In the non-relativistic limit, the Lorentz factor γ tends to 1, and Eqs. (135) acquire an even simpler form

$$\mathbf{E}' \to \mathbf{E} + \mathbf{v}\times\mathbf{B}, \qquad \mathbf{B}' \to \mathbf{B} - \frac{1}{c^2}\mathbf{v}\times\mathbf{E}. \tag{9.136}$$

Thus we see that the electric and magnetic fields are transformed to each other even in the first order of the v/c ratio. For example, if we fly across the field lines of a uniform, static, purely electric field $\mathbf{E}$ (e.g., the one in a plane capacitor) we will see not only the electric field's renormalization (in the second order of the v/c ratio), but also a non-zero dc magnetic field $\mathbf{B}'$ perpendicular to both the vector $\mathbf{E}$ and the vector $\mathbf{v}$, i.e. to the direction of our motion. This is of course what might be expected from the relativity principle: from the point of view of the moving observer (which is as legitimate as that of a stationary observer), the surface charges of the capacitor's plates, that create the field $\mathbf{E}$, move back creating the dc currents (114), which induce the magnetic field $\mathbf{B}'$. Similarly, motion across a magnetic field creates, from the point of view of the moving observer, an electric field.

This fact is very important conceptually. One may say there is no such thing in Mother Nature as an electric field (or a magnetic field) all by itself. Not only can the electric field induce the magnetic field (and vice versa) in dynamics, but even in an apparently static configuration, what exactly we measure depends on our speed relative to the field sources – justifying once again the term *electromagnetism* for the field of physics we are studying in this course.

Another simple but very important application of Eqs. (134)-(135) is the calculation of the fields created by a charged particle moving in free space by inertia, i.e. along a straight line with constant velocity $\mathbf{u}$, at the *impact parameter*[51] (the closest distance) b from the observer. Selecting the reference frame 0' to move with the particle in its origin, and the reference frame 0 to reside in the "lab" in which the fields $\mathbf{E}$ and $\mathbf{B}$ are measured, we can use the above formulas with $\mathbf{v} = \mathbf{u}$. In this case, the fields $\mathbf{E}'$ and $\mathbf{B}'$ may be calculated from, respectively, electro- and magnetostatics:

$$\mathbf{E}' = \frac{q}{4\pi\varepsilon_0}\frac{\mathbf{r}'}{r'^3}, \qquad \mathbf{B}' = 0, \tag{9.137}$$

because in frame 0', the particle does not move. Selecting the coordinate axes so that at the measurement point, $x = 0$, $y = b$, $z = 0$ (Fig. 11a), for this point we may write $x' = -ut'$, $y' = b$, $z' = 0$, so $r' = (u^2t'^2 + b^2)^{1/2}$, and the Cartesian components of the fields (137) are:

[51] This term is very popular in the theory of particle scattering – see, e.g., CM Sec. 3.7.

$$E'_x = -\frac{q}{4\pi\varepsilon_0}\frac{ut'}{\left(u^2t'^2+b^2\right)^{3/2}}, \qquad E'_y = \frac{q}{4\pi\varepsilon_0}\frac{b}{\left(u^2t'^2+b^2\right)^{3/2}}, \qquad E'_z = 0, \tag{9.138}$$

$$B'_x = B'_y = B'_z = 0.$$

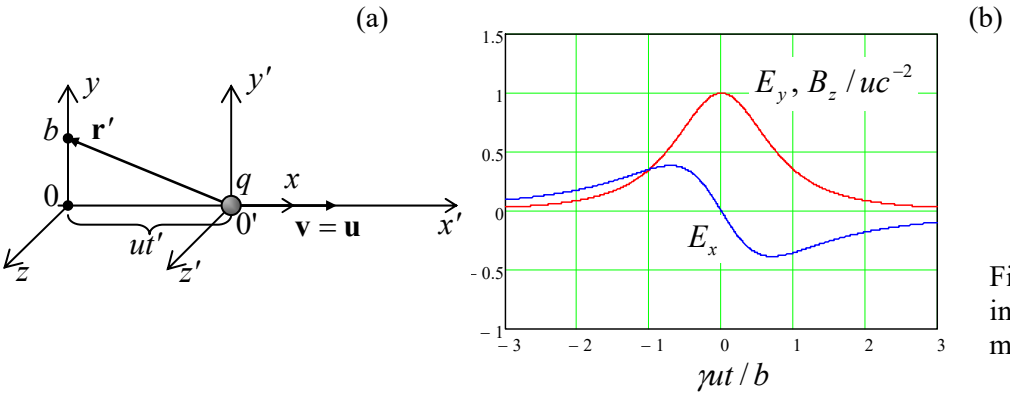


Fig. 9.11. The field pulses induced by a uniformly moving charge.

Now using the last of Eqs. (19b) with $x = 0$, giving $t' = \gamma t$, and the relations reciprocal to Eqs. (134) for the field transform (they are similar to the direct transform but with v replaced with $-v = -u$), in the lab frame we get

$$E_x = E'_x = -\frac{q}{4\pi\varepsilon_0}\frac{u\gamma t}{\left(u^2\gamma^2t^2+b^2\right)^{3/2}}, \qquad E_y = \gamma E'_y = \frac{q}{4\pi\varepsilon_0}\frac{\gamma b}{\left(u^2\gamma^2t^2+b^2\right)^{3/2}}, \qquad E_z = 0, \tag{9.139}$$

$$B_x = 0, \qquad B_y = 0, \qquad B_z = \frac{\gamma u}{c^2}E'_y = \frac{u}{c^2}\frac{q}{4\pi\varepsilon_0}\frac{\gamma b}{\left(u^2\gamma^2t^2+b^2\right)^{3/2}} \equiv \frac{u}{c^2}E_y. \tag{9.140}$$

These results,[52] plotted in Fig. 11b in the units of $\gamma q^2/4\pi\varepsilon_0 b^2$, reveal two major effects. First, the charge passage by the observer generates not only an electric field pulse but also a magnetic field pulse. This is natural, because, as was repeatedly discussed in Chapter 5, any charge motion is essentially an electric current.[53] Second, Eqs. (139)-(140) show that the pulse duration scale is

$$\Delta t = \frac{b}{\gamma u} \equiv \frac{b}{u}\left(1-\frac{u^2}{c^2}\right)^{1/2}, \tag{9.141}$$

i.e. shrinks to virtually zero as the charge's velocity u approaches the speed of light. This is of course a direct corollary of the relativistic length contraction. Indeed, in the frame 0′ moving with the charge, the longitudinal spread of its electric field at distance b from the motion line is of the order of $\Delta x' = b$. When observed from the lab frame 0, this interval, in accordance with Eq. (20), shrinks to $\Delta x = \Delta x'/\gamma = b/\gamma$, and hence so does the pulse duration scale $\Delta t = \Delta x/u = b/\gamma u$.

[52] In the next chapter, we will re-derive them in a different way.

[53] It is straightforward to use Eqs. (140) and the linear superposition principle to calculate, for example, the magnetic field of a string of charges moving along the same line and separated by equal distances $\Delta x = a$ (so the average current, as measured in frame 0, is qu/a), and to show that the time-average of the magnetic field is given by the familiar Eq. (5.20) of magnetostatics, with b instead of ρ.

9.6. Relativistic particles in electric and magnetic fields

Now let us analyze the dynamics of charged particles in electric and magnetic fields. Inspired by "our" success in forming the 4-vector (75) of energy-momentum, with the contravariant form

$$p^{\alpha} = \left\{\frac{\mathscr{E}}{c}, \mathbf{p}\right\} = \gamma\{mc, \mathbf{p}\} = m\frac{dx^{\alpha}}{d\tau} \equiv mu^{\alpha}, \tag{9.142}$$

where u^{α} is the contravariant form of the 4-velocity (63) of the particle,

$$u^{\alpha} \equiv \frac{dx^{\alpha}}{d\tau}, \qquad u_{\alpha} \equiv \frac{dx_{\alpha}}{d\tau}, \tag{9.143}$$

we may notice that the non-relativistic equation of motion, resulting from the Lorentz-force formula (5.10) for the three spatial components of p^{α}, for a charged particle's motion in an electromagnetic field,

Particle's equation of motion

$$\boxed{\frac{d\mathbf{p}}{dt} = q\left(\mathbf{E} + \mathbf{u}\times\mathbf{B}\right),} \tag{9.144}$$

is fully consistent with the following 4-vector equality (which is evidently form-invariant with respect to the Lorentz transform):

Particle's dynamics: 4-form

$$\boxed{\frac{dp^{\alpha}}{d\tau} = qF^{\alpha\beta}u_{\beta}.} \tag{9.145}$$

For example, according to Eq. (125), the $\alpha = 1$ component of this equation reads

$$\frac{dp^{1}}{d\tau} = qF^{1\beta}u_{\beta} = q\left[\frac{E_x}{c}\gamma c + 0\cdot(-\gamma u_x) + (-B_z)(-\gamma u_y) + B_y(-\gamma u_z)\right] = q\gamma\left[\mathbf{E} + \mathbf{u}\times\mathbf{B}\right]_x, \tag{9.146}$$

and similarly for two other spatial components ($\alpha = 2$ and $\alpha = 3$). It may look that these expressions differ from the 2nd Newton law (144) by an extra factor of γ. However, plugging into Eq. (146) the definition of the proper time interval, $d\tau = dt/\gamma$, and canceling γ in both parts, we recover Eq. (144) exactly – for *any* velocity of the particle! The only caveat is that if u is comparable with c, the vector $\mathbf{p}$ in Eq. (144) has to be understood as the relativistic momentum (70), proportional to the velocity-dependent mass $M = \gamma m \geq m$ rather than to the rest mass m.

The only remaining general task is to examine the meaning of the 0th component of Eq. (145). Let us spell it out:

$$\frac{dp^{0}}{d\tau} = qF^{0\beta}u_{\beta} = q\left[0\cdot\gamma c + \left(-\frac{E_x}{c}\right)(-\gamma u_x) + \left(-\frac{E_y}{c}\right)(-\gamma u_y) + \left(-\frac{E_z}{c}\right)(-\gamma u_z)\right] = q\gamma\frac{\mathbf{E}}{c}\cdot\mathbf{u}. \tag{9.147}$$

Recalling that $p^0 = \mathscr{E}/c$, and using the basic relation $d\tau = dt/\gamma$ again, we see that Eq. (147) looks exactly like the non-relativistic relation for the kinetic energy change (what is sometimes called the *work-energy principle*, in our case for the Lorentz force only[54]):

[54] See, e.g., CM Eq. (1.20) divided by dt, and with $d\mathbf{p}/dt = \mathbf{F} = q\mathbf{E}$. (As a reminder, the magnetic field cannot affect the particle's energy, because the magnetic component of the Lorentz force is perpendicular to its velocity.)

$$\frac{d\mathscr{E}}{dt} = q\mathbf{E}\cdot\mathbf{u}\,, \tag{9.148}$$

Particle's energy: evolution

besides that in the relativistic case, the energy has to be taken in the general form (73).

Without question, the 4-component equation (145) of the relativistic dynamics is absolutely beautiful in its simplicity. However, for the solution of particular problems, Eqs. (144) and (148) are frequently more convenient. As an illustration of this point, let us now use these equations to explore relativistic effects at charged particle motion in uniform, time-independent electric and magnetic fields. In doing that, we will, for the time being, neglect the contributions into the field by the particle itself.[55]

(i) Uniform magnetic field. Let the magnetic field be constant and uniform in the "lab" reference frame 0 that is used for measurements. Then in this frame, Eqs. (144) and (148) yield

$$\frac{d\mathbf{p}}{dt} = q\mathbf{u}\times\mathbf{B}, \qquad \frac{d\mathscr{E}}{dt} = 0. \tag{9.149}$$

From the second equation, $\mathscr{E}$ = const, we get u = const, $\beta \equiv u/c$ = const, $\gamma \equiv (1-\beta^2)^{-1/2}$ = const, and $M \equiv \gamma m$ = const, so the first of Eqs. (149) may be rewritten as

$$\frac{d\mathbf{u}}{dt} = \mathbf{u}\times\boldsymbol{\omega}_c\,, \tag{9.150}$$

where $\boldsymbol{\omega}_c$ is the vector directed along the magnetic field $\mathbf{B}$, with the magnitude equal to the following *cyclotron frequency* (sometimes called "gyrofrequency"):

$$\omega_c \equiv \frac{qB}{M} = \frac{qB}{\gamma m} = \frac{qc^2B}{\mathscr{E}}. \tag{9.151}$$

Cyclotron frequency

If the particle's initial velocity $\mathbf{u}_0$ is perpendicular to the magnetic field, Eq. (150) describes its circular motion, with a constant speed $u = u_0$, in a plane normal to $\mathbf{B}$, with the angular velocity (151). In the non-relativistic limit $u << c$, when $\gamma \to 1$, i.e. $M \to m$, the cyclotron frequency ω_c equals qB/m, i.e. is independent of the speed. However, as the kinetic energy of the particle is increased to become comparable with its rest energy mc^2, the frequency decreases, and in the ultra-relativistic limit,

$$\omega_c \approx qc\frac{B}{p} << \frac{qB}{m}, \qquad \text{for } u \approx c\,. \tag{9.152}$$

The cyclotron motion's radius may be calculated as $R = u/\omega_c$; in the non-relativistic limit, it is proportional to the particle's speed, i.e. to the square root of its kinetic energy. However, as Eq. (151) shows, in the general case the radius is proportional to the particle's relativistic momentum rather than its speed:

$$R = \frac{u}{\omega_c} = \frac{Mu}{qB} = \frac{m\gamma u}{qB} = \frac{1}{q}\frac{p}{B}, \tag{9.153}$$

Cyclotron radius

so in the ultra-relativistic limit, when $p \approx \mathscr{E}/c$, R is proportional to the kinetic energy.

[55] As was emphasized earlier in this course, in statics this contribution is formally infinite and has to be ignored. In dynamics, this is generally not true; these *self-action effects* (which are, in most cases, negligible) will be discussed in the next chapter.

These dependencies of ω_c and R on energy are the major factors in the design of circular accelerators of charged particles. In the simplest of these machines (the *cyclotron*, invented in 1929 by Ernest Orlando Lawrence), the frequency ω of the accelerating ac electric field is constant, so even if it is tuned to the ω_c of the initially injected particles, the drop of the cyclotron frequency with energy eventually violates this tuning. Due to this reason, the largest achievable particle's speed is limited to just ~0.1 c (for protons, corresponding to the kinetic energy of just ~15 MeV). This problem may be addressed in several ways. In particular, in *synchrotrons* (such as Fermilab's Tevatron and the CERN's Large Hadron Collider, LHC[56]) the magnetic field is gradually increased in time to compensate for the momentum increase ($B \propto p$), so both R (148) and ω_c (147) stay constant, enabling proton acceleration to energies as high as ~ 7 TeV, i.e. ~2,000 mc^2.[57]

Returning to our initial problem, if the particle's initial velocity has a component $u_{\|}$ along the magnetic field, then it is conserved in time, so the trajectory is a spiral around the magnetic field lines. As Eqs. (149) show, in this case, Eq. (150) remains valid but in Eqs. (151) and (153) the full speed and momentum have to be replaced with magnitudes of their (also time-conserved) components, $u_\perp$ and $p_\perp$, normal to **B**, while the Lorentz factor γ in those formulas still includes the full speed of the particle.

Finally, in the special case when the particle's initial velocity is directed *exactly* along the magnetic field's direction, it continues to move straight along the vector **B**. In this case, the cyclotron frequency still has the non-zero value (151) but does not correspond to any real motion, because $R = 0$.

(ii) Uniform electric field. This problem is (technically) more complex than the previous one because in the electric field, the particle's energy changes. Directing the z-axis along the field **E**, from Eq. (144) we get

$$\frac{dp_z}{dt} = qE, \qquad \frac{d\mathbf{p}_\perp}{dt} = 0\,. \tag{9.154}$$

If E does not change in time, the first integration of these equations is elementary,

$$p_z(t) = p_z(0) + qEt, \qquad \mathbf{p}_\perp(t) = \text{const} = \mathbf{p}_\perp(0)\,, \tag{9.155}$$

but the further integration requires care because the effective mass $M = \gamma m$ of the particle depends on its full speed u, with

$$u^2 = u_z^2 + u_\perp^2\,, \tag{9.156}$$

making the two motions, along and across the field, mutually dependent.

If the initial velocity is perpendicular to the field **E**, i.e. if $p_z(0) = 0$, $p_\perp(0) = p(0) \equiv p_0$, the easiest way to proceed is to calculate the kinetic energy first:

$$\mathscr{E}^2 = \left(mc^2\right)^2 + c^2 p^2(t) \equiv \mathscr{E}_0^2 + c^2\left(qEt\right)^2, \qquad \text{where } \mathscr{E}_0 \equiv \left[(mc^2)^2 + c^2 p_0^2\right]^{1/2}. \tag{9.157}$$

On the other hand, we can calculate the same energy by integrating Eq. (148),

---

[56] See https://home.cern/topics/large-hadron-collider.

[57] I am sorry I have no more time/space to discuss particle accelerator physics, and have to refer the interested reader to special literature, for example, either S. Lee, *Accelerator Physics*, 2nd ed., World Scientific, 2004, or E. Wilson, *An Introduction to Particle Accelerators*, Oxford U. Press, 2001.

$$\frac{d\mathscr{E}}{dt} = q\mathbf{E}\cdot\mathbf{u} \equiv qE\frac{dz}{dt}, \tag{9.158}$$

over time, with a simple result:

$$\mathscr{E} = \mathscr{E}_0 + qEz(t), \tag{9.159}$$

where (just for the notation simplicity) I took $z(0) = 0$. Requiring Eq. (159) to give the same $\mathscr{E}^2$ as Eq. (157), we get a quadratic equation for the function $z(t)$,

$$\mathscr{E}_0^2 + c^2(qEt)^2 = \left[\mathscr{E}_0 + qEz(t)\right]^2, \tag{9.160}$$

whose solution (with the sign before the square root corresponding to $E > 0$, i.e. to $z \geq 0$) is

$$z(t) = \frac{\mathscr{E}_0}{qE}\left\{\left[1 + \left(\frac{cqEt}{\mathscr{E}_0}\right)^2\right]^{1/2} - 1\right\}. \tag{9.161}$$

Now let us find the particle's trajectory. Directing the x-axis so that the initial velocity vector (and hence the velocity vector at any further instant) is within the $[x, z]$ plane, i.e. that $y(t) = 0$ identically, we may use Eqs. (155) to calculate the trajectory's slope, at its arbitrary point, as

$$\frac{dz}{dx} \equiv \frac{dz/dt}{dx/dt} \equiv \frac{Mu_z}{Mu_x} \equiv \frac{p_z}{p_x} = \frac{qEt}{p_0}. \tag{9.162}$$

Now let us use Eq. (160) to express the numerator of this fraction, qEt, as a function of z:

$$qEt = \frac{1}{c}\left[\left(\mathscr{E}_0 + qEz\right)^2 - \mathscr{E}_0^2\right]^{1/2}. \tag{9.163}$$

Plugging this expression into Eq. (162), we get

$$\frac{dz}{dx} = \frac{1}{cp_0}\left[\left(\mathscr{E}_0 + qEz\right)^2 - \mathscr{E}_0^2\right]^{1/2}. \tag{9.164}$$

This differential equation may be readily integrated separating the variables z and x, and using the substitution $\xi \equiv \cosh^{-1}(qEz/\mathscr{E}_0 + 1)$. Selecting the origin of axis x at the initial point, so $x(0) = 0$, we finally get the trajectory:

$$z = \frac{\mathscr{E}_0}{qE}\left(\cosh\frac{qEx}{cp_0} - 1\right). \tag{9.165}$$

This curve is usually called the *catenary*, but sometimes the "chainette" – because it (with the proper constant replacement) describes, in particular, the stationary shape of a heavy uniform chain in a uniform gravity field directed along the z-axis. At the initial part of the trajectory, where $qEx << cp_0(0)$, this expression may be approximated with the first non-zero term of its Taylor expansion in small x, giving the following parabola:

$$z = \frac{\mathscr{E}_0 qE}{2}\left(\frac{x}{cp_0}\right)^2, \tag{9.166}$$

so if the initial velocity of the particle is much lower than c (i.e. $p_0 \approx mu_0$, $\mathscr{E}_0 \approx mc^2$), we get the very familiar non-relativistic formula:

$$z = \frac{qE}{2mu_0^2}x^2 \equiv \frac{a}{2}t^2, \qquad \text{with } a = \frac{F}{m} = \frac{qE}{m}. \tag{9.167}$$

The generalization of this solution to the case of an arbitrary direction of the particle's initial velocity is left for the reader's exercise.

(iii) Crossed uniform magnetic and electric fields (**E** ⊥ **B**). In view of the somewhat bulky solution of the previous problem (i.e. the particular case of the current problem for **B** = 0), one might think that this problem, with **B** ≠ 0, should be forbiddingly complex for an analytical solution. Counter-intuitively, this is not the case, due to the help from the field transform relations (135). Let us consider two possible cases.

Case 1: $E/c < B$. Let us consider an inertial reference frame 0' moving (relatively the "lab" reference frame 0 in that the fields **E** and **B** are measured) with the following velocity:

$$\mathbf{v} = \frac{\mathbf{E}\times\mathbf{B}}{B^2}, \tag{9.168}$$

and hence the speed $v = c(E/c)/B < c$. Selecting the coordinate axes as shown in Fig. 12, so

$$E_x = 0, \quad E_y = E, \quad E_z = 0; \qquad B_x = 0, \quad B_y = 0, \quad B_z = B, \tag{9.169}$$

we see that the Cartesian components of this velocity are $v_x = v$, $v_y = v_z = 0$.

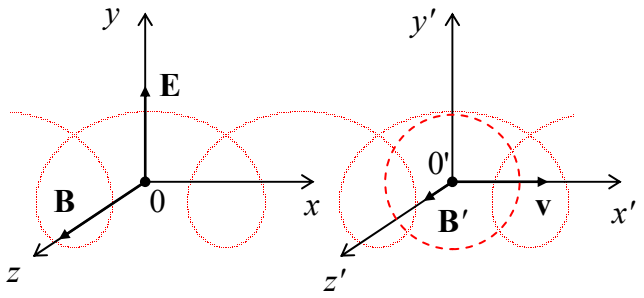


Fig. 9.12. Particle's trajectory in crossed electric and magnetic fields (at $E/c < B$).

Since this choice of the coordinates complies with the one used to derive Eqs. (134), we can readily use that simple form of the Lorentz transform to calculate the field components in the moving reference frame:

$$E'_x = 0, \qquad E'_y = \gamma(E - vB) \equiv \gamma\left(E - \frac{E}{B}B\right) \equiv 0, \qquad E'_z = 0, \tag{9.170}$$

$$B'_x = 0, \qquad B'_y = 0, \qquad B'_z = \gamma\left(B - \frac{vE}{c^2}\right) \equiv \gamma B\left(1 - \frac{vE}{Bc^2}\right) \equiv \gamma B\left(1 - \frac{v^2}{c^2}\right) \equiv \frac{B}{\gamma} \le B, \tag{9.171}$$

where the Lorentz parameter $\gamma \equiv (1 - v^2/c^2)^{-1/2}$ corresponds to the velocity (168) rather than that of the particle. These relations show that in this special reference frame, the particle only "sees" the re-normalized uniform magnetic field $B' \le B$, parallel to the initial field, i.e. normal to the velocity (168). Using the result of the above case (i), we see that in this frame the particle moves along either a circle or a spiral winding about the direction of the magnetic field, with the angular velocity (151):

$$\omega'_c = \frac{qB'}{\mathscr{E}'/c^2}, \tag{9.172}$$

and the radius (153):

$$R' = \frac{p'_\perp}{qB'}. \tag{9.173}$$

Hence in the lab frame, the particle performs this orbital/spiral motion plus a "drift" with the constant velocity **v** (Fig. 12). As a result, the lab-frame trajectory of the particle (or rather its projection onto the plane normal to the magnetic field) is a *trochoid*-like curve[58] that, depending on the initial velocity, may be either *prolate* (self-crossing), as in Fig. 12, or *curtate* (drift-stretched so much that it is not self-crossing).

Such looped motion of electrons is used, in particular, in *magnetrons* – very popular generators of microwave radiation. In such a device (Fig. 13), the magnetic field, usually created by specially-shaped permanent magnets, is nearly uniform (in the region of electron motion) and directed along the magnetron's axis (in Fig. 13, normal to the plane of the drawing), while the electric field of magnitude $E \ll cB$, created by the dc voltage applied between the anode and the cathode, is virtually radial.

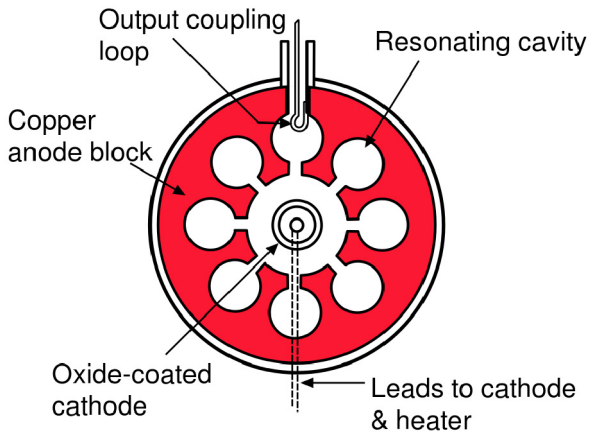


Fig. 9.13. Schematic cross-section of a typical magnetron. (Figure adapted from https://en.wikipedia.org/wiki/Cavity_magnetron under the Free GNU Documentation License.)

As a result, the above simple theory is only approximately valid, and the electron trajectories are close to *epicycloids* rather than trochoids. The applied electric field is adjusted so that these looped trajectories pass close to the anode's surface, and hence to the gap openings of the cylindrical microwave cavities drilled in the anode's bulk. The fundamental mode of such a cavity is quasi-lumped, with the cylindrical walls working mostly as inductances, and the gap openings as capacitances, with the microwave electric field concentrated in these openings. This is why the mode is strongly coupled to the electrons "licking" the anode's surface, and their interaction creates large positive feedback (equivalent to negative damping), which results in intensive microwave self-oscillations at the cavities' own frequency.[59] The oscillation energy, of course, is taken from the dc-field-accelerated electrons; due to this energy loss, the looped trajectory of each electron gradually moves closer to the anode and finally

[58] As a reminder, a trochoid may be described as the trajectory of a point on a rigid disk rolled along a straight line. It's canonical parametric representation is $x = \Theta + a\cos\Theta$, $y = a\sin\Theta$. (For $a > 1$, the trochoid is *prolate*, if $a < 1$, it is *curtate*, and if $a = 1$, it is called the *cycloid*.) Note, however, that for our problem, the trajectory in the lab frame is exactly trochoidal only in the non-relativistic limit $v \ll c$ (i.e. $E/c \ll B$).

[59] See, e.g., CM Sec. 5.4.

lands on its surface. The wide use of such generators (in particular, in microwave ovens, which operate in a narrow frequency band around 2.45 GHz, allocated for these devices to avoid their interference with wireless communication systems) is due to their simplicity and high (up to 65%) efficiency of the dc-to-rf energy transfer.

Case 2: $E/c > B$. In this case, the speed given by Eq. (168) would be above the speed of light, so let us introduce a reference frame moving with a different velocity,

$$\mathbf{v} = \frac{\mathbf{E}\times\mathbf{B}}{(E/c)^2}, \tag{9.174}$$

whose direction is the same as before (Fig. 12), and magnitude $v = c\times B/(E/c)$ is again below c. A calculation absolutely similar to the one performed above for Case 1, yields

$$E'_x = 0, \qquad E'_y = \gamma(E - vB) = \gamma E\left(1 - \frac{vB}{E}\right) = \gamma E\left(1 - \frac{v^2}{c^2}\right) = \frac{E}{\gamma} \le E, \qquad E'_z = 0, \tag{9.175}$$

$$B'_x = 0, \qquad B'_y = 0, \qquad B'_z = \gamma\left(B - \frac{vE}{c^2}\right) = \gamma\left(B - \frac{EB}{E}\right) = 0. \tag{9.176}$$

so in the moving frame the particle "sees" only the electric field $E' \le E$. According to the solution of our previous problem (ii), the trajectory of the particle in the moving frame is the catenary (165), so in the lab frame it has an "open", hyperbolic character as well.

To conclude this section, let me note that if the electric and magnetic fields are nonuniform, the particle motion may be much more complex, and in most cases, the integration of the system of equations (144) and (148) may be carried out only numerically. However, if the field's nonuniformity is small, approximate analytical methods may be very effective. For example, if $\mathbf{E} = 0$, and the magnetic field has a small *transverse* gradient ∇B in a direction normal to the vector $\mathbf{B}$ itself, such that

$$\eta \equiv \frac{|\nabla B|}{B} << \frac{1}{R}, \tag{9.177}$$

where R is the cyclotron radius (153), then it is straightforward to use Eq. (150) to show[60] that the cyclotron orbit drifts perpendicular to both $\mathbf{B}$ and ∇B, with the drift speed

$$v_{\rm d} \approx \frac{\eta}{\omega_{\rm c}}\left(\frac{1}{2}u_\perp^2 + u_{||}^2\right) << u. \tag{9.178}$$

The physics of this drift is rather simple: according to Eq. (153), the instant curvature of the cyclotron orbit is proportional to the local value of the field. Hence if the field is nonuniform, the trajectory bends slightly more on its parts passing through a stronger field, thus acquiring a shape close to a curate trochoid.

For experimental physics and engineering practice, the effects of *longitudinal* gradients of magnetic field on the charged particle motion are much more important, but it is more convenient for me to postpone their discussion until we have developed a little bit more analytical tools in the next section.

[60] See, e.g., Sec. 12.4 in J. Jackson, *Classical Electrodynamics*, 3rd ed., Wiley, 1999.

9.7. Analytical mechanics of charged particles

The general Eq. (145) gives a full description of relativistic particle dynamics in electric and magnetic fields, just as the 2nd Newton law (1) does it in the non-relativistic limit. However, we know that in the latter case, the Lagrange formalism of analytical mechanics allows an easier solution of many problems.[61] We can expect that to be true in relativistic mechanics as well, so let us expand the analysis of Sec. 3 (which was valid only for free particles) to particles in the field.

For a free particle, our main result was Eq. (68), which may be rewritten as

$$\gamma \mathscr{L} = -mc^2 , \tag{9.179}$$

with $\gamma \equiv (1 - u^2/c^2)^{-1/2}$, showing that the product on the left-hand side is Lorentz-invariant. How can the electromagnetic field affect this relation? In non-relativistic electrostatics, we could write

$$\mathscr{L} = T - U = T - q\phi . \tag{9.180}$$

However, in relativity, the scalar potential ϕ is just one component of the potential 4-vector (116). The only way to get from this full 4-vector a Lorentz-invariant contribution to $\gamma\mathscr{L}$, which would be also proportional to the first power of the particle's velocity (to account for the magnetic component of the Lorentz force), is evidently

$$\gamma \mathscr{L} = -mc^2 + \text{const} \times u^\alpha A_\alpha , \tag{9.181}$$

where u^α is the 4-velocity (63). To comply with Eq. (180) at $u << c$, the constant factor should be equal to $(-q)$, so Eq. (181) becomes

$$\gamma \mathscr{L} = -mc^2 - qu^\alpha A_\alpha , \tag{9.182}$$

and with the account of Eqs. (63) and the second of Eqs. (116), we get very important equality

$$\boxed{\mathscr{L} = -\frac{mc^2}{\gamma} - q\phi + q\mathbf{u}\cdot\mathbf{A} ,} \tag{9.183}$$

Particle's Lagrangian function

whose Cartesian form is

$$\mathscr{L} = -mc^2\left(1 - \frac{u_x^2 + u_y^2 + u_z^2}{c^2}\right)^{1/2} - q\phi + q\left(u_x A_x + u_y A_y + u_z A_z\right). \tag{9.184}$$

Let us see whether this relation (which admittedly was derived by an educated guess rather than by a strict derivation) passes a natural sanity check. For the case of an unconstrained motion of a particle, we can select its three Cartesian coordinates r_j (j = 1, 2, 3) as the generalized coordinates, and its linear velocity components u_j as the corresponding generalized velocities. In this case, the Lagrange equations of motion are

$$\frac{d}{dt}\frac{\partial \mathscr{L}}{\partial u_j} - \frac{\partial \mathscr{L}}{\partial r_j} = 0. \tag{9.185}$$

For example, for $r_1 = x$, Eq. (184) yields

$$\frac{\partial \mathscr{L}}{\partial u_x} = \frac{mu_x}{\left(1 - u^2/c^2\right)^{1/2}} + qA_x \equiv p_x + qA_x , \qquad \frac{\partial \mathscr{L}}{\partial x} = -q\frac{\partial \phi}{\partial x} + q\mathbf{u}\cdot\frac{\partial \mathbf{A}}{\partial x} , \tag{9.186}$$

[61] See, e.g., CM Sec. 2.2 and on.

so Eq. (185) takes the form

$$\frac{dp_x}{dt} = -q\frac{\partial \phi}{\partial x} + q\mathbf{u}\cdot\frac{\partial \mathbf{A}}{\partial x} - q\frac{dA_x}{dt}. \tag{9.187}$$

In the equations of motion, the field values have to be taken at the instant position of the particle, so the last (full) derivative has components due to both the actual field's change (at a fixed point of space) and the particle's motion. Such addition is described by the so-called *convective derivative*[62]

Convective derivative

$$\boxed{\frac{d}{dt} = \frac{\partial}{\partial t} + \mathbf{u}\cdot\nabla}. \tag{9.188}$$

Spelling out both scalar products, we may group the terms remaining after cancellations as follows:

$$\frac{dp_x}{dt} = q\left[\left(-\frac{\partial \phi}{\partial x} - \frac{\partial A_x}{\partial t}\right) + u_y\left(\frac{\partial A_y}{\partial x} - \frac{\partial A_x}{\partial y}\right) - u_z\left(\frac{\partial A_x}{\partial z} - \frac{\partial A_z}{\partial x}\right)\right]. \tag{9.189}$$

But taking into account the relations (121) between the electric and magnetic fields and potentials, this expression is nothing more than

$$\frac{dp_x}{dt} = q\left(E_x + u_y B_z - u_z B_y\right) = q\left(\mathbf{E} + \mathbf{u}\times\mathbf{B}\right)_x, \tag{9.190}$$

i.e. the x-component of Eq. (144). Since other Cartesian coordinates participate in Eq. (184) similarly, it is evident that the Lagrangian equations of motion along other coordinates yield other components of the same vector equation of motion.

So, Eq. (183) does indeed give the correct Lagrangian function, and we can use it for further analysis, in particular to discuss the first of Eqs. (186). This relation shows that in the electromagnetic field, the generalized momentum corresponding to the particle's coordinate x is *not* $p_x = m\gamma u_x$, but[63]

$$P_x \equiv \frac{\partial \mathscr{L}}{\partial u_x} = p_x + qA_x. \tag{9.191}$$

Thus, as was already discussed (at that point, without proof) in Sec. 6.4, the particle's motion in a magnetic field may be is described by two different linear momentum vectors: the *kinetic momentum* $\mathbf{p}$ defined by Eq. (70), and the *canonical* (or "conjugate") *momentum*[64]

Particle's canonical momentum

$$\boxed{\mathbf{P} = \mathbf{p} + q\mathbf{A}}. \tag{9.192}$$

In order to facilitate discussion of this notion, let us generalize Eq. (72) for the Hamiltonian function $\mathscr{H}$ of a free particle to the case of a particle in the field:

$$\mathscr{H} \equiv \mathbf{P}\cdot\mathbf{u} - \mathscr{L} = (\mathbf{p} + q\mathbf{A})\cdot\mathbf{u} - \left(-\frac{mc^2}{\gamma} + q\mathbf{u}\cdot\mathbf{A} - q\phi\right) \equiv \mathbf{p}\cdot\mathbf{u} + \frac{mc^2}{\gamma} + q\phi. \tag{9.193}$$

[62] Alternatively called the "Lagrangian derivative"; for its (rather simple) derivation see, e.g., CM Sec. 8.3.

[63] With regrets, I have to use for the generalized momentum the same (very common) notation as was used earlier in the course for the electric polarization – which will not be discussed here and in the balance of these notes.

[64] In the Gaussian units, Eq. (192) has the form $\mathbf{P} = \mathbf{p} + q\mathbf{A}/c$.

Merging the first two terms of the last expression exactly as it was done in Eq. (72), we get an extremely simple result,

$$\mathcal{H} = \gamma mc^2 + q\phi, \tag{9.194a}$$

which may be spelled out as

$$\mathcal{H} = \left[1 + \left(\frac{p}{mc}\right)^2\right]^{1/2} mc^2 + q\phi, \qquad \text{i.e. } (\mathcal{H} - q\phi)^2 = (mc^2)^2 + c^2 p^2. \tag{9.194b}$$

These expressions may leave the reader wondering: where is the vector potential **A** here – and the magnetic field effects it has to describe? The resolution of this puzzle is easy: as we know from analytical mechanics,[65] for most applications, for example for an alternative derivation of the equations of motion, $\mathcal{H}$ has to be represented as a function of the particle's generalized coordinates (in the case of unconstrained motion, these may be the Cartesian components of the vector **r** that serves as an argument for the potentials **A** and ϕ), and the generalized momenta, i.e. the components of the vector **P** – generally, plus time. For that, the kinematic momentum **p** in Eq. (194b) has to be expressed via these variables. This may be done using Eq. (192), giving us the following generalization of Eq. (78):[66]

$$(\mathcal{H} - q\phi)^2 = (mc^2)^2 + c^2(\mathbf{P} - q\mathbf{A})^2. \tag{9.195}$$

Particle's Hamiltonian function

It is straightforward to verify that the Hamilton equations of motion for three Cartesian coordinates of the particle, obtained in a regular way from this $\mathcal{H}$, may be merged into the same vector equation (144). In the non-relativistic limit, performing the expansion of Eqs. (194b) into the Taylor series in p^2, and limiting it to two leading terms, we get the following generalization of Eq. (74):

$$\mathcal{H} \approx mc^2 + \frac{p^2}{2m} + q\phi, \qquad \text{i.e. } \mathcal{H} - mc^2 \approx \frac{1}{2m}(\mathbf{P} - q\mathbf{A})^2 + U, \qquad \text{with } U = q\phi. \tag{9.196}$$

These expressions for $\mathcal{H}$, and Eq. (183) for $\mathcal{L}$, give a clear view of the electromagnetic field effects' description in analytical mechanics. The electric part $q\mathbf{E}$ of the total Lorentz force can perform mechanical work on the particle, i.e. change its kinetic energy – see Eq. (148) and its discussion. As a result, the scalar potential ϕ, whose gradient gives a contribution to **E**, may be directly associated with the potential energy $U = q\phi$ of the particle. On the contrary, the magnetic component $q\mathbf{u}\times\mathbf{B}$ of the Lorentz force is always perpendicular to the particle's velocity **u**, and cannot perform a non-zero work on it, and as a result, cannot be described by a contribution to U. However, if **A** did not participate in the functions $\mathcal{L}$ and/or $\mathcal{H}$ at all, the analytical mechanics would be unable to describe effects of the magnetic field $\mathbf{B} = \nabla\times\mathbf{A}$ on the particle's motion. The relations (183) and (195)-(196) show the wonderful way in which physics (with some help from Mother Nature herself :-) solves this problem: the vector potential gives such contributions to the functions $\mathcal{L}$ and $\mathcal{H}$ that cannot be uniquely attributed to either kinetic or potential energy, but ensure both the Lagrange and Hamilton formalisms yield the correct equation of motion (144) – including the magnetic field effects.

[65] See, e.g., CM Sec. 10.1.

[66] Alternatively, this relation may be obtained from the expression for the Lorentz-invariant norm, $p^\alpha p_\alpha = (mc)^2$, of the 4-momentum (75), $p^\alpha = \{\mathcal{E}/c, \mathbf{p}\} = \{(\mathcal{H} - q\phi)/c, \mathbf{P} - q\mathbf{A}\}$.

I believe I still owe the reader some discussion of the physical sense of the canonical momentum **P**. For that, let us consider a charged particle moving near a region of localized magnetic field **B**(**r**,*t*), but not entering this region (see Fig. 14), so on its trajectory $\nabla\times\mathbf{A} \equiv \mathbf{B} = 0$.

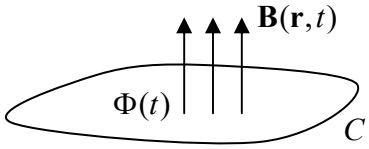


Fig. 9.14. Particle's motion around a localized magnetic field with a time-dependent flux.

If there is no electrostatic field affecting the particle (i.e. no other electric charges nearby), we may select such a local gauge that $\phi(\mathbf{r}, t) = 0$ and $\mathbf{A} = \mathbf{A}(t)$, so Eq. (144) is reduced to

$$\frac{d\mathbf{p}}{dt} = q\mathbf{E} = -q\frac{d\mathbf{A}}{dt}, \tag{9.197}$$

and Eq. (192) immediately gives

$$\frac{d\mathbf{P}}{dt} \equiv \frac{d\mathbf{p}}{dt} + q\frac{d\mathbf{A}}{dt} = 0. \tag{9.198}$$

Hence, even if the magnetic field is changed in time, so that the induced electric field **E** does accelerate the particle, its canonical momentum does not change. Hence **P** is a variable more stable to magnetic field changes than its kinetic counterpart **p**. This conclusion may be criticized because it relies on a specific gauge, and generally $\mathbf{P} \equiv \mathbf{p} + q\mathbf{A}$ is not gauge–invariant, because the vector potential **A** is not.[67] However, as was already discussed in Sec. 5.3, the integral $\int\mathbf{A}\cdot d\mathbf{r}$ over a closed contour *is* gauge-invariant and is equal to the magnetic flux Φ through the area limited by the contour – see Eq. (5.65). So, integrating Eq. (197) over a closed trajectory of a particle (Fig. 14), and over the time of one orbit, we get

$$\Delta\oint_C \mathbf{p}\cdot d\mathbf{r} = -q\Delta\Phi, \qquad \text{so that } \Delta\oint_C \mathbf{P}\cdot d\mathbf{r} = 0, \tag{9.199}$$

where ΔΦ is the change of flux during that time. This gauge-invariant result confirms the above conclusion about the stability of the canonical momentum to magnetic field variations.

Generally, Eq. (199) is invalid if a particle moves inside a magnetic field and/or changes its trajectory at the field variation. However, if the field is *almost* uniform, i.e. its gradient is small in the sense of Eq. (177), this result is (approximately) applicable. Indeed, analytical mechanics[68] tells us that for any canonical coordinate-momentum pair $\{q_j, p_j\}$, the corresponding *action variable*,

$$J_j \equiv \frac{1}{2\pi}\oint p_j dq_j, \tag{9.200}$$

remains virtually constant at slow variations of motion conditions. According to Eq. (191), for a particle in a magnetic field, the generalized momentum corresponding to the Cartesian coordinate r_j is P_j rather than p_j. Thus forming the net action variable $J \equiv J_x + J_y + J_z$, we may write

[67] In contrast, the kinetic momentum $\mathbf{p} = M\mathbf{u}$ is evidently gauge- (though not Lorentz-) invariant.
[68] See, e.g., CM Sec. 10.2.

$$2\pi J = \oint \mathbf{P}\cdot d\mathbf{r} = \oint \mathbf{p}\cdot d\mathbf{r} + q\Phi = \text{const}\,. \tag{9.201}$$

Let us apply this relation to the motion of a non-relativistic particle in an almost uniform magnetic field, with a relatively small longitudinal velocity, $u_{||}/u_{\perp} \to 0$ – see Fig. 15.

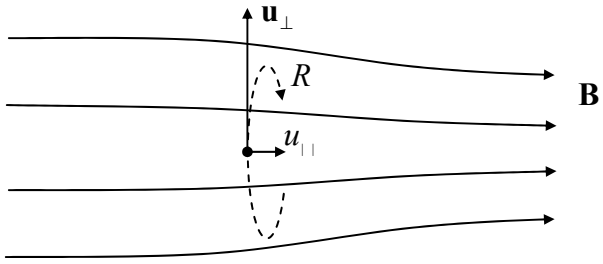


Fig. 9.15. Particle in a magnetic field with a small longitudinal gradient $\nabla B \,||\, \mathbf{B}$.

In this case, Φ in Eq. (201) is the flux encircled by the particle's cyclotron orbit, $\Phi = -\pi R^2 B$, where R is its radius given by Eq. (153), and the negative sign accounts for the fact that in our case, the "correct" direction of the normal vector $\mathbf{n}$ in the definition of flux, $\Phi = \int \mathbf{B}\cdot\mathbf{n} d^2 r$, is antiparallel to the vector $\mathbf{B}$. At $u << c$, the kinetic momentum is just $p_{\perp} = mu_{\perp}$, while Eq. (153) yields

$$mu_{\perp} = qBR\,. \tag{9.202}$$

Plugging these relations into Eq. (201), we get

$$2\pi J = mu_{\perp} 2\pi R - q\pi R^2 B = m\frac{qRB}{m} 2\pi R - q\pi R^2 B \equiv (2-1)q\pi R^2 B \equiv -q\Phi\,. \tag{9.203}$$

This means that even if the circular orbit slowly moves through the magnetic field, the flux encircled by the cyclotron orbit should remain virtually constant. One manifestation of this effect is the result already mentioned at the end of Sec. 6: if a small gradient of the magnetic field is perpendicular to the field itself, then the particle orbit's drift direction is perpendicular to ∇B, so Φ stays constant.

Now let us analyze the case of a small longitudinal gradient, $\nabla B \,||\, \mathbf{B}$ (Fig. 15). If a small initial longitudinal velocity $u_{||}$ is directed toward the higher field region, the cyclotron orbit has to gradually shrink to keep Φ constant. Rewriting Eq. (202) as

$$mu_{\perp} = q\frac{\pi R^2 B}{\pi R} = q\frac{|\Phi|}{\pi R}\,, \tag{9.204}$$

we see that this reduction of R (at constant Φ) increases the orbiting speed $u_{\perp}$. But since the magnetic field cannot perform any work on the particle, its kinetic energy,

$$\mathscr{E} = \frac{m}{2}\left(u_{||}^2 + u_{\perp}^2\right), \tag{9.205}$$

should stay constant, so the longitudinal velocity $u_{||}$ has to decrease. Hence eventually the orbit's drift has to stop, and then it has to start moving back toward the region of lower fields, being essentially repulsed from the high-field region. This effect is very important, in particular, for plasma confinement systems. In the simplest of such systems, two coaxial magnetic coils, inducing magnetic fields of the same direction (Fig. 16), naturally form a "magnetic bottle", which traps charged particles injected, with sufficiently low longitudinal velocities, into the region between the coils. More complex systems of this

type, but working on the same basic principle, are the most essential components of the persisting large-scale efforts to achieve controllable nuclear fusion.[69]

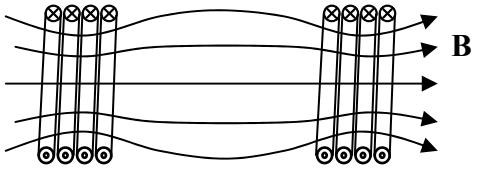


Fig. 9.16. A simple magnetic bottle (schematically).

Returning to the constancy of the magnetic flux encircled by free particles, it reminds us of the Meissner-Ochsenfeld effect, which was discussed in Sec. 6.4, and gives a motivation for a brief revisit of the electrodynamics of superconductivity. As was emphasized in that section, superconductivity is a substantially quantum phenomenon; nevertheless, the classical notion of the conjugate momentum **P** helps to understand its theoretical description. Indeed, the general rule of quantization of physical systems[70] is that each canonical pair $\{q_j, p_j\}$ of a generalized coordinate q_j and the corresponding generalized momentum p_j is described by quantum-mechanical operators that obey the following commutation relation:

$$\left[\hat{q}_j, \hat{p}_{j'}\right] = i\hbar\delta_{jj'}. \tag{9.206}$$

According to Eq. (191), for the Cartesian coordinates r_j of a particle in the magnetic field, the corresponding generalized momenta are P_j, so their operators should obey the similar commutation relations:

$$\left[\hat{r}_j, \hat{P}_{j'}\right] = i\hbar\delta_{jj'}. \tag{9.207}$$

In the coordinate representation of quantum mechanics, the canonical operators of the Cartesian components of the linear momentum are described by the corresponding components of the vector operator $-i\hbar\nabla$. As a result, ignoring the rest energy mc^2 (which gives an inconsequential phase factor $\exp\{-imc^2t/\hbar\}$ in the wavefunction), we can use Eq. (196) to rewrite the usual non-relativistic Schrödinger equation,

$$i\hbar\frac{\partial\psi}{\partial t} = \hat{\mathscr{H}}\psi, \tag{9.208}$$

as follows:

$$i\hbar\frac{\partial\psi}{\partial t} = \left(\frac{\hat{p}^2}{2m} + U\right)\psi \equiv \left[\frac{1}{2m}\left(-i\hbar\nabla - q\mathbf{A}\right)^2 + q\phi\right]\psi. \tag{9.209}$$

Thus, I believe I have finally delivered on my promise to justify the replacement (6.50), which had been used in Secs. 6.4 and 6.5 to discuss the electrodynamics of superconductors, including the Meissner-Ochsenfeld effect. The Schrödinger equation (209) may be also used as the basis for the quantum-mechanical description of other magnetic field phenomena, including the so-called Aharonov-Bohm and quantum Hall effects – see, e.g., QM Secs. 3.1-3.2.

[69] For further reading on this technology, the reader may be referred, for example, to the simple monograph by F. Chen, *Introduction to Plasma Physics and Controllable Fusion*, vol. 1, 2nd ed., Springer, 1984, and/or the graduate-level theoretical treatment by R. Hazeltine and J. Meiss, *Plasma Confinement*, Dover, 2003.

[70] See, e.g., CM Sec. 10.1.

9.8. Analytical mechanics of the electromagnetic field

We have just seen that the analytical mechanics of a *particle* in an electromagnetic field may be used to get some important results. The same is true for the analytical mechanics of the electromagnetic *field* as such, and the *field-particle system* as a whole. For such space-distributed systems as fields, governed by local dynamics laws (in our case, the Maxwell equations), we need to apply analytical mechanics to the *local densities* $\mathcal{l}$ and $\mathcal{h}$ of the Lagrangian and Hamiltonian functions, defined by relations

$$\mathscr{L} = \int \mathcal{l}\, d^3r, \quad \mathscr{H} = \int \mathcal{h}\, d^3r. \tag{9.210}$$

Let us start, as usual, from the Lagrange formalism. Some clues on the possible structure of the Lagrangian function density $\mathcal{l}$ may be obtained from that of the particle-field interaction description in this formalism, discussed in the last section. As we have seen, for the case of a single particle, the interaction is described by the last two terms of Eq. (183):

$$\mathscr{L}_{\text{int}} = -q\phi - q\mathbf{u}\cdot\mathbf{A}. \tag{9.211}$$

Obviously, if the charge q is continuously distributed over some volume, we may represent this $\mathscr{L}_{\text{int}}$ as a volume integral of the following Lagrangian function density:

$$\boxed{\mathcal{l}_{\text{int}} = -\rho\phi + \mathbf{j}\cdot\mathbf{A} \equiv -j_\alpha A^\alpha.} \tag{9.212}$$

Interaction Lagrangian density

Notice that this density (in contrast to $\mathscr{L}_{\text{int}}$ itself!) is Lorentz-invariant. (This is due to the contraction of the longitudinal coordinate, and hence volume, at the Lorentz transform.) Hence we may expect the density of the *field's* part of the Lagrangian to be Lorentz-invariant as well. Moreover, given the local structure of the Maxwell equations (containing only the first spatial and temporal derivatives of the fields), $\mathcal{l}_{\text{field}}$ should be a function of the potential's 4-vector and its 4-derivative:

$$\mathcal{l}_{\text{field}} = \mathcal{l}_{\text{field}}\left(A^\alpha, \partial_\alpha A^\beta\right). \tag{9.213}$$

Also, the density should be selected in such a way that the 4-vector analog of the Lagrangian equation of motion,

$$\partial_\alpha \frac{\partial \mathcal{l}_{\text{field}}}{\partial\left(\partial_\alpha A^\beta\right)} - \frac{\partial \mathcal{l}_{\text{field}}}{\partial A^\beta} = 0, \tag{9.214}$$

gave us the correct inhomogeneous Maxwell equations (127).[71] The field part $\mathcal{l}_{\text{field}}$ of the total Lagrangian density $\mathcal{l}$ should be a scalar and a quadratic form of the field strengths, i.e. of the tensor $F^{\alpha\beta}$, so the natural choice is

$$\mathcal{l}_{\text{field}} = \text{const}\times F_{\alpha\beta}F^{\alpha\beta}. \tag{9.215}$$

with the implied summation over both indices. Indeed, adding to this expression the interaction Lagrangian (212),

$$\mathcal{l} = \mathcal{l}_{\text{field}} + \mathcal{l}_{\text{int}} = \text{const}\times F_{\alpha\beta}F^{\alpha\beta} - j_\alpha A^\alpha, \tag{9.216}$$

[71] Here the implicit summation over the index α plays a role similar to the convective derivative (188) in replacing the full derivative over time, in a way that reflects the symmetry of time and space in special relativity. I do not want to spend more time justifying Eq. (214), because of the reasons that will be clear imminently.

and performing the differentiations, we see that Eqs. (214)-(215) indeed yield Eqs. (127), provided that the constant factor equals $(-1/4\mu_0)$.[72] So, the field's Lagrangian density is

Field's Lagrangian density

$$\mathcal{l}_{\text{field}} = -\frac{1}{4\mu_0} F_{\alpha\beta} F^{\alpha\beta} = \frac{1}{2\mu_0}\left(\frac{E^2}{c^2} - B^2\right) \equiv \frac{\varepsilon_0}{2} E^2 - \frac{B^2}{2\mu_0} \equiv u_{\text{e}} - u_{\text{m}}, \tag{9.217}$$

where u_{e} is the electric field energy density (1.65), and u_{m} is the magnetic field energy density (5.57). Let me hope the reader agrees that Eq. (217) is a wonderful result because the Lagrangian function has a structure absolutely similar to the well-known expression $\mathcal{L} = T - U$ of classical mechanics. So, for the field alone, the "potential" and "kinetic" energies are separable again.[73]

Now let us explore whether we can calculate the 4-form of the field's Hamiltonian function $\mathcal{H}$. In the generic analytical mechanics,

$$\mathcal{H} = \sum_j \frac{\partial \mathcal{L}}{\partial \dot{q}_j} \dot{q}_j - \mathcal{L}. \tag{9.218}$$

However, just as for the Lagrangian function, for a field we should find the spatial density $\mathcal{h}$ of the Hamiltonian, defined by the second of Eqs. (210), for which the natural 4-form of Eq. (218) is

$$\mathcal{h}^{\alpha\beta} = \frac{\partial \mathcal{l}}{\partial(\partial_\alpha A^\gamma)} \partial^\beta A^\gamma - g^{\alpha\beta} \mathcal{l}. \tag{9.219}$$

Calculated for the field alone, i.e. using Eq. (217) for $\mathcal{l}$, this definition yields

$$\mathcal{h}^{\alpha\beta}_{\text{field}} = \theta^{\alpha\beta} - \tau_D^{\alpha\beta}, \tag{9.220}$$

where the tensor

Symmetric energy-momentum tensor

$$\theta^{\alpha\beta} \equiv \frac{1}{\mu_0}\left(g^{\alpha\gamma} F_{\gamma\delta} F^{\delta\beta} + \frac{1}{4} g^{\alpha\beta} F_{\gamma\delta} F^{\gamma\delta}\right), \tag{9.221}$$

is gauge-invariant, while the remaining term,

$$\tau_D^{\alpha\beta} \equiv \frac{1}{\mu_0} g^{\alpha\gamma} F_{\gamma\delta} \partial^\delta A^\beta, \tag{9.222}$$

is not, so it cannot correspond to any measurable variables. Fortunately, it is straightforward to verify that the last tensor may be represented in the form

$$\tau_D^{\alpha\beta} = \frac{1}{\mu_0} \partial_\gamma \left(F^{\gamma\alpha} A^\beta\right), \tag{9.223}$$

and as a result, obeys the following relations:

$$\partial_\alpha \tau_D^{\alpha\beta} = 0, \qquad \int \tau_D^{0\beta} d^3 r = 0, \tag{9.224}$$

[72] In the Gaussian units, this coefficient is $(-1/16\pi)$.

[73] Since the Lagrange equations of motion are homogeneous, the simultaneous change of the signs of T and U does not change them. Thus, it is not important which of the two energy densities, u_{e} or u_{m}, we count as the potential energy, and which as the kinetic energy. (Actually, such duality of the two energy components is typical for all analytical mechanics – see, e.g., the discussion of this issue in CM Sec. 2.2.)

so it does not interfere with the conservation properties of the gauge-invariant, symmetric *energy-momentum tensor* (also called the *symmetric stress tensor*) $\theta^{\alpha\beta}$, to be discussed below.

Let us use Eqs. (125) to express the elements of the latter tensor via the electric and magnetic fields. For $\alpha = \beta = 0$, we get

$$\theta^{00} = \frac{\varepsilon_0}{2} E^2 + \frac{B^2}{2\mu_0} = u_{\rm e} + u_{\rm m} \equiv u\,, \tag{9.225}$$

i.e. the expression for the total energy density u – see Eq. (6.113). The other 3 elements of the same row/column turn out to be just the Cartesian components of the Poynting vector (6.114), divided by c:

$$\theta^{j0} = \frac{1}{\mu_0}\left(\frac{\mathbf{E}}{c}\times\mathbf{B}\right)_j = \left(\frac{\mathbf{E}}{c}\times\mathbf{H}\right)_j \equiv \frac{S_j}{c}, \qquad \text{for } j = 1, 2, 3\,. \tag{9.226}$$

The remaining 9 elements $\theta_{jj'}$ of the tensor, with j, $j' = 1, 2, 3$, are usually represented as

$$\theta^{jj'} = -\tau^{(\mathrm{M})}_{jj'}, \tag{9.227}$$

where $\tau^{(\mathrm{M})}$ is the so-called *Maxwell stress tensor*:

$$\tau^{(\mathrm{M})}_{jj'} = \varepsilon_0\left(E_j E_{j'} - \frac{\delta_{jj'}}{2}E^2\right) + \frac{1}{\mu_0}\left(B_j B_{j'} - \frac{\delta_{jj'}}{2}B^2\right), \tag{9.228}$$

Maxwell stress tensor

so the whole symmetric energy-momentum tensor (221) may be conveniently represented in the following symbolic way:

$$\theta^{\alpha\beta} = \left(\begin{array}{c|c} u & \leftarrow\ \mathbf{S}/c\ \rightarrow \\ \hline \uparrow & \\ \mathbf{S}/c & -\tau^{(\mathrm{M})}_{jj'} \\ \downarrow & \end{array}\right). \tag{9.229}$$

The physical meaning of this tensor may be revealed in the following way. Considering Eq. (221) as the *definition* of the tensor $\theta^{\alpha\beta}$,[74] and using the 4-vector form of Maxwell equations given by Eqs. (127) and (129), it is straightforward to verify an extremely simple result for the 4-derivative of the symmetric tensor:

$$\partial_\alpha \theta^{\alpha\beta} = -F^{\beta\gamma} j_\gamma\,. \tag{9.230}$$

This expression is valid in the presence of electromagnetic field sources, e.g., for any system of charged particles and the fields they have created. Of these four equations (for four values of the index β), the temporal one (with $\beta = 0$) may be simply expressed via the energy density (225) and the Poynting vector (226):

$$\frac{\partial u}{\partial t} + \nabla\cdot\mathbf{S} = -\mathbf{j}\cdot\mathbf{E}\,, \tag{9.231}$$

while three spatial equations (with $\beta = j = 1, 2, 3$) may be represented in the form

[74] In this way, we are using Eq. (219) just as a useful guess, which has led us to the definition of $\theta^{\alpha\beta}$, and may leave its strict justification for more in-depth field theory courses.

$$\frac{\partial}{\partial t}\frac{S_j}{c^2}-\sum_{j'=1}^{3}\frac{\partial}{\partial r_{j'}}\tau_{jj'}^{(\mathrm{M})}=-\left(\rho\mathbf{E}+\mathbf{j}\times\mathbf{B}\right)_j. \tag{9.232}$$

If integrated over a volume V limited by surface S, with the account of the divergence theorem, Eq. (231) returns us to the Poynting theorem (6.111):

$$\int_V\left(\frac{\partial u}{\partial t}+\mathbf{j}\cdot\mathbf{E}\right)d^3r+\oint_S S_n d^2r=0, \tag{9.233}$$

while Eq. (232) yields[75]

$$\int_V\left[\frac{\partial}{\partial t}\frac{\mathbf{S}}{c^2}+\mathbf{f}\right]_j d^3r=\sum_{j'=1}^{3}\oint_S\tau_{jj'}^{(\mathrm{M})}dA_{j'},\quad\text{with }\mathbf{f}\equiv\rho\mathbf{E}+\mathbf{j}\times\mathbf{B}, \tag{9.234}$$

where $dA_j = n_j dA = n_j d^2r$ is the j^{th} component of the elementary area vector $d\mathbf{A} = \mathbf{n}dA = \mathbf{n}d^2\mathrm{r}$ that is normal to the volume's surface, and directed out of the volume – see Fig. 17.[76]

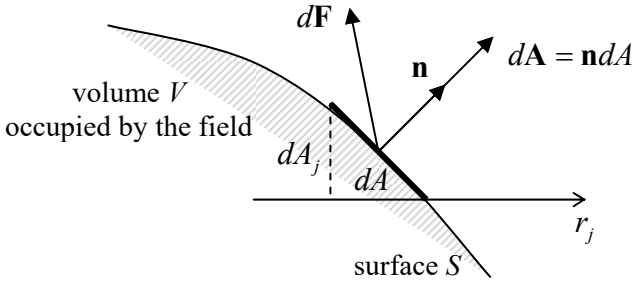


Fig. 9.17. The force $d\mathbf{F}$ exerted on a boundary element $d\mathbf{A}$ of the volume V occupied by the field.

Since, according to Eq. (5.10), the vector $\mathbf{f}$ in Eq. (234) is nothing other than the density of volume-distributed Lorentz forces exerted by the field on the charged particles, we can use the 2nd Newton law, in its relativistic form (144), to rewrite Eq. (234), for a stationary volume V, as

Field momentum's dynamics

$$\boxed{\frac{d}{dt}\left[\int_V\frac{\mathbf{S}}{c^2}d^3r+\mathbf{p}_{\text{part}}\right]=\mathbf{F},} \tag{9.235}$$

where $\mathbf{p}_{\text{part}}$ is the total mechanical (relativistic) momentum of all particles in the volume V, and the vector $\mathbf{F}$ is defined by its Cartesian components:

Force via the Maxwell tensor

$$\boxed{F_j=\sum_{j'=1}^{3}\oint_S\tau_{jj'}^{(\mathrm{M})}dA_{j'}.} \tag{9.236}$$

Relations (235)-(236) are our main new results. The first of them shows that the vector

[75] Just like the Poynting theorem (233), Eq. (234) may be obtained directly from the Maxwell equations, without resorting to the 4-vector formalism – see, e.g., Sec. 8.2.2 in D. Griffiths, *Introduction to Electrodynamics*, 3rd ed., Prentice-Hall, 1999. However, the derivation discussed above is superior because it shows the wonderful unity between the laws of conservation of energy and momentum.

[76] The same notions are used in the mechanical stress theory – see, e.g., CM Sec. 7.2.

$$\mathbf{g} \equiv \frac{\mathbf{S}}{c^2}, \tag{9.237}$$

already discussed in Sec. 6.8 without derivation, may be indeed interpreted as the density of momentum of the electromagnetic field (per unit volume). This classical relation is consistent with the quantum-mechanical picture of photons as ultra-relativistic particles, with a momentum of magnitude $\mathscr{E}/c$, because then the flux of the momentum carried by photons through a unit normal area per unit time may be represented either as S_n/c or as $g_n c$. It also allows us to revisit the Poynting vector paradox that was discussed in Sec. 6.8 – see Fig. 611 and its discussion. As was emphasized in that discussion, in this case, the vector $\mathbf{S} = \mathbf{E}\times\mathbf{H}$ does not correspond to any measurable energy flow. However, the corresponding momentum of the field, equal to the integral of the density (237) over a volume of interest,[77] is not only real but may be measured by the recoil impulse it gives to the field sources – say, to the magnetic coil inducing the field **H**, or to the capacitor plates creating the field **E**.

Now let us turn to our second result, Eq. (236). It tells us that the 3×3-element Maxwell stress tensor complies with the general definition of the stress tensor[78] characterizing the forces exerted on the boundaries of a volume, in our current case the volume occupied by the electromagnetic field (Fig. 17). Let us use this important result to analyze two simple examples of static fields.

(i) *Electrostatic field's* effect on a perfect conductor. Since Eq. (235) has been derived for a free space region, we have to select volume V outside the conductor, but we may align one of its faces with the conductor's surface (Fig. 18).

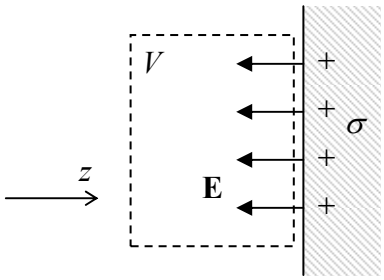


Fig. 9.18. The electrostatic field near a conductor's surface.

From Chapter 2, we know that the electrostatic field just outside the conductor's surface has to be normal to it. Selecting the z-axis in this direction, we have $E_x = E_y = 0$, $E_z = \pm E$, so only diagonal elements of the tensor (228) are not equal to zero:

$$\tau_{xx}^{(\mathrm{M})} = \tau_{yy}^{(\mathrm{M})} = -\frac{\varepsilon_0}{2}E^2, \qquad \tau_{zz}^{(\mathrm{M})} = \frac{\varepsilon_0}{2}E^2, \tag{9.238}$$

Since the elementary surface area vector has just one non-zero component, dA_z, according to Eq. (236), only the last component (that is positive regardless of the sign of E) gives a contribution to the surface force **F**. We see that the force exerted *by the conductor* (and eventually by the external forces that hold the conductor in its equilibrium position) on the field is normal to the conductor and directed out of the field volume: $dF_z \geq 0$. Hence, by the 3rd Newton law, the force exerted *by the field* on the conductor's surface is directed toward the field-filled space:

[77] It is sometimes called *hidden momentum*.
[78] See, e.g., CM Sec. 7.2.

Electric field's pull

$$dF_{\text{surface}} = -dF_z = -\frac{\varepsilon_0}{2}E^2 dA\,. \tag{9.239}$$

This important result could be obtained by simpler means as well. (Actually, this was the task of one of the exercise problems assigned in Chapter 2.) For example, one could argue, quite convincingly, that the local relation between the force and the field should not depend on the global configuration creating the field, and thus consider the simplest configuration, a planar capacitor (see, e.g. Fig. 2.3) with surfaces of both plates charged by equal and opposite charges of density $\sigma = \pm\varepsilon_0 E$. According to the Coulomb law, the charges should attract each other, pulling each plate toward the field region, so the Maxwell-tensor result gives the correct direction of the force. Now the force's magnitude given by Eq. (239) may be verified either by the direct integration of the Coulomb law or by the following simple reasoning. In the plane capacitor, the inner field $E_z = \sigma/\varepsilon_0$ is equally contributed by two surface charges; hence the field created by the negative charge of the counterpart plate (not shown in Fig. 18) is $E_- = -\sigma/2\varepsilon_0$, and the force it exerts of the elementary surface charge $dQ = \sigma dA$ of the positively charged plate is $dF_{\text{surface}} = EdQ = -\sigma^2 dA/2\varepsilon_0 = \varepsilon_0 E^2 dA/2$, in accordance with Eq. (239).[79]

Quantitatively, even for such a high electric field as $E = 10^5$ V/m (close to the electric breakdown's threshold in the air at a frequency of 10 GHz[80]), the "negative pressure" (dF/dA) given by Eq. (239) is of the order of 0.05 Pa (N/m^2), i.e. many orders below the ambient atmospheric pressure of 1 bar $\approx 10^5$ Pa. Still, this negative pressure may be substantial (well above 1 bar) in some cases, for example in good dielectrics (such as the high-quality SiO_2 grown at high temperature, which is broadly used in integrated circuits), which can withstand electric fields up to ~10^9 V/m.

(ii) *Static magnetic field's* effect on its source[81] – say a solenoid's wall or a superconductor's surface (Fig. 19). With the Cartesian coordinates' choice shown in that figure, we have $B_x = B$, $B_y = B_z = 0$, so the Maxwell stress tensor (228) is diagonal again:

$$\tau_{xx}^{(\text{M})} = \frac{1}{2\mu_0}B^2, \qquad \tau_{yy}^{(\text{M})} = \tau_{zz}^{(\text{M})} = -\frac{1}{2\mu_0}B^2\,. \tag{9.240}$$

However, since for this geometry, only dA_z differs from 0 in Eq. (236), the sign of the resulting force is opposite to that in electrostatics: $dF_z \le 0$, and the force exerted by the magnetic field upon the conductor's surface,

Magnetic field's push

$$dF_{\text{surface}} = -dF_z = \frac{1}{2\mu_0}B^2 dA\,, \tag{9.241}$$

[79] By the way, repeating these arguments for a plane capacitor filled with a linear dielectric, we may readily see that Eq. (239) may be generalized for this case by replacing ε_0 with ε. A similar replacement ($\mu_0 \to \mu$) is valid for Eq. (241) in a linear magnetic medium.

[80] Note that the breakdown field E_t in is a strong function of frequency. In the ambient air, it drops from its dc value of ~3×10^6 V/m to ~1.5×10^5 V/m at microwave frequencies and then rises to as much as ~6×10^9 V/m at optical frequencies. The reason of the rise is that at very high frequencies, the amplitude of the field-induced oscillations of the rare free electrons becomes much smaller than their mean free path, inhibiting the bulk impact-ionization of neutral atoms. (Because of this reason, E_t also depends on the air's pressure.)

[81] The causal relation is not important here. Especially in the case of a superconductor, the magnetic field may be induced by another source, with the surface supercurrent **j** just shielding the superconductor's bulk from its penetration – see Sec. 6.

corresponds to positive pressure. For good laboratory magnets ($B \sim 10$ T), this pressure is of the order of 4×10^7 Pa $\approx$ 400 bars, i.e. is very substantial, so the magnets require solid mechanical design.

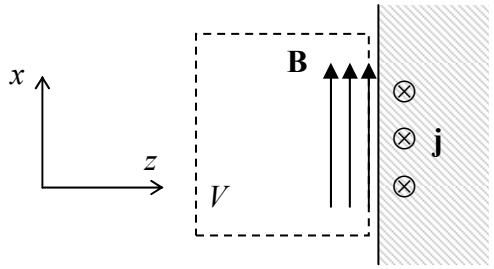


Fig. 9.19. The magnetostatic field near a current-carrying surface.

The direction of the force (241) could be also readily predicted using elementary magnetostatics arguments. Indeed, we can imagine the magnetic field volume limited by another, parallel wall with the opposite direction of surface current. According to the starting point of magnetostatics, Eq. (5.1), such surface currents of opposite directions have to repulse each other – doing that via the magnetic field.

Another explanation of the fundamental sign difference between the electric and magnetic field pressures may be provided using the electric circuit language. As we know from Chapter 2, the potential energy of the electric field stored in a capacitor may be represented in two equivalent forms,

$$U_{\rm e} = \frac{CV^2}{2} = \frac{Q^2}{2C}. \tag{9.242}$$

Similarly, the magnetic field energy of an inductive coil is

$$U_{\rm m} = \frac{LI^2}{2} = \frac{\Phi^2}{2L}. \tag{9.243}$$

If we do not want to consider the work of external sources at a virtual change of the system dimensions, we should use the last forms of these relations, i.e. consider a galvanically detached capacitor (Q = const) and an externally-shorted inductance (Φ = const).[82] Now if we let the electric field forces (239) drag the capacitor's plates in the direction they "want", i.e. toward each other, this would lead to a *reduction* of the capacitor thickness, and hence to an *increase* of its capacitance C, and hence to a *decrease* of $U_{\rm e}$. Similarly, for a solenoid, allowing the positive pressure (241) to move its walls from each other would lead to an *increase* of the solenoid's volume, and hence of its inductance L, so the potential energy $U_{\rm m}$ would be also *reduced* – as it should be. It is remarkable (actually, beautiful!) how the local field formulas (239) and (241) "know" about these global circumstances.

Finally, let us see whether the major results (237) and (241) obtained in this section, match each other. For that, let us return to the normal incidence of a plane, monochromatic wave from the free space upon the plane surface of a perfect conductor (see, e.g., Fig. 7.8 and its discussion), and use those results to calculate the time average of the pressure $dF_{\rm surface}/dA$ imposed by the wave on the surface. At elastic reflection from the conductor's surface, the electromagnetic field's momentum retains its amplitude but reverses its sign, so the average momentum transferred to a unit area of the surface in a unit time (i.e. the average pressure) is

[82] Of course, this condition may hold "forever" only for solenoids with superconducting wiring, but even in normal-metal solenoids with practicable inductances, the flux relaxation constants L/R may be rather large (practically, up to a few minutes), quite sufficient to carry out the force measurement.

$$\overline{\frac{dF_{\text{surface}}}{dA}} = 2cg_{\text{incident}} = 2c\frac{\overline{S_{\text{incident}}}}{c^2} = 2c\frac{\overline{EH}}{c^2} = E_\omega H_\omega^*, \tag{9.244}$$

where E_ω and H_ω are complex amplitudes of the incident wave. Using the relation (7.7) between these amplitudes (for $\varepsilon = \varepsilon_0$ and $\mu = \mu_0$ giving $E_\omega = cB_\omega$), we get

$$\overline{\frac{dF_{\text{surface}}}{dA}} = \frac{1}{c}cB_\omega\frac{B_\omega^*}{\mu_0} \equiv \frac{|B_\omega|^2}{\mu_0}. \tag{9.245}$$

On the other hand, as was discussed in Sec. 7.3, at the surface of a perfect mirror the electric field vanishes while the magnetic field doubles, so we can use Eq. (241) with $B \to B(t) = 2\text{Re}[B_\omega \exp\{-i\omega t\}]$. Averaging the pressure given by Eq. (241) over time, we get

$$\overline{\frac{dF_{\text{surface}}}{dA}} = \frac{1}{2\mu_0}\overline{\left(2\,\text{Re}\left[B_\omega e^{-i\omega t}\right]\right)^2} = \frac{|B_\omega|^2}{\mu_0}, \tag{9.246}$$

i.e. the same result as Eq. (245).

For physics intuition development, it is useful to evaluate the electromagnetic radiation pressure. Even for a relatively high wave intensity S_n of 1 kW/m^2 (close to that of the direct sunlight at the Earth's surface), the pressure $2cg_n = 2S_n/c$ is somewhat below 10^{-5} Pa ~ 10^{-10} bar. Still, this extremely small effect was experimentally observed (by P. Lebedev) as early as 1899, giving one more confirmation of Maxwell's theory. Currently, there are ongoing attempts to use the pressure of the Sun's light for propelling small spacecraft, e.g., the *LightSail 2* satellite with a 32-m^2 sail, launched in 2019.

9.9. Exercise problems

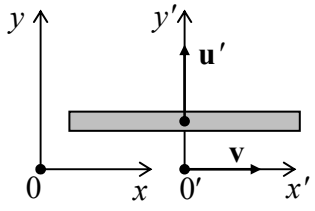
9.1. Use the pre-relativistic picture of light propagation with velocity c in a Sun-bound aether to derive Eq. (4).

9.2. Show that two successive Lorentz space/time transforms, with velocities u' and v in the same direction, are equivalent to the single transform with the velocity u given by Eq. (25).

9.3. $N + 1$ reference frames numbered by index n (taking values 0, 1, …, N) move in the same direction as a particle. Express the particle's velocity in the frame number 0 via its velocity u_N in the frame number N and the velocities v_n of the frame number n relative to the frame number $(n - 1)$.

9.4. A spaceship moving with a constant velocity v directly from the Earth sends back brief flashes of light with a period Δt_s – as measured by the spaceship's clock. Calculate the period with that an Earth-based observer may receive these signals – as measured by their clock.

9.5. From the point of view of observers in a reference frame 0′, a straight thin rod, parallel to the x'-axis, is moving without rotation with a constant velocity $\mathbf{u}'$ directed along the y'-axis. The reference frame 0′ is itself moving relative to another ("lab") reference frame 0 with a constant velocity $\mathbf{v}$ along the x-axis, also without rotation – see the figure on the right.



Calculate:

(i) the direction of the rod's velocity, and
(ii) the orientation of the rod on the [x, y] plane,

– both as observed from the lab reference frame. Is the velocity, in this frame, perpendicular to the rod?

9.6. Starting from the rest at $t = 0$, a spaceship moves directly from the Earth, with a constant acceleration as measured in its *instantaneous rest frame*. Find its displacement $x(t)$ from the Earth, as measured from the Earth's reference frame, and interpret the result.

Hint: The instantaneous rest frame of a moving particle is the inertial reference frame that, at the considered moment of time, has the same velocity as the particle.

9.7. Analyze the twin paradox for the simplest case of 1D travel with a piecewise-constant acceleration.

Hint: You may use an intermediate result of the solution of the previous problem.

9.8. Suggest a natural definition of the 4-vector of acceleration (commonly called the *4-acceleration*) of a point and calculate its components for of a relativistic point moving with velocity $\mathbf{u} = \mathbf{u}(t)$.

9.9. Calculate the first relativistic correction to the frequency of a harmonic oscillator as a function of its amplitude.

9.10. An atom with an initial rest mass m has been excited to an internal state with an additional energy $\Delta\mathscr{E}$, while still being at rest. Next, it returns to its initial state, emitting a photon. Calculate the photon's frequency, taking into account the relativistic recoil of the atom.

Hint: In this problem, and also in Problems 13-15 below, you may treat photons as classical ultra-relativistic point particles with zero rest mass, energy $\mathscr{E} = \hbar\omega$, and momentum $\mathbf{p} = \hbar\mathbf{k}$.

9.11. A particle of mass m, initially at rest, decays into two particles with rest masses m_1 and m_2. Calculate the total energy of the first product particle, in the c.o.m. reference frame.

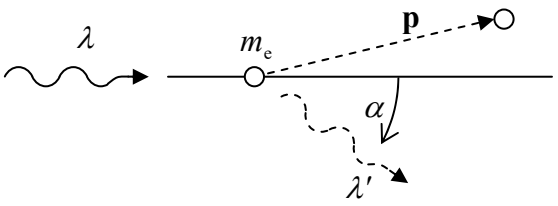
9.12. A relativistic particle with a rest mass m, moving with velocity u, decays into two particles with zero rest mass.

(i) Calculate the smallest possible angle between the decay product velocities (in the lab frame, in that the velocity u is measured).
(ii) What is the largest possible energy of one product particle?

9.13. A relativistic particle flying in free space with velocity $\mathbf{u}$ decays into two photons.[83] Calculate the angular dependence of the photon detection probability, as measured in the lab frame.

[83] Such a decay may happen, for example, with a neutral pion.

9.14. A photon with wavelength λ is scattered by an electron, initially at rest. Calculate the wavelength λ' of the scattered photon as a function of the scattering angle α – see the figure on the right.[84]



9.15. Calculate the threshold energy of a γ-photon for the reaction

$$\gamma + p \to p + \pi^0,$$

if the proton was initially at rest.

Hint: For protons, $m_p c^2 \approx 938$ MeV, while for neutral pions, $m_\pi c^2 \approx 135$ MeV.

9.16. Calculate the largest possible velocity of the electrons emitted by (initially, resting) neutrons at their *β-decays*:

$$n \to p + e + \bar{\nu}_e .$$

Hint: Electron neutrinos ν_e and antineutrinos $\bar{\nu}_e$ are virtually massless (on the energy scale of this problem); the rest energies $\mathscr{E} \equiv mc^2$ of the other involved particles are as follows: 939.565 MeV for the neutron, 938.272 MeV for the proton, 0.511 MeV for the electron.

9.17. A relativistic particle with a rest mass m and an energy $\mathscr{E}$ collides with a similar particle, initially at rest in the laboratory reference frame. Calculate:

(i) the final velocity of the center of mass of the system, in the lab frame,

(ii) the total energy of the system, in the center-of-mass frame, and

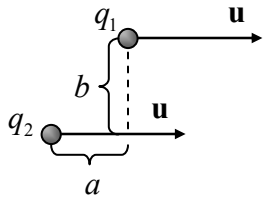
(iii) the final velocities of both particles (in the lab frame), provided that they move along the same direction.

9.18. A "primed" reference frame moves, relative to the "lab" frame, with a reduced velocity $\boldsymbol{\beta} \equiv \mathbf{v}/c = \mathbf{n}_x\beta$. Use Eq. (109) to express the elements T'^{00} and T'^{0j} (with $j = 1, 2, 3$) of an arbitrary contravariant 4-tensor $T^{\gamma\delta}$ via its elements in the lab frame.

9.19. Prove that quantities $E^2 - c^2B^2$ and $\mathbf{E}\cdot\mathbf{B}$ are Lorentz-invariant.

9.20. Consider the situation when static fields $\mathbf{E}$ and $\mathbf{B}$ are uniform but arbitrary (both in magnitude and in direction). What should be the velocity of an inertial reference frame to have the vectors $\mathbf{E}'$ and $\mathbf{B}'$, observed from that frame, parallel? Is this solution unique?

9.21. Two charged particles moving with equal constant velocities $\mathbf{u}$ are offset by a constant vector $\mathbf{R} = \{a, b\}$ (see the figure on the right), as measured in the lab frame. Calculate the force of interaction between the particles – also in the lab frame.



[84] This is the famous *Compton scattering* effect, whose discovery in 1923 was one of the major motivations for the development of quantum mechanics – see, e.g., QM Sec. 1.1.

9.22. Each of two thin, long, parallel particle beams of the same velocity **u**, separated by distance d, carries electric charges with a constant density λ per unit length, as measured in the reference frame moving with the particles.

(i) Calculate the distribution of the electric and magnetic fields in the system (outside the beams), as measured in the lab reference frame.

(ii) Calculate the interaction force between the beams (per particle) and the resulting acceleration, both in the lab reference frame and in the frame moving with the particles.

(iii) Compare the results and give a brief discussion of their relation.

9.23.

(i) Spell out the Lorentz transform of the Cartesian components of the scalar potential and the vector potential of an arbitrary electromagnetic field.

(ii) Use this general result to calculate the potentials of the field created by a point charge q moving with a constant velocity **u**, as measured in the lab reference frame.

9.24. Calculate the scalar and vector potentials created by a time-independent electric dipole **p**, as measured in a reference frame that moves relative to the dipole with a constant velocity **v**, with the shortest distance ("impact parameter") equal to b.

9.25. Solve the previous problem, in the limit $v << c$, for a time-independent magnetic dipole **m**.

9.26. Review the solution of Problem 23 (on the hypothetical magnetic monopole passing through a superconducting ring) for the case when this particle moves with an arbitrary constant velocity.

9.27. Re-derive Eq. (161) for the simplest case $\mathbf{p}(0) = 0$, by using the 4-vector form (145) of the equation of motion and the notion of rapidity $\varphi \equiv \tanh^{-1}\beta$ that was briefly discussed in Sec. 2.

9.28.* Calculate the trajectory of a relativistic particle in a uniform electrostatic field **E**, for an arbitrary direction of its initial velocity $\mathbf{u}(0)$, by using two different ways – at least one of them different from the approach described in Sec. 6 for the case $\mathbf{u}(0) \perp \mathbf{E}$.

9.29. A charged relativistic particle the rest mass m performs planar cyclotron rotation, with velocity u, in a uniform external magnetic field of magnitude B. How much would the velocity and the orbit's radius change at a slow change of the field to a new magnitude B'?

9.30.* Analyze the motion of a relativistic particle in uniform, mutually perpendicular fields **E** and **B**, for the particular case when E is *exactly* equal to cB.

9.31. Find the law of motion of a relativistic particle in uniform static fields **E** and **B** parallel to each other.

9.32. An external Lorentz force **F** is exerted on a relativistic particle with an electric charge q and a rest mass m, moving with velocity **u**, as observed from some inertial "lab" frame. Calculate its acceleration as observed from that frame.

9.33. Neglecting relativistic kinetic effects, calculate the lowest voltage V that has to be applied between the anode and cathode of a magnetron (see Fig. 13 and its discussion) to enable electrons to reach the anode, at negligible electron-electron interactions (including the space-charge effects) and collisions with the residual gas molecules. You may:

(i) model the cathode and anode as two coaxial round cylinders, of radii R_1 and R_2, respectively;
(ii) assume that the magnetic field **B** is uniform and directed along their common axis; and
(iii) neglect the initial velocity of the electrons emitted by the cathode.

After the solution, estimate the validity of the last assumption and of the non-relativistic approximation, for reasonable values of parameters.

9.34. A charged relativistic particle has been injected into a region with a uniform electric field whose magnitude oscillates in time with frequency ω. Calculate the time dependence of the particle's velocity, as observed from the lab reference frame.

9.35.* A linearly-polarized plane electromagnetic wave of frequency ω is incident on an otherwise free relativistic particle with electric charge q. Analyze the dynamics of the particle's momentum and compare the result with those of the previous problem and Problem 7.5.

9.36. Analyze the motion of a non-relativistic particle in a region where the electric and magnetic fields are both uniform and constant in time, but not necessarily parallel or perpendicular to each other.

9.37. A static distribution of electric charge in otherwise free space has created a time-independent distribution $\mathbf{E}(\mathbf{r})$ of the electric field. Use two different approaches to express the field energy density u' and the Poynting vector $\mathbf{S}'$, as observed from a reference frame moving with a constant velocity $\mathbf{v}$, via the Cartesian components of the vector $\mathbf{E}$. In particular, is $\mathbf{S}'$ equal to $(-\mathbf{v}u')$?

9.38. A traveling plane wave of frequency ω and intensity S is normally incident on a perfect mirror moving with velocity v in the same direction as the wave.

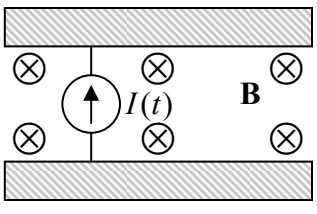
(i) Calculate the reflected wave's frequency, and
(ii) use the Lorentz transform of the fields to calculate the reflected wave's intensity

– both as observed from the lab reference frame.

9.39. Perform the second task of the previous problem by using general relations between the wave's energy, power, and momentum.

Hint: As a byproduct, this approach should also give you the pressure exerted by the wave on the moving mirror.

9.40. For the simple model of capacitor charging by a lumped source of current $I(t)$, shown in the figure on the right, prove that the momentum given by a constant, uniform external magnetic field **B** to the current-carrying conductor is equal and opposite to the momentum of the electromagnetic field that this current builds up in the capacitor. (You may assume that the capacitor is planar and very broad, and hence neglect the fringe field effects.)



9.41. Consider an electromagnetic plane wave packet propagating in free space, with its electric field represented as the Fourier integral

$$\mathbf{E}(\mathbf{r},t) = \mathrm{Re}\int_{-\infty}^{+\infty}\mathbf{E}_k e^{i\Psi_k}\,dk, \qquad \text{with } \Psi_k \equiv kz - \omega_k t, \quad \text{and } \omega_k \equiv c|k|.$$

Express the full linear momentum (per unit area of wave's front) of the packet via the complex amplitudes $\mathbf{E}_k$ of its Fourier components. Does the momentum depend on time? (In contrast with Problem 7.8, the wave packet is not necessarily narrow.)

9.42. Calculate the forces exerted on well-conducting walls of a waveguide with a rectangular ($a\times b$) cross-section, by a wave propagating along it in the fundamental (H_{10}) mode. Give an interpretation of the results.

Chapter 10. Radiation by Relativistic Charges

The discussion of special relativity in the previous chapter enables us to revisit the analysis of electromagnetic radiation by charged particles, now for arbitrary velocities. For a single point particle, it turns out to be possible to calculate the radiated wave fields in an explicit form and analyze the results for such important particular cases as synchrotron radiation and the "Bremsstrahlung" (brake radiation). After that, we will discuss the apparently unrelated effect of the so-called Coulomb losses of energy by a particle moving in condensed matter, because this discussion will naturally lead us to such important phenomena as the Cherenkov radiation and the transitional radiation. At the end of the chapter, I will briefly review the effects of the back action of the emitted radiation on the emitting particle, whose analysis reveals some limitations of classical electrodynamics.

10.1. Liénard-Wiechert potentials

A convenient starting point for the discussion of radiation by relativistic charges is provided by Eqs. (8.17) for the retarded potentials. In free space, these formulas with the integration variable notation changed from $\mathbf{r}'$ to $\mathbf{r}''$ for the clarity of what follows, are reduced to

$$\phi(\mathbf{r},t)=\frac{1}{4\pi\varepsilon_0}\int\frac{\rho(\mathbf{r}'',t-R/c)}{R}d^3r'',\quad \mathbf{A}(\mathbf{r},t)=\frac{\mu_0}{4\pi}\int\frac{\mathbf{j}(\mathbf{r}'',t-R/c)}{R}d^3r'',\quad \text{with } \mathbf{R}\equiv\mathbf{r}-\mathbf{r}''. \tag{10.1a}$$

As a reminder, Eqs. (1a) were derived from the Maxwell equations without any restrictions, and are very natural for situations with continuous distributions of the electric charge and/or current. However, for a single charged particle, whose charge and current distributions may be described as

$$\rho(\mathbf{r},t)=q\delta(\mathbf{r}-\mathbf{r}'),\qquad \mathbf{j}(\mathbf{r},t)=q\mathbf{u}\delta(\mathbf{r}-\mathbf{r}'),\qquad \text{with } \mathbf{u}\equiv\dot{\mathbf{r}}', \tag{10.1b}$$

where $\mathbf{r}'=\mathbf{r}'(t)$ is the instantaneous position of the charge, it is more convenient to recast Eqs. (1a) into an explicit form that would not require integration in each particular case. Indeed, as Eqs. (1) show, the potentials at a given observation point $\{\mathbf{r}, t\}$ are contributed by only one specific point $\{\mathbf{r}'(t_{\rm ret}), t_{\rm ret}\}$ of the particle's 4D trajectory (called its *world line*), which satisfies the following condition:

$$t_{\rm ret}\equiv t-\frac{R_{\rm ret}}{c}, \tag{10.2}$$

where $t_{\rm ret}$ is called the *retarded time*, and $R_{\rm ret}$ is the length of the following distance vector

$$\mathbf{R}_{\rm ret}\equiv\mathbf{r}(t)-\mathbf{r}'(t_{\rm ret}) \tag{10.3}$$

– physically, the distance covered by the electromagnetic wave from its emission to observation.

The reduction of Eqs. (1a) to such a simpler form, however, requires some care. Their naïve integration over $\mathbf{r}''$ would yield the following apparent but wrong results:

$$\phi(\mathbf{r},t)=\frac{1}{4\pi\varepsilon_0}\frac{q}{R_{\rm ret}},\quad \text{i.e. } \frac{\phi(\mathbf{r},t)}{c}=\frac{\mu_0}{4\pi}\frac{qc}{R_{\rm ret}};\qquad \mathbf{A}(\mathbf{r},t)=\frac{\mu_0}{4\pi}\frac{q\mathbf{u}_{\rm ret}}{R_{\rm ret}},\quad \text{(WRONG!)} \tag{10.4}$$

where $\mathbf{u}_{\rm ret}$ is the particle's velocity at the retarded point $\mathbf{r}'(t_{\rm ret})$. Eqs. (4) is a good example of how the relativity theory (even the special one :-) cannot be taken too lightly. Indeed, the strings (9.84)-(9.85), formed from the apparent potentials (4), would not obey the Lorentz transform rule (9.91), because according to Eqs. (2)-(3), the distance $R_{\rm ret}$ also depends on the reference frame it is measured in.

In order to correct the error, we need, first of all, to discuss the conditions (2)-(3). Combining them by eliminating $R_{\rm ret}$, we get the following equation for $t_{\rm ret}$:

$$c(t - t_{\rm ret}) = \left| \mathbf{r}(t) - \mathbf{r}'(t_{\rm ret}) \right|. \qquad (10.5)$$

Retarded time

Figure 1 depicts the graphical solution of this self-consistency equation as the only[1] point of intersection of the light cone of the observation point (see Fig. 9.9 and its discussion) and the particle's world line.

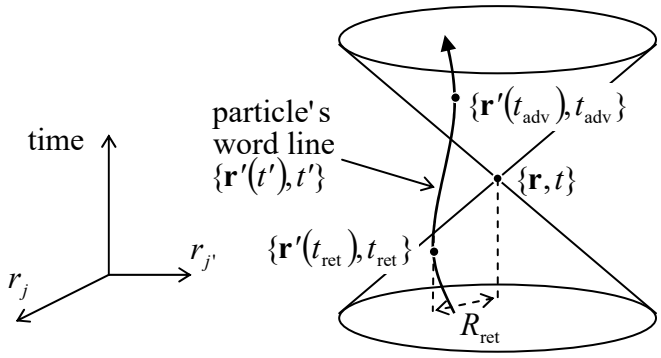


Fig. 10.1. Graphical solution of Eq. (5).

In Eq. (5), just as in Eqs. (1)-(3), all variables have to be measured in the inertial ("lab") reference frame in which the observation point $\mathbf{r}$ rests. Now let us write Eqs. (1) for a point charge in another inertial frame the frame 0' whose velocity (as measured in the lab frame) coincides, at the moment $t' = t_{\rm ret}$, with the velocity $\mathbf{u}_{\rm ret}$ of the charge.[2] In that frame, the charge rests, so, as we know from the electro- and magnetostatics,

$$\phi' = \frac{q}{4\pi\varepsilon_0 R'}, \qquad \mathbf{A}' = 0. \qquad (10.6a)$$

(Remember that this R' may not be equal to $R_{\rm ret}$, because the latter distance is measured in the "lab" reference frame.) Let us use the identity $1/\varepsilon_0 \equiv \mu_0 c^2$ again to rewrite Eqs. (6a) in the form of components of a 4-vector similar in structure to the last two of Eqs. (4):

$$\frac{\phi'}{c} = \frac{\mu_0}{4\pi}\frac{qc}{R'}, \qquad \mathbf{A}' = 0. \qquad (10.6b)$$

Now it is easy to guess the correct answer for the 4-potential for an arbitrary reference frame:

[1] As Fig. 1 shows, there is always another, "advanced" point $\{\mathbf{r}'(t_{\rm adv}), t_{\rm adv}\}$ of the particle's world line, with $t_{\rm adv} > t$, which is also a solution of Eq. (5), but it does not fit Eqs. (1), because the observation, at the point $\{\mathbf{r}, t < t_{\rm adv}\}$, of the field induced at the advanced point, would violate the causality principle.

[2] This is just a particular case of *the instantaneous reference frame* –the notion that was encountered in several exercise problems of the previous chapter, and indeed was implied (though admittedly not sufficiently advertised) as the derivation of the key Eq. (9.60).

$$A^\alpha = \frac{\mu_0}{4\pi}\frac{qcu^\alpha}{u_\beta R^\beta}, \tag{10.7}$$

where (as a reminder) $A^\alpha \equiv \{\phi/c, \mathbf{A}\}$, $u^\alpha \equiv \gamma\{c, \mathbf{u}\}$, and R^α is the 4-vector of the inter-event distance, formed similarly to that of a single event – cf. Eq. (9.48):

$$R^\alpha \equiv \{c(t-t'), \mathbf{R}'\} \equiv \{c(t-t'), \mathbf{r}-\mathbf{r}'\}. \tag{10.8}$$

Indeed, we needed the 4-vector A^α that would:

(i) obey the Lorentz transform,
(ii) have its spatial components A_j scaling, at low velocity, as u_j, and
(iii) be reduced to the correct result (6) in the instantaneous reference frame of the charge.

Eq. (7) evidently satisfies all these requirements, because the scalar product in its denominator is just

$$u_\beta R^\beta = \gamma\{c,-\mathbf{u}\}\cdot\{c(t-t'),\mathbf{R}\} \equiv \gamma[c^2(t-t') - \mathbf{u}\cdot\mathbf{R}] = \gamma c\left(R - \boldsymbol{\beta}\cdot\mathbf{R}\right) \equiv \gamma cR\left(1-\boldsymbol{\beta}\cdot\mathbf{n}\right), \tag{10.9}$$

where $\mathbf{n} \equiv \mathbf{R}/R$ is a unit vector in the observer's direction, $\boldsymbol{\beta} \equiv \mathbf{u}/c$ is the normalized velocity of the particle, and $\gamma \equiv 1/(1- u^2/c^2)^{1/2}$. In the instantaneous reference frame of the charge (in which $\boldsymbol{\beta} = 0$ and $\gamma = 1$), the expression (9) is reduced to cR, so Eq. (7) is correctly reduced to Eq. (6b). Now let us spell out the components of Eq. (7) for the lab frame (in which $t' = t_{\rm ret}$ and $R = R_{\rm ret}$):

Liénard-Wiechert potentials

$$\phi(\mathbf{r},t) = \frac{1}{4\pi\varepsilon_0}\frac{q}{(R-\boldsymbol{\beta}\cdot\mathbf{R})_{\rm ret}} \equiv \frac{1}{4\pi\varepsilon_0}q\left[\frac{1}{R(1-\boldsymbol{\beta}\cdot\mathbf{n})}\right]_{\rm ret}, \tag{10.10a}$$

$$\mathbf{A}(\mathbf{r},t) = \frac{\mu_0}{4\pi}q\left(\frac{\mathbf{u}}{R-\boldsymbol{\beta}\cdot\mathbf{R}}\right)_{\rm ret} \equiv \frac{\mu_0}{4\pi}qc\left[\frac{\boldsymbol{\beta}}{R(1-\boldsymbol{\beta}\cdot\mathbf{n})}\right]_{\rm ret} \equiv \phi(\mathbf{r},t)\frac{\mathbf{u}_{\rm ret}}{c^2}. \tag{10.10b}$$

These formulas are called the *Liénard-Wiechert potentials*.[3] In the non-relativistic limit, they coincide with the naïve guess (4), but in the general case include the additional factor $1/(1 - \boldsymbol{\beta}\cdot\mathbf{n})_{\rm ret}$. Its physical origin may be illuminated by one more formal calculation – whose result we will need anyway. Let us differentiate the geometric relation (5), rewritten as

$$R_{\rm ret} = c(t - t_{\rm ret}), \tag{10.11}$$

over $t_{\rm ret}$ and then, independently, over t, assuming that $\mathbf{r}$ is fixed. For that, let us first differentiate, over $t_{\rm ret}$, both sides of the identity $R_{\rm ret}^2 = \mathbf{R}_{\rm ret}\cdot\mathbf{R}_{\rm ret}$:

$$2R_{\rm ret}\frac{\partial R_{\rm ret}}{\partial t_{\rm ret}} = 2\mathbf{R}_{\rm ret}\cdot\frac{\partial \mathbf{R}_{\rm ret}}{\partial t_{\rm ret}}. \tag{10.12}$$

If $\mathbf{r}$ is fixed, then $\partial\mathbf{R}_{\rm ret}/\partial t_{\rm ret} \equiv \partial(\mathbf{r} - \mathbf{r}')/\partial t_{\rm ret} = -\partial\mathbf{r}'/\partial t_{\rm ret} \equiv -\mathbf{u}_{\rm ret}$, and Eq. (12) yields

$$\frac{\partial R_{\rm ret}}{\partial t_{\rm ret}} = \frac{\mathbf{R}_{\rm ret}}{R_{\rm ret}}\cdot\frac{\partial \mathbf{R}_{\rm ret}}{\partial t_{\rm ret}} = -(\mathbf{n}\cdot\mathbf{u})_{\rm ret}. \tag{10.13}$$

Now let us differentiate the same $R_{\rm ret}$ over t. On one hand, Eq. (11) yields

[3] They were derived in 1898 by Alfred-Marie Liénard and (independently) in 1900 by Emil Wiechert.

$$\frac{\partial R_{\rm ret}}{\partial t} = c - c\frac{\partial t_{\rm ret}}{\partial t}. \tag{10.14}$$

On the other hand, according to Eq. (5), at the partial differentiation over time, i.e. if $\mathbf{r}$ is fixed, $t_{\rm ret}$ is a function of t alone, so (using Eq. (13) at the second step), we may write

$$\frac{\partial R_{\rm ret}}{\partial t_{\rm ret}} = \frac{\partial R_{\rm ret}}{\partial t_{\rm ret}}\frac{\partial t_{\rm ret}}{\partial t} = -(\mathbf{n}\cdot\mathbf{u})_{\rm ret}\frac{\partial t_{\rm ret}}{\partial t}. \tag{10.15}$$

Now requiring Eqs. (14) and (15) to give the same result, we get:[4]

$$\frac{\partial t_{\rm ret}}{\partial t} = \frac{c}{c-(\mathbf{n}\cdot\mathbf{u})_{\rm ret}} \equiv \left(\frac{1}{1-\boldsymbol{\beta}\cdot\mathbf{n}}\right)_{\rm ret}. \tag{10.16}$$

$\partial t_{\rm ret}/\partial t$

This important relation may be readily re-derived (and more clearly understood) for the particular case when the charge's velocity is directed straight toward the observation point. In this case, its vector $\mathbf{u}$ resides in the same space-time plane as the observation point's world line $\mathbf{r}$ = const – say, the plane [x, t] shown in Fig. 2.

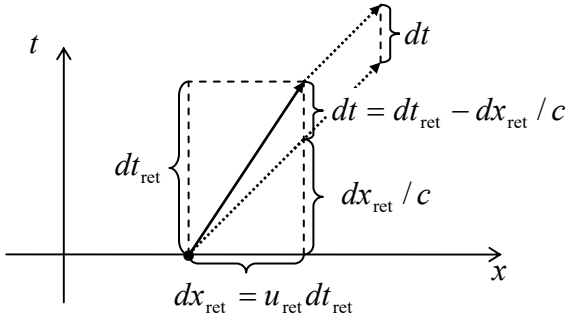


Fig. 10.2. Deriving Eq. (16) for the case $\boldsymbol{\beta}\cdot\mathbf{n} = \beta$.

Let us consider an elementary time interval $dt_{\rm ret} \equiv dt'$, during which the particle would travel the space interval $dx_{\rm ret} = u_{\rm ret}dt_{\rm ret}$. In Fig. 2, the corresponding segment of its world line is shown with a solid vector. The dotted vectors in this figure show the world lines of the radiation emitted by the particle in the beginning and at the end of this interval, and propagating with the speed of light c. As it follows from the drawing, the time interval dt between the instants of the arrival of the radiation from these two points to any time-independent spatial point of observation is

$$dt = dt_{\rm ret} - \frac{dx_{\rm ret}}{c} = dt_{\rm ret} - \frac{u_{\rm ret}}{c}dt_{\rm ret}, \qquad \text{so that } \frac{dt_{\rm ret}}{dt} = \frac{1}{1-u_{\rm ret}/c} \equiv \frac{1}{1-\beta_{\rm ret}}. \tag{10.17}$$

This expression coincides with Eq. (16) for our particular case when the directions of the vectors $\boldsymbol{\beta} \equiv \mathbf{u}/c$ and $\mathbf{n} \equiv \mathbf{R}/R$ (both taken at time $t_{\rm ret}$) coincide, and hence $(\boldsymbol{\beta}\cdot\mathbf{n})_{\rm ret} = \beta_{\rm ret}$. The difference between Eqs. (16) and (17) may be interpreted by saying that the particle's velocity in the transverse directions (normal to the vector $\mathbf{n}$) is not important for this kinematic effect[5] – the fact almost evident from Fig. 1.

[4] This relation may be used for an alternative derivation of Eqs. (10) directly from Eqs (1) – the calculation left for the reader's exercise.

[5] Note that this effect (linear in β) has nothing to do with the Lorentz time dilation (9.21), which is quadratic in β. (Indeed, all our arguments above referred to the same, lab frame.) Rather, it is close in nature to the Doppler effect.

So, the additional factor in the Liénard-Wiechert potentials is just the derivative $\partial t_{\rm ret}/\partial t$. The reason for its appearance in Eqs. (10) is usually interpreted along the following lines. Let the charge q be spread along the direction of the vector $\mathbf{R}_{\rm ret}$ (in Fig. 2, along the x-axis) by an infinitesimal speed-independent interval $\delta x_{\rm ret}$, so the linear density λ of its charge is proportional to $1/\delta x_{\rm ret}$. Then the time rate of *charge's* arrival at some spatial point is $\lambda u_{\rm ret} = \lambda dx_{\rm ret}/dt_{\rm ret}$, i.e. scales as $1/dt_{\rm ret}$. However, the rate of *radiation's* arrival at the observation point scales as $1/dt$, so due to the non-zero velocity $\mathbf{u}_{\rm ret}$ of the particle, this rate differs from the charge arrival rate by the factor of $dt_{\rm ret}/dt$, given by Eq. (16). (If the particle moves toward the observation point, $(\boldsymbol{\beta}\cdot\mathbf{n})_{\rm ret} > 0$, as shown in Fig. 2, this factor is larger than 1.) This radiation compression effect leads to the field change (at $(\boldsymbol{\beta}\cdot\mathbf{n})_{\rm ret} > 0$, its enhancement) by the same factor (16) – as described by Eqs. (10).

So, the 4-vector formalism was very instrumental for the calculation of field potentials. It may be also used to calculate the fields $\mathbf{E}$ and $\mathbf{B}$ – by plugging Eq. (7) into Eq. (9.124) to calculate the field strength tensor. This calculation yields

$$F^{\alpha\beta} = \frac{\mu_0 q}{4\pi}\frac{1}{u_\gamma R^\gamma}\frac{d}{d\tau}\left[\frac{R^\alpha u^\beta - R^\beta u^\alpha}{u_\delta R^\delta}\right]. \tag{10.18}$$

Now using Eq. (9.125) to identify the elements of this tensor with the field components, we may bring the result to the following vector form: [6]

Relativistic particle's fields

$$\mathbf{E} = \frac{q}{4\pi\varepsilon_0}\left[\frac{\mathbf{n}-\boldsymbol{\beta}}{\gamma^2(1-\boldsymbol{\beta}\cdot\mathbf{n})^3 R^2} + \frac{\mathbf{n}\times\left\{(\mathbf{n}-\boldsymbol{\beta})\times\dot{\boldsymbol{\beta}}\right\}}{(1-\boldsymbol{\beta}\cdot\mathbf{n})^3 cR}\right]_{\rm ret}, \tag{10.19}$$

$$\mathbf{B} = \frac{\mathbf{n}_{\rm ret}\times\mathbf{E}}{c}, \qquad \text{i.e. } \mathbf{H} = \frac{\mathbf{n}_{\rm ret}\times\mathbf{E}}{Z_0}. \tag{10.20}$$

Thus the magnetic and electric fields of a relativistic particle are always proportional and perpendicular to each other, and related just as in a plane wave – cf. Eq. (7.6), with the difference that now the vector $\mathbf{n}_{\rm ret}$ may be a function of time. Superficially, this result contradicts the electro- and magnetostatics, because, for a particle at rest, $\mathbf{B}$ should vanish while $\mathbf{E}$ stays finite. However, note that according to the Coulomb law for a point charge, in this case, $\mathbf{E} = E\mathbf{n}_{\rm ret}$, so $\mathbf{B} \propto \mathbf{n}_{\rm ret}\times\mathbf{E} \propto \mathbf{n}_{\rm ret}\times\mathbf{n}_{\rm ret} = 0$. (Actually, in these relations, the subscript "ret" is unnecessary.)

As a sanity check, let us use Eq. (19) as an alternative way to find the electric field of a charge moving without acceleration, i.e. uniformly, along a straight line – see Fig. 9.11a reproduced, with minor changes, in Fig. 3. (This calculation will also illustrate the technical challenges of practical applications of the Liénard-Wiechert formulas for even simple cases.) In this case, the vector $\boldsymbol{\beta}$ does not change in time, so the second term in Eq. (19) vanishes, and all we need to do is to spell out the Cartesian components of the first term.

[6] An alternative way of deriving these formulas (highly recommended to the reader as an exercise) is to plug Eqs. (10) into the general relations (9.121), and carry out the required temporal and spatial differentiations directly, using Eq. (16) and its spatial counterpart (which may be derived absolutely similarly):

$$\nabla t_{\rm ret} = -\left[\frac{\mathbf{n}}{c(1-\boldsymbol{\beta}\cdot\mathbf{n})}\right]_{\rm ret}.$$

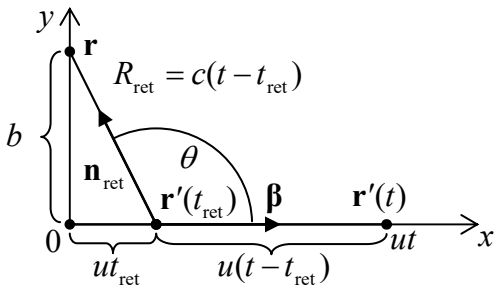


Fig. 10.3. The linearly moving charge problem.

Let us select the coordinate axes and the time origin as shown in Fig. 3, and make a clear distinction between the actual position, $\mathbf{r}'(t) = \{ut, 0, 0\}$ of the charged particle at the instant t we are considering, and its position $\mathbf{r}'(t_{ret})$ at the retarded instant defined by Eq. (5), i.e. the moment when the particle's field had to be radiated to reach the observation point $\mathbf{r}$ at the given time t, propagating with the speed of light. In these coordinates

$$\boldsymbol{\beta} = \{\beta, 0, 0\}, \quad \mathbf{r} = \{0, b, 0\}, \quad \mathbf{r}'(t_{ret}) = \{ut_{ret}, 0, 0\}, \quad \mathbf{n}_{ret} = \{\cos\theta, \sin\theta, 0\}, \tag{10.21}$$

with $\cos\theta = -ut_{ret}/R_{ret}$, so $[(\mathbf{n} - \boldsymbol{\beta})_x]_{ret} = -ut_{ret}/R_{ret} - \beta$, and Eq. (19) yields, in particular:

$$E_x = \frac{q}{4\pi\varepsilon_0} \frac{-ut_{ret}/R_{ret} - \beta}{\gamma^2 \left[(1 - \boldsymbol{\beta}\cdot\mathbf{n})^3 R^2\right]_{ret}} \equiv \frac{q}{4\pi\varepsilon_0} \frac{-ut_{ret} - \beta R_{ret}}{\gamma^2 \left[(1 - \boldsymbol{\beta}\cdot\mathbf{n})^3 R^3\right]_{ret}}. \tag{10.22}$$

But according to Eq. (5), the product βR_{ret} may be represented as $\beta c(t - t_{ret}) \equiv u(t - t_{ret})$. Plugging this expression into Eq. (22), we may eliminate the explicit dependence of E_x on time t_{ret}:

$$E_x = \frac{q}{4\pi\varepsilon_0} \frac{-ut}{\gamma^2 \left[(1 - \boldsymbol{\beta}\cdot\mathbf{n})R\right]^3_{ret}}. \tag{10.23}$$

The only non-zero transverse component of the field also has a similar form:

$$E_y = \frac{q}{4\pi\varepsilon_0} \left[\frac{\sin\theta}{\gamma^2 (1 - \boldsymbol{\beta}\cdot\mathbf{n})^3 R^2}\right]_{ret} = \frac{q}{4\pi\varepsilon_0} \frac{b}{\gamma^2 \left[(1 - \boldsymbol{\beta}\cdot\mathbf{n})R\right]^3_{ret}}, \tag{10.24}$$

while $E_z = 0$. From Fig. 3, $\boldsymbol{\beta} - \mathbf{n}_{ret} = \beta\cos\theta = -\beta ut_{ret}/R_{ret}$, so $(1 - \boldsymbol{\beta}\cdot\mathbf{n})R_{ret} \equiv R_{ret} + \beta ut_{ret}$, and we may again use Eq. (5) to get $(1 - \boldsymbol{\beta}\cdot\mathbf{n})R_{ret} = c(t - t_{ret}) + \beta ut_{ret} \equiv ct - ct_{ret}/\gamma^2$. What remains is to calculate t_{ret} from the self-consistency equation (5), whose square in our current case (Fig. 3) takes the form

$$R^2_{ret} \equiv b^2 + (ut_{ret})^2 = c^2 (t - t_{ret})^2. \tag{10.25}$$

This is a simple quadratic equation for t_{ret}, which (with the appropriate negative sign before the square root, to get $t_{ret} < t$) yields:

$$t_{ret} = \gamma^2 t - \left[\left(\gamma^2 t\right)^2 - \gamma^2 \left(t^2 - b^2/c^2\right)\right]^{1/2} \equiv \gamma^2 t - \frac{\gamma}{c}\left(u^2\gamma^2 t^2 + b^2\right)^{1/2}, \tag{10.26}$$

so the only retarded-function combination that participates in Eqs. (23)-(24) is

$$\left[(1 - \boldsymbol{\beta}\cdot\mathbf{n})R\right]_{ret} = \frac{c}{\gamma^2}\left(u^2\gamma^2 t^2 + b^2\right)^{1/2}, \tag{10.27}$$

and, finally, the electric field components are

$$E_x = -\frac{q}{4\pi\varepsilon_0}\frac{\gamma u t}{\left(b^2+\gamma^2u^2t^2\right)^{3/2}}, \qquad E_y = \frac{q}{4\pi\varepsilon_0}\frac{\gamma b}{\left(b^2+\gamma^2u^2t^2\right)^{3/2}}, \qquad E_z = 0. \tag{10.28}$$

But these are exactly Eqs. (9.139),[7] which had been obtained in Sec. 9.5 by much simpler means, without the necessity to solve the self-consistency equation (5). However, that alternative approach was essentially based on the inertial motion of the particle, and cannot be used in problems in which it moves with acceleration. In such problems, the second term in Eq. (19), dropping with distance more slowly, as $1/R_{\text{ret}}$, and hence describing wave radiation, is frequently the most important one.

10.2. Radiation power

Let us calculate the angular distribution of the particle's radiation. For that, we need to return to Eqs. (19)-(20) to find the Poynting vector $\mathbf{S} = \mathbf{E}\times\mathbf{H}$, and in particular, its radial component $S_n = \mathbf{S}\cdot\mathbf{n}_{\text{ret}}$, at large distances R from the particle. Following tradition,[8] let us express the result as the energy radiated into unit solid angle per unit time interval dt_{rad} of the *radiation*, rather than that (dt) of its *measurement*. (We will need to return to the measurement time t in the next section to calculate the observed radiation spectrum.) Using Eq. (16), we get

$$\frac{d\mathscr{P}}{d\Omega} \equiv -\frac{d\mathscr{E}}{d\Omega dt_{\text{ret}}} = \left(R^2S_n\right)_{\text{ret}}\frac{\partial t}{\partial t_{\text{ret}}} = \left(\mathbf{E}\times\mathbf{H}\right)\cdot\left[R^2\mathbf{n}\left(1-\boldsymbol{\beta}\cdot\mathbf{n}\right)\right]_{\text{ret}}. \tag{10.29}$$

At sufficiently large distances from the particle, i.e. in the limit $R_{\text{ret}} \to \infty$ (in the *radiation zone*), the contribution of the first (essentially, the Coulomb-field) term in the square brackets of Eq. (19) vanishes as $1/R^2$, and the substitution of the remaining term into Eqs. (20) and then Eq. (29) yields the following formula, which is valid for an arbitrary law of the particle's motion:[9]

Radiation power density

$$\boxed{\frac{d\mathscr{P}}{d\Omega} = \frac{Z_0q^2}{(4\pi)^2}\frac{\left|\mathbf{n}\times\left[\left(\mathbf{n}-\boldsymbol{\beta}\right)\times\dot{\boldsymbol{\beta}}\right]\right|^2}{\left(1-\mathbf{n}\cdot\boldsymbol{\beta}\right)^5}}. \tag{10.30}$$

Now, let us apply this important result to some simple cases. First of all, Eq. (30) says that a charge moving with a constant velocity $\boldsymbol{\beta}$ does not radiate at all. This might be expected from our analysis of this case in Sec. 9.5 because in the reference frame moving with the charge it produces only the Coulomb electrostatic field, i.e. no radiation.

Next, let us consider a linear motion of a point charge with a non-zero acceleration directed along the straight line of the motion. In this case, with the coordinate axes selected as shown in Fig. 4a, each of the vectors involved in Eq. (30) has at most two non-zero Cartesian components:

[7] A similar calculation of magnetic field components from Eq. (20) gives results identical to Eqs. (9.140).

[8] This tradition may be reasonably justified. Indeed, we may say that the radiation field "detaches" from the particle at times close to t_{ret}, while the observation time t depends on the detector's position, and hence is less relevant for the radiation process as such.

[9] If the direction of radiation, $\mathbf{n}$, does not change in time, this formula does not depend on the observer's position $\mathbf{R}$. Hence, from this point on, the index "ret" may be safely dropped for brevity, though we should always remember that $\boldsymbol{\beta}$ in Eq. (30) is the reduced velocity of the particle at the instant of the radiation's *emission*, not of its observation.

$$\mathbf{n} = \{\sin\theta, 0, \cos\theta\}, \qquad \boldsymbol{\beta} = \{0, 0, \beta\}, \qquad \dot{\boldsymbol{\beta}} = \{0, 0, \dot{\beta}\}, \tag{10.31}$$

where θ is the angle between the directions of the particle's motion and of the radiation's propagation. Plugging these expressions into Eq. (30) and performing the vector multiplications, we readily get

$$\frac{d\mathscr{P}}{d\Omega} = \frac{Z_0 q^2}{(4\pi)^2}\dot{\beta}^2 \frac{\sin^2\theta}{(1-\beta\cos\theta)^5}. \tag{10.32}$$

Figure 4b shows the angular distribution of such radiation, for three values of the particle's speed u.

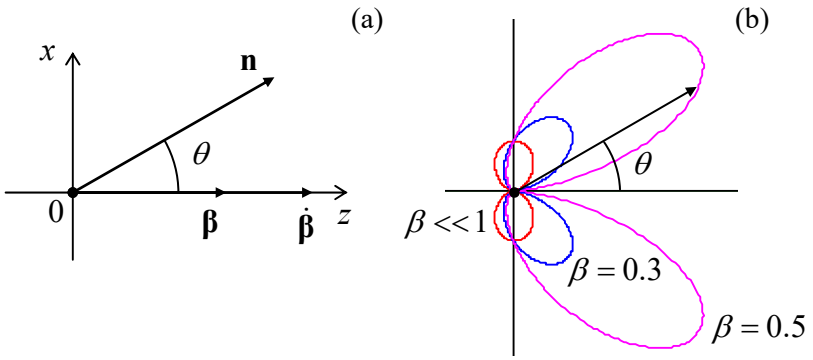


Fig. 10.4. Particle's radiation at its linear acceleration: (a) the problem's geometry, and (b) the last fraction of Eq. (32) as a function of the angle θ.

If the speed is relatively low ($u << c$, i.e. $\beta << 1$), the denominator in Eq. (32) is very close to 1 for all observation angles θ, so the angular distribution of the radiation power is close to $\sin^2\theta$ – just as it follows from the general non-relativistic Larmor formula (8.26), for our current case with $\Theta = \theta$. However, as the velocity is increased, the denominator becomes less than 1 for $\theta < \pi/2$, i.e. for the forward-looking directions, and larger than 1 for back directions. As a result, the radiation in the direction of the particle's motion is increased (somewhat counter-intuitively, regardless of the acceleration's sign!), while that in the back direction is suppressed. For ultra-relativistic particles ($\beta \to 1$), this trend is strongly exacerbated, and radiation to very small forward angles dominates. To describe this main part of the angular distribution, we may expand the trigonometric functions of θ participating in Eq. (32) in the Taylor series in small θ, and keep only their leading terms: $\sin\theta \approx \theta$, $\cos\theta \approx 1 - \theta^2/2$, so $(1 - \beta\cos\theta) \approx (1 + \gamma^2\theta^2)/2\gamma^2$. The resulting expression,

$$\frac{d\mathscr{P}}{d\Omega} \approx \frac{2Z_0 q^2}{\pi^2}\dot{\beta}^2\gamma^8 \frac{(\gamma\theta)^2}{\left(1+\gamma^2\theta^2\right)^5}, \quad \text{for } \gamma >> 1, \tag{10.33}$$

describes a narrow "hollow cone" distribution of radiation, with its maximum at the angle

$$\theta_0 = \frac{1}{2\gamma} << 1. \tag{10.34}$$

Another important aspect of Eq. (33) is how extremely fast (as γ^8) the radiation density grows with the Lorentz factor γ, i.e. with the particle's energy $\mathscr{E} = \gamma mc^2$.

Still, the total radiated power $\mathscr{P}$ (into all observation angles) at linear acceleration is not too high for any practicable values of parameters. To show this, let us first calculate $\mathscr{P}$ for an arbitrary motion of the particle. To start, let me demonstrate how $\mathscr{P}$ may be found (or rather guessed) from the general relativistic arguments. In Sec. 8.2, we have derived Eq. (8.27) for the power of the electric dipole radiation for a non-relativistic particle motion. That result is valid, in particular, for one charged particle,

whose electric dipole moment's derivative over time may be expressed as $d(q\mathbf{r})/dt = (q/m)\mathbf{p}$, where $\mathbf{p}$ is the particle's linear mechanical momentum (*not* its electric dipole moment). As a result, the Larmor formula (8.27) in free space, i.e. with $v = c$ (but $u << c$) reduces to

$$\mathscr{P} = \frac{Z_0}{6\pi c^2}\left(\frac{q}{m}\frac{dp}{dt}\right)^2 \equiv \frac{Z_0 q^2}{6\pi m^2 c^2}\left(\frac{d\mathbf{p}}{dt}\cdot\frac{d\mathbf{p}}{dt}\right), \quad \text{for } u << c. \tag{10.35}$$

This is evidently not a Lorentz-invariant result, but it gives a clear hint of how such an invariant, which would be reduced to Eq. (35) in the non-relativistic limit, may be formed:

$$\mathscr{P} = -\frac{Z_0 q^2}{6\pi m^2 c^2}\left(\frac{dp_\alpha}{d\tau}\cdot\frac{dp^\alpha}{d\tau}\right) \equiv \frac{Z_0 q^2}{6\pi m^2 c^2}\left[\left(\frac{d\mathbf{p}}{d\tau}\right)^2 - \frac{1}{c^2}\left(\frac{d\mathscr{E}}{d\tau}\right)^2\right]. \tag{10.36}$$

Using the relativistic expressions $\mathbf{p} = \gamma mc\boldsymbol{\beta}$, $\mathscr{E} = \gamma mc^2$, and $d\tau = dt/\gamma$, the last formula may be recast into the so-called *Liénard extension* of the Larmor formula:[10]

Total radiation power via $\boldsymbol{\beta}$

$$\mathscr{P} = \frac{Z_0 q^2}{6\pi}\gamma^6\left[\left(\dot{\boldsymbol{\beta}}\right)^2 - \left(\boldsymbol{\beta}\times\dot{\boldsymbol{\beta}}\right)^2\right] \equiv \frac{Z_0 q^2}{6\pi}\gamma^4\left[\left(\dot{\boldsymbol{\beta}}\right)^2 + \gamma^2\left(\boldsymbol{\beta}\cdot\dot{\boldsymbol{\beta}}\right)^2\right]. \tag{10.37}$$

It may be also obtained by direct integration of Eq. (30) over the full solid angle, thus confirming our guess.

However, for some applications, it is beneficial to express $\mathscr{P}$ via the time evolution of the particle's momentum alone. For that, we may differentiate the fundamental relativistic relation (9.78), $\mathscr{E}^2 = (mc^2)^2 + (pc)^2$, over the proper time τ to get

$$2\mathscr{E}\frac{d\mathscr{E}}{d\tau} = 2c^2 p\frac{dp}{d\tau}, \qquad \text{i.e.} \quad \frac{d\mathscr{E}}{d\tau} = \frac{c^2 p}{\mathscr{E}}\frac{dp}{d\tau} = u\frac{dp}{d\tau}, \tag{10.38}$$

where the last step used the relativistic relation $c^2\mathbf{p}/\mathscr{E} = \mathbf{u}$ mentioned in Sec. 9.3. Plugging Eq. (38) into Eq. (36), we may rewrite it as

Total radiation power via $\mathbf{p}$

$$\mathscr{P} = \frac{Z_0 q^2}{6\pi m^2 c^2}\left[\left(\frac{d\mathbf{p}}{d\tau}\right)^2 - \beta^2\left(\frac{dp}{d\tau}\right)^2\right]. \tag{10.39}$$

Please note the difference between the squared derivatives in this expression: in the first of them we have to differentiate the momentum's vector $\mathbf{p}$ first, and only then form a scalar by squaring the resulting vector derivative, while in the second case, only the magnitude of the vector has to be differentiated. For example, for circular motion with a constant speed (to be analyzed in detail in the next section), the second term vanishes, while the first one does not.

However, if we return to the simplest case of linear acceleration (Fig. 4), then $(d\mathbf{p}/d\tau)^2 = (dp/d\tau)^2$, and Eq. (39) is reduced to

[10] The second form of Eq. (10.37), which is frequently more convenient for applications, may be readily obtained from the first one by applying MA Eq. (7.7a) to the vector product.

$$\mathscr{P} = \frac{Z_0 q^2}{6\pi m^2 c^2}\left(\frac{dp}{d\tau}\right)^2\left(1-\beta^2\right) \equiv \frac{Z_0 q^2}{6\pi m^2 c^2}\left(\frac{dp}{d\tau}\right)^2 \frac{1}{\gamma^2} \equiv \frac{Z_0 q^2}{6\pi m^2 c^2}\left(\frac{dp}{dt_{\text{ret}}}\right)^2, \tag{10.40}$$

i.e. formally coincides with the non-relativistic relation (35). To get a better feeling of the magnitude of this radiation, we may combine Eq. (9.144) with $\mathbf{B} = 0$, and Eq. (9.148) with $\mathbf{E} \parallel \mathbf{u}$ to get $dp/dt_{\text{ret}} = d\mathscr{E}/dz'$, where z' is the particle's coordinate at the moment t_{ret}. The last relation allows us to rewrite Eq. (40) in the following form:

$$\mathscr{P} = \frac{Z_0 q^2}{6\pi m^2 c^2}\left(\frac{d\mathscr{E}}{dz}\right)^2 \equiv \frac{Z_0 q^2}{6\pi m^2 c^2}\frac{d\mathscr{E}}{dz'}\frac{d\mathscr{E}}{dt_{\text{ret}}}\frac{dt_{\text{ret}}}{dz'} \equiv \frac{Z_0 q^2}{6\pi m^2 c^2 u}\frac{d\mathscr{E}}{dz'}\frac{d\mathscr{E}}{dt_{\text{ret}}}. \tag{10.41}$$

For the most important case of ultra-relativistic motion ($u \to c$), this result reduces to

$$\frac{\mathscr{P}}{d\mathscr{E}/dt_{\text{ret}}} \approx \frac{2}{3}\frac{d(\mathscr{E}/mc^2)}{d(z'/r_c)}, \tag{10.42}$$

where r_c is the classical radius of the particle, defined by Eq. (8.41). This formula shows that the radiated power, i.e. the change of the particle's energy due to radiation, is much smaller than that due to the accelerating field unless energy as large as $\sim mc^2$ is gained on the classical radius of the particle. For example, for an electron, with $r_c \approx 3\times10^{-15}$ m and $mc^2 = m_e c^2 \approx 0.5$ MeV, such an acceleration would require the accelerating electric field of the order of (0.5 MV)/(3×10^{-15} m) ~ 10^{14} MV/m, while practicable accelerating fields are below 10^2 MV/m – limited by the electric breakdown effects. (As described by the factor m^2 in the denominator of Eq. (41), for heavier particles such as protons, the relative losses are even lower.) Such negligible radiative losses of energy are actually a large advantage of linear accelerators – such as the famous two-mile-long SLAC,[11] which can accelerate electrons or positrons to energies up to 50 GeV, i.e. to $\gamma \approx 10^5$. If obtaining radiation from the accelerated particles is the goal, it may be readily achieved by bending their trajectories using additional magnetic fields – see the next section.

10.3. Synchrotron radiation

Now let us consider a charged particle being accelerated in the direction perpendicular to its velocity $\mathbf{u}$ (for example by the magnetic component of the Lorentz force), so its speed u, and hence the magnitude p of its momentum, do not change. In this case, the second term in the square brackets of Eq. (39) vanishes, and it yields

$$\mathscr{P} = \frac{Z_0 q^2}{6\pi m^2 c^2}\left(\frac{d\mathbf{p}}{d\tau}\right)^2 = \frac{Z_0 q^2}{6\pi m^2 c^2}\left(\frac{d\mathbf{p}}{dt_{\text{ret}}}\right)^2 \gamma^2. \tag{10.43}$$

Comparing this expression with Eq. (40), we see that for the same acceleration magnitude, the electromagnetic radiation is a factor of γ^2 larger. For modern accelerators, with $\gamma \sim 10^4$-10^5, such a factor creates an enormous difference. For example, if a particle is on a cyclotron orbit in a constant magnetic field (as was analyzed in Sec. 9.6), both $\mathbf{u}$ and $\mathbf{p} = \gamma m\mathbf{u}$ obey Eq. (9.150), so

[11] See, e.g., https://www6.slac.stanford.edu/.

$$\left|\frac{d\mathbf{p}}{dt_{\rm ret}}\right| = \omega_c p = \frac{u}{R} p = \beta^2 \gamma \frac{mc^2}{R}, \tag{10.44}$$

(where R is the orbit's radius), so for the power of this *synchrotron radiation*, Eq. (43) yields

Synchrotron radiation: total power

$$\mathscr{P} = \frac{Z_0 q^2}{6\pi} \beta^4 \gamma^4 \frac{c^2}{R^2} \equiv \frac{1}{4\pi\varepsilon_0} \frac{2}{3} \frac{q^4 B^2}{m^2 c} \beta^2 \gamma^2 . \tag{10.45}$$

Note that for ultrarelativistic particles ($\beta \approx 1$), the power grows as γ^2, i.e. as the square of the particle's energy $\mathscr{E} \propto \gamma$. For example, for typical parameters of the first electron cyclotrons (such as the General Electric's machine in which the synchrotron radiation was first noticed in 1947), $R \sim 1$ m, $\mathscr{E} \sim 0.3$ GeV ($\gamma \sim 600$), Eq. (45) gives a very modest electron energy loss per one revolution: $\mathscr{P}\mathscr{T} \equiv \mathscr{P}(2\pi R/u) \approx 2\pi \mathscr{P} R/c \sim 1$ keV. However, already by the mid-1970s, electron accelerators, with $R \sim 100$ m, could give each particle energy $\mathscr{E} \sim 10$ GeV, and the energy loss per revolution grew to ~ 10 MeV, becoming the major energy loss mechanism. For proton accelerators, such energy loss is much less of a problem, because the γ of an ultra-relativistic particle (at fixed $\mathscr{E}$) is proportional to $1/m$, so the estimates, at the same R, should be scaled back by $(m_p/m_e)^4 \sim 10^{13}$. Nevertheless, in the giant modern accelerators such as the LHC (with $R \approx 4.3$ km and $\mathscr{E}$ up to 7 TeV), the synchrotron radiation loss per revolution is rather noticeable ($\mathscr{P}\mathscr{T} \sim 6$ keV), leading not as much to particle deceleration as to a substantial photoelectron emission from the beam tube's walls, creating harmful defocusing effects.

However, what is bad for particle accelerators and storage rings is good for the so-called *synchrotron light sources* – the electron accelerators designed for the generation of intensive synchrotron radiation – with the spectrum extending well beyond the visible light range. Let us analyze the angular and spectral distributions of such radiation. To calculate the angular distribution, let us select the coordinate axes as shown in Fig. 5, with the origin at the current location of the orbiting particle, the z-axis directed along its instant velocity (i.e. the vector $\boldsymbol{\beta}$), and the x-axis, toward the orbit's center.

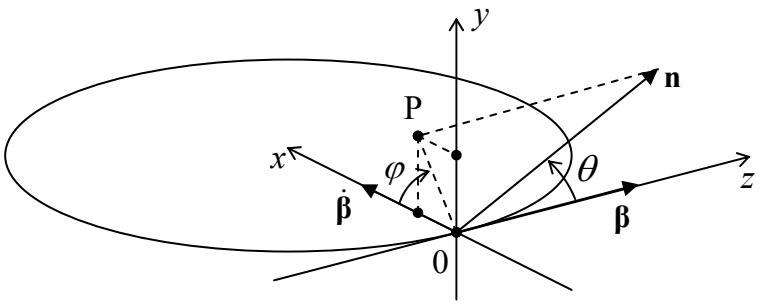


Fig. 10.5. The synchrotron radiation problem's geometry.

In the general case, when the unit vector $\mathbf{n}$ toward the radiation's observer is not within any of the coordinate planes, it has to be described by two angles – the polar angle θ, and the azimuthal angle φ between the x-axis and the projection 0P of the vector $\mathbf{n}$ onto the $[x, y]$-plane. Since the length of the segment 0P is $\sin\theta$, the Cartesian components of the relevant vectors are as follows:

$$\mathbf{n} = \{\sin\theta\cos\varphi, \sin\theta\sin\varphi, \cos\theta\}, \qquad \boldsymbol{\beta} = \{0, 0, \beta\}, \qquad \text{and} \quad \dot{\boldsymbol{\beta}} = \{\dot{\beta}, 0, 0\}. \tag{10.46}$$

Plugging these expressions into the general Eq. (30), we get

$$\frac{d\mathscr{P}}{d\Omega} = \frac{2Z_0 q^2}{\pi^2}\left|\dot{\boldsymbol{\beta}}\right|^2 \gamma^6 f(\theta,\varphi), \quad \text{where}$$
$$f(\theta,\varphi) \equiv \frac{1}{8\gamma^6(1-\beta\cos\theta)^3}\left[1-\frac{\sin^2\theta\cos^2\varphi}{\gamma^2(1-\beta\cos\theta)^2}\right], \tag{10.47}$$

Synchrotron radiation: angular distribution

According to this result, just as at the linear acceleration, in the ultra-relativistic limit, most radiation goes into a narrow cone (of a width $\Delta\theta \sim \gamma^{-1} << 1$) around the vector $\boldsymbol{\beta}$, i.e. around the instant direction of the particle's propagation. For such small angles, and $\gamma >> 1$,

$$f(\theta,\varphi) \approx \frac{1}{(1+\gamma^2\theta^2)^3}\left[1-\frac{4\gamma^2\theta^2\cos^2\varphi}{(1+\gamma^2\theta^2)^2}\right]. \tag{10.48}$$

The left panel of Fig. 6 shows a color-coded contour map of this angular distribution $f(\theta, \varphi)$, as observed on a distant plane normal to the particle's instant velocity (in Fig. 5, parallel to the $[x, y]$-plane), while its right panel shows the factor f as a function of θ in two perpendicular directions: within the particle's rotation plane (in the direction parallel to the x-axis, i.e. at $\varphi = 0$) and perpendicular to this plane (along the y-axis, i.e. at $\varphi = \pm\pi/2$). The result shows, first of all, that, in contrast to the case of linear acceleration, the narrow radiation cone is now not hollow: the intensity maximum is reached at $\theta = 0$, i.e. exactly in the direction of the particle's motion direction. Second, the radiation cone is not axially symmetric: within the particle rotation plane, the intensity drops faster (and even has nodes at $\theta = \pm 1/\gamma$).

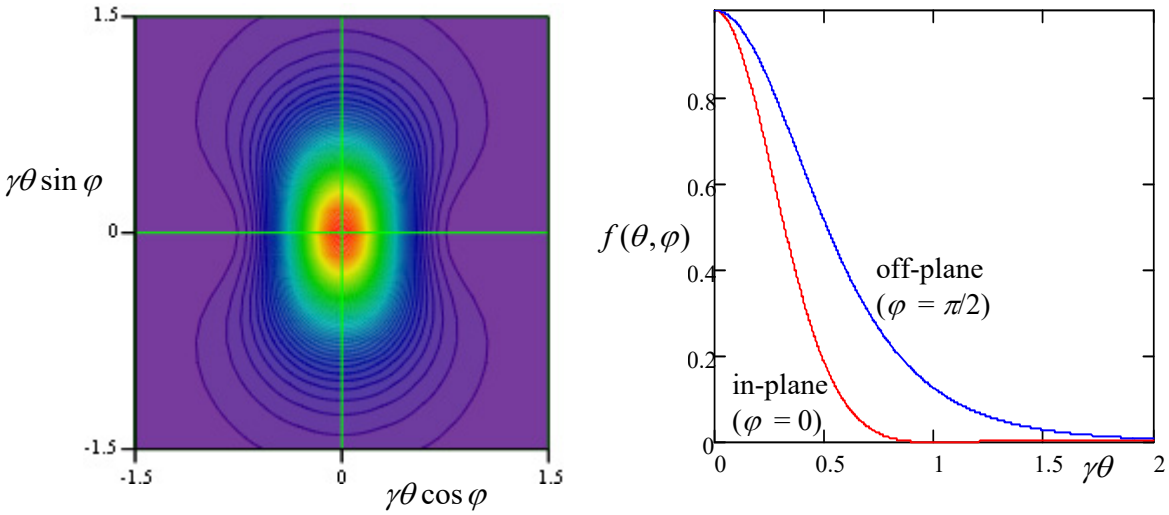


Fig. 10.6. The angular distribution of the synchrotron radiation at $\gamma >> 1$.

The angular distribution (47) of the synchrotron radiation was calculated for the (inertial) reference frame whose origin coincides with the particle's position at this particular instant, i.e. its radiation pattern is time-independent in the frame moving with the particle. This pattern enables a semi-quantitative description of the radiation by an ultra-relativistic particle from the point of view of a stationary observer: if the observation point is on (or very close to) the rotation plane,[12] it is being

[12] It is easy (and hence is left for the reader's exercise) to show that if the observation point is much off-plane (say, is located on the particle orbit's axis), the radiation is virtually monochromatic, with frequency ω_c. (As we know from Sec. 8.2, in the non-relativistic limit $u << c$, this is true for *any* observation point.)

"struck" by the narrow radiation cone once each rotation period $\mathcal{T} \approx 2\pi R/c$, each "strike" giving a field pulse of a short duration $\Delta t_{\rm ret} << 1/\omega_{\rm c}$ – see Fig. 7.[13]

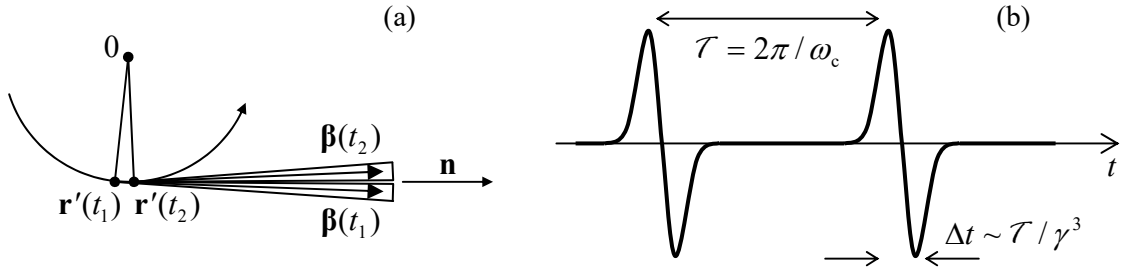


Fig. 10.7. (a) The synchrotron radiation cones (at $\gamma >> 1$) for two close values of $t_{\rm ret}$, and (b) the in-plane component of the electric field observed in the rotation plane, as a function of time t – schematically.

The evaluation of the time duration Δt of each pulse requires some care: its estimate $\Delta t_{\rm ret} \sim 1/\gamma\omega_{\rm c}$ is correct for the duration of the retarded time interval during which its cone is aimed at the observer. However, due to the time compression effect discussed in detail in Sec. 1 and described by Eq. (16), the pulse duration as seen by the observer is a factor of $1/(1-\beta)$ shorter, so

$$\Delta t = (1-\beta)\Delta t_{\rm ret} \sim \frac{1-\beta}{\gamma\omega_{\rm c}} \sim \frac{1}{\gamma^3\omega_{\rm c}} \sim \gamma^{-3}\mathcal{T}, \qquad \text{for } \gamma >> 1. \tag{10.49}$$

From the Fourier theorem, we can expect the frequency spectrum of such radiation to consist of numerous ($N \sim \gamma^3 >> 1$) harmonics of the particle rotation frequency $\omega_{\rm c}$, with comparable amplitudes. However, if the orbital frequency fluctuates even slightly ($\delta\omega_{\rm c}/\omega_{\rm c} > 1/N \sim 1/\gamma^3$), as it happens in most practical systems, the radiation pulses are not coherent, so the average radiation power spectrum may be calculated as that of one pulse, multiplied by the number of pulses per second. In this case, the spectrum is continuous, extending from low frequencies all the way to approximately

$$\omega_{\rm max} \sim 1/\Delta t \sim \gamma^3\omega_{\rm c}. \tag{10.50}$$

In order to verify and quantify this result, let us calculate the spectrum of radiation due to a single pulse. For that, we should first make the general notion of the radiation spectrum quantitative. Let us represent an arbitrary electric field (say that of the synchrotron radiation we are studying now) observed at a fixed point $\mathbf{r}$, as a function of the *observation* time t, as a Fourier integral:[14]

$$\mathbf{E}(t) = \int_{-\infty}^{+\infty} \mathbf{E}_\omega e^{-i\omega t}\, dt. \tag{10.51}$$

[13] The fact that the in-plane component of each electric field's pulse $\mathbf{E}(t)$ is antisymmetric with respect to its central point, and hence vanishes at that point (as Fig. 7b shows), readily follows from Eq. (19).

[14] In contrast to the single-frequency case (i.e. a monochromatic wave), we may avoid taking the real part of the complex function ($\mathbf{E}_\omega e^{-i\omega t}$) by requiring that in Eq. (51), $\mathbf{E}_{-\omega} = \mathbf{E}_\omega^*$. However, it is important to remember the factor ½ required for the transition to a monochromatic wave of frequency ω_0 and with real amplitude $\mathbf{E}_0$: $\mathbf{E}_\omega = \mathbf{E}_0 [\delta(\omega - \omega_0) + \delta(\omega + \omega_0)]/2$.

This expression may be plugged into the formula for the total energy of the radiation pulse (i.e. of the loss of particle's energy $\mathscr{E}$) per unit solid angle:[15]

$$-\frac{d\mathscr{E}}{d\Omega} \equiv \int_{-\infty}^{+\infty} S_n(t) R^2 dt = \frac{R^2}{Z_0} \int_{-\infty}^{+\infty} \left|\mathbf{E}(t)\right|^2 dt \,. \tag{10.52}$$

This substitution, followed by a natural change of the integration order, yields

$$-\frac{d\mathscr{E}}{d\Omega} = \frac{R^2}{Z_0} \int_{-\omega}^{+\omega} d\omega \int_{-\omega}^{+\omega} d\omega' \, \mathbf{E}_\omega \cdot \mathbf{E}_{\omega'} \int_{-\infty}^{+\infty} dt \; e^{-i(\omega+\omega')t} \,. \tag{10.53}$$

But the inner integral (over t) is just $2\pi\delta(\omega + \omega')$.[16] This delta function kills one of the frequency integrals (say, one over ω'), and Eq. (53) gives us a result that may be recast as

$$-\frac{d\mathscr{E}}{d\Omega} = \int_0^{+\infty} I(\omega) d\omega, \qquad \text{with } I(\omega) \equiv \frac{4\pi R^2}{Z_0} \mathbf{E}_\omega \cdot \mathbf{E}_{-\omega} \equiv \frac{4\pi R^2}{Z_0} \mathbf{E}_\omega \cdot \mathbf{E}_\omega^*, \tag{10.54}$$

where the evident frequency symmetry of the scalar product $\mathbf{E}_\omega \cdot \mathbf{E}_{-\omega}$ has been utilized to fold the integral of $I(\omega)$ to positive frequencies only. The first of Eqs. (54) makes the physical sense of the function $I(\omega)$ very clear: this is the so-called *spectral density* of the electromagnetic radiation (per unit solid angle).[17]

To calculate the spectral density, we can express the function $\mathbf{E}_\omega$ via $\mathbf{E}(t)$ using the Fourier transform reciprocal to Eq. (51):

$$\mathbf{E}_\omega = \frac{1}{2\pi} \int_{-\infty}^{+\infty} \mathbf{E}(t) e^{i\omega t} dt \,. \tag{10.55}$$

In the particular case of radiation by a single point charge, we may use here the second (radiative) term of Eq. (19):

$$\mathbf{E}_\omega = \frac{1}{2\pi} \frac{q}{4\pi\varepsilon_0} \frac{1}{cR} \int_{-\infty}^{+\infty} \left[\frac{\mathbf{n} \times \{(\mathbf{n} - \boldsymbol{\beta}) \times \dot{\boldsymbol{\beta}}\}}{(1 - \boldsymbol{\beta} \cdot \mathbf{n})^3} \right]_{\text{ret}} e^{i\omega t} dt \,. \tag{10.56}$$

Since the vectors $\mathbf{n}$ and $\boldsymbol{\beta}$ are more natural functions of the radiation's emission (retarded) time t_{ret}, let us use Eqs. (5) and (16) to exclude the observation time t from this integral:

$$\mathbf{E}_\omega = \frac{q}{4\pi\varepsilon_0} \frac{1}{2\pi} \frac{1}{cR} \int_{-\infty}^{+\infty} \left[\frac{\mathbf{n} \times \{(\mathbf{n} - \boldsymbol{\beta}) \times \dot{\boldsymbol{\beta}}\}}{(1 - \boldsymbol{\beta} \cdot \mathbf{n})^2} \right]_{\text{ret}} \exp\left\{ i\omega \left(t_{\text{ret}} + \frac{R_{\text{ret}}}{c} \right) \right\} dt_{\text{ret}} \,. \tag{10.57}$$

Assuming that the observer is sufficiently far from the particle,[18] we may treat the unit vector $\mathbf{n}$ as a constant and also use the approximation (8.19) to reduce Eq. (57) to

---

[15] Note that the expression under this integral differs from $d\mathscr{P}/d\Omega$ defined by Eq. (29) by the absence of the term $(1 - \boldsymbol{\beta}\cdot\mathbf{n}) = \partial t_{\text{ret}}/\partial t$ – see Eq. (16). This is natural because now we are calculating the wave energy arriving at the observation point $\mathbf{r}$ during the time interval dt rather than dt_{ret}.

[16] See, e.g. MA Eq. (14.4).

[17] The notion of spectral density may be readily generalized to random processes – see, e.g., SM Sec. 5.4.

[18] According to the estimate (49), for a synchrotron radiation pulse, this restriction requires the observer to be much farther than $\Delta r' \sim c\Delta t \sim R/\gamma^3$ from the particle. With the values $R \sim 10^4$ m and $\gamma \sim 10^5$ mentioned above, $\Delta r' \sim 10^{-11}$ m, so this requirement is satisfied for any realistic radiation detector.

$$\mathbf{E}_\omega = \frac{q}{4\pi\varepsilon_0}\frac{1}{2\pi}\frac{1}{cR}\exp\left\{\frac{i\omega r}{c}\right\}\int_{-\infty}^{+\infty}\left[\frac{\mathbf{n}\times\{(\mathbf{n}-\boldsymbol{\beta})\times\dot{\boldsymbol{\beta}}\}}{(1-\boldsymbol{\beta}\cdot\mathbf{n})^2}\exp\left\{i\omega\left(t-\frac{\mathbf{n}\cdot\mathbf{r}'}{c}\right)\right\}\right]_{\rm ret} dt_{\rm ret}\,. \tag{10.58}$$

Plugging this expression into Eq. (54), and then using the definitions $c \equiv 1/(\varepsilon_0\mu_0)^{1/2}$ and $Z_0 \equiv (\mu_0/\varepsilon_0)^{1/2}$, we get[19]

$$I(\omega) = \frac{Z_0 q^2}{16\pi^3}\left|\int_{-\infty}^{+\infty}\left[\frac{\mathbf{n}\times\{(\mathbf{n}-\boldsymbol{\beta})\times\dot{\boldsymbol{\beta}}\}}{(1-\boldsymbol{\beta}\cdot\mathbf{n})^2}\exp\left\{i\omega\left(t-\frac{\mathbf{n}\cdot\mathbf{r}'}{c}\right)\right\}\right]_{\rm ret} dt_{\rm ret}\right|^2. \tag{10.59}$$

This result may be further simplified by noticing that the fraction before the exponent may be represented as a full derivative over $t_{\rm ret}$,

$$\left[\frac{\mathbf{n}\times\{(\mathbf{n}-\boldsymbol{\beta})\times\dot{\boldsymbol{\beta}}\}}{(1-\boldsymbol{\beta}\cdot\mathbf{n})^2}\right]_{\rm ret} \equiv \left[\frac{\mathbf{n}\times\{(\mathbf{n}-\boldsymbol{\beta})\times d\boldsymbol{\beta}/dt\}}{(1-\boldsymbol{\beta}\cdot\mathbf{n})^2}\right]_{\rm ret} \equiv \frac{d}{dt}\left[\frac{\mathbf{n}\times(\mathbf{n}\times\boldsymbol{\beta})}{1-\boldsymbol{\beta}\cdot\mathbf{n}}\right]_{\rm ret}, \tag{10.60}$$

and working out the resulting integral by parts. At this operation, the time differentiation of the parentheses in the exponent gives $d[t_{\rm ret} - \mathbf{n}\cdot\mathbf{r}'(t_{\rm ret})/c]/dt_{\rm ret} = (1 - \mathbf{n}\cdot\mathbf{u}/c)_{\rm ret} \equiv (1 - \boldsymbol{\beta}\cdot\mathbf{n})_{\rm ret}$, leading to the cancellation of the remaining factor in the denominator and hence to a very simple general result: [20]

Relativistic radiation: spectral density

$$I(\omega) = \frac{Z_0 q^2 \omega^2}{16\pi^3}\left|\int_{-\infty}^{+\infty}\left[\mathbf{n}\times(\mathbf{n}\times\boldsymbol{\beta})\exp\left\{i\omega\left(t-\frac{\mathbf{n}\cdot\mathbf{r}'}{c}\right)\right\}\right]_{\rm ret} dt_{\rm ret}\right|^2. \tag{10.61}$$

Now returning to the particular case of the synchrotron radiation, it is beneficial to choose the origin of time $t_{\rm ret}$ so that at $t_{\rm ret} = 0$, the angle θ between the vectors $\mathbf{n}$ and $\boldsymbol{\beta}$ takes its smallest value θ_0, i.e., in terms of Fig. 5, the vector $\mathbf{n}$ is within the $[y, z]$-plane. Fixing this direction of the axes so that they do not move, we can redraw that figure as shown in Fig. 8.

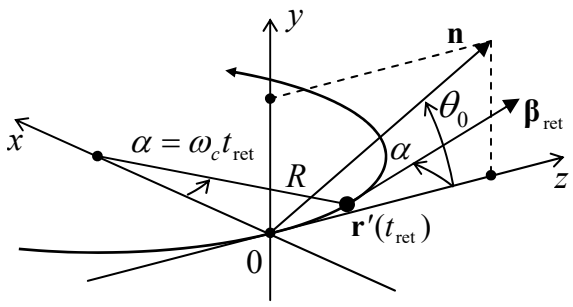


Fig. 10.8. Deriving the synchrotron radiation's spectral density. The vector $\mathbf{n}$ is static within the $[y, z]$-plane, while the vectors $\mathbf{r}'(t_{\rm ret})$ and $\boldsymbol{\beta}_{\rm ret}$ rotate, within the $[x, z]$-plane, with the angular velocity ω_c of the particle.

In this "lab" reference frame, the vector $\mathbf{n}$ does not depend on time, while the vectors $\mathbf{r}'(t_{\rm ret})$ and $\boldsymbol{\beta}_{\rm ret}$ do depend on it via the angle $\alpha \equiv \omega_c t_{\rm ret}$:

[19] Note that for our current purposes of calculation of the spectral density of radiation by a single particle, the factor $\exp\{i\omega r/c\}$ got canceled. However, as we have seen in Chapter 8, this factor plays a central role in the interference of radiation from several (many) sources. Such interference is important, in particular, in *undulators* and *free-electron lasers* – the devices to be (qualitatively) discussed below.

[20] Actually, this simplification is not occasional. According to Eq. (10b), the expression under the derivative in the last form of Eq. (60) is just the transverse component of the vector potential $\mathbf{A}$ (give or take a constant factor), and from the discussion in Sec. 8.2 we know that this component determines the electric dipole radiation of a system, which dominates the radiation in our current case of a single particle with a non-zero electric charge.

$$\mathbf{n} = \{0, \sin\theta_0, \cos\theta_0\}, \quad \mathbf{r}'(t_{\rm ret}) = \{R(1-\cos\alpha), 0, R\sin\alpha\}, \quad \boldsymbol{\beta}_{\rm ret} \equiv \{\beta\sin\alpha, 0, \beta\cos\alpha\}. \quad (10.62)$$

Now an easy multiplication yields

$$[\mathbf{n}\times(\mathbf{n}\times\boldsymbol{\beta})]_{\rm ret} = \beta\{\sin\alpha,\ \sin\theta_0\cos\theta_0\cos\alpha,\ -\sin^2\theta_0\sin\alpha\}, \quad (10.63)$$

$$\left[\exp\left\{i\omega\left(t - \frac{\mathbf{n}\cdot\mathbf{r}'}{c}\right)\right\}\right]_{\rm ret} = \exp\left\{i\omega\left(t_{\rm ret} - \frac{R}{c}\cos\theta_0\sin\alpha\right)\right\}. \quad (10.64)$$

As we already know, in the (most interesting) ultra-relativistic limit $\gamma >> 1$, most radiation is confined to short pulses, so only small angles $\alpha \sim \omega_c \Delta t_{\rm ret} \sim \gamma^{-1}$ may contribute to the integral in Eq. (61). Moreover, since most radiation goes to small angles $\theta \sim \theta_0 \sim \gamma^{-1}$, it makes sense to consider only such small angles. Expanding both trigonometric functions of these small angles, participating in parentheses of Eq. (64), into the Taylor series, and keeping only the leading terms, we get

$$t_{\rm ret} - \frac{R}{c}\cos\theta_0\sin\alpha \approx t_{\rm ret} - \frac{R}{c}\omega_c t_{\rm ret} + \frac{R}{c}\frac{\theta_0^2}{2}\omega_c t_{\rm ret} + \frac{R}{c}\frac{\omega_c^3}{6}t_{\rm ret}^3. \quad (10.65)$$

Since $(R/c)\omega_c = u/c = \beta \approx 1$, in the two last terms, we may approximate this parameter by 1. However, it is crucial to distinguish the difference between the two first terms, proportional to $(1 - \beta)t_{\rm ret}$, from zero; as we have done before, we may approximate it with $t_{\rm ret}/2\gamma^2$. On the right-hand side of Eq. (63), which does not have such a critical difference, we may be bolder, taking[21]

$$\beta\{\sin\alpha, \sin\theta_0\cos\theta_0\cos\alpha, -\sin^2\theta_0\sin\alpha\} \approx \{\alpha, \theta_0, 0\} \equiv \{\omega_c t_{\rm ret}, \theta_0, 0\}. \quad (10.66)$$

As a result, Eq. (61) is reduced to

$$I(\omega) = \frac{Z_0 q^2}{16\pi^3}\left|a_x\mathbf{n}_x + a_y\mathbf{n}_y\right|^2 \equiv \frac{Z_0 q^2}{16\pi^3}\left(\left|a_x\right|^2 + \left|a_y\right|^2\right), \quad (10.67)$$

where a_x and a_y are the following dimensionless factors:

$$\begin{aligned} a_x &\equiv \omega\int_{-\infty}^{+\infty}\omega_c t_{\rm ret}\exp\left\{\frac{i\omega}{2}\left((\theta_0^2+\gamma^{-2})t_{\rm ret} + \frac{\omega_c^2}{3}t_{\rm ret}^3\right)\right\}dt_{\rm ret}, \\ a_y &\equiv \omega\int_{-\infty}^{+\infty}\theta_0\exp\left\{\frac{i\omega}{2}\left((\theta_0^2+\gamma^{-2})t_{\rm ret} + \frac{\omega_c^2}{3}t_{\rm ret}^3\right)\right\}dt_{\rm ret}, \end{aligned} \quad (10.68)$$

that describe the frequency spectra of two components of the synchrotron radiation, with mutually perpendicular polarization planes. Defining the following dimensionless parameter

$$\nu \equiv \frac{\omega}{3\omega_c}\left(\theta_0^2 + \gamma^{-2}\right)^{3/2}, \quad (10.69)$$

[21] This expression confirms that the in-plane (x) component of the electric field is an odd function of $t_{\rm ret}$ and hence of $t - t_0$ (see its sketch in Fig. 7b), while the normal (y) component is an even function of this difference. Also, note that for an observer exactly in the rotation plane ($\theta_0 = 0$) the latter component equals zero for all times – the fact which could be predicted from the very beginning because of the evident mirror symmetry of the problem with respect to the particle's rotation plane.

which is proportional to the observation frequency, and changing the integration variable to $\xi \equiv \omega_c t_{ret}/(\theta_0^2 + \gamma^{-2})^{1/2}$, the integrals (68) may be reduced to the modified Bessel functions of the second kind, but with fractional indices:

$$a_x = \frac{\omega}{\omega_c}\left(\theta_0^2 + \gamma^{-2}\right)\int_{-\infty}^{+\infty} \xi \exp\left\{\frac{3}{2} i\nu\left(\xi + \frac{\xi^3}{3}\right)\right\} d\xi = \frac{2\sqrt{3}\, i}{\left(\theta_0^2 + \gamma^{-2}\right)^{1/2}} \nu K_{2/3}(\nu),$$
$$a_y = \frac{\omega}{\omega_c}\theta_0\left(\theta_0^2 + \gamma^{-2}\right)^{1/2}\int_{-\infty}^{+\infty} \exp\left\{\frac{3}{2} i\nu\left(\xi + \frac{\xi^3}{3}\right)\right\} d\xi = \frac{2\sqrt{3}\,\theta_0}{\theta_0^2 + \gamma^{-2}} \nu K_{1/3}(\nu) \qquad (10.70)$$

Figure 9a shows the dependence of the Bessel factors defining the amplitudes a_x and a_y on the normalized observation frequency ν. It shows that the radiation intensity changes with frequency relatively slowly (note the log-log scale of the plot!) until the normalized frequency defined by Eq. (69) is increased beyond ~1. For the most important observation angles $\theta_0 \sim \gamma$, this means that our estimate (50) is indeed correct, though formally the frequency spectrum extends to infinity.[22]

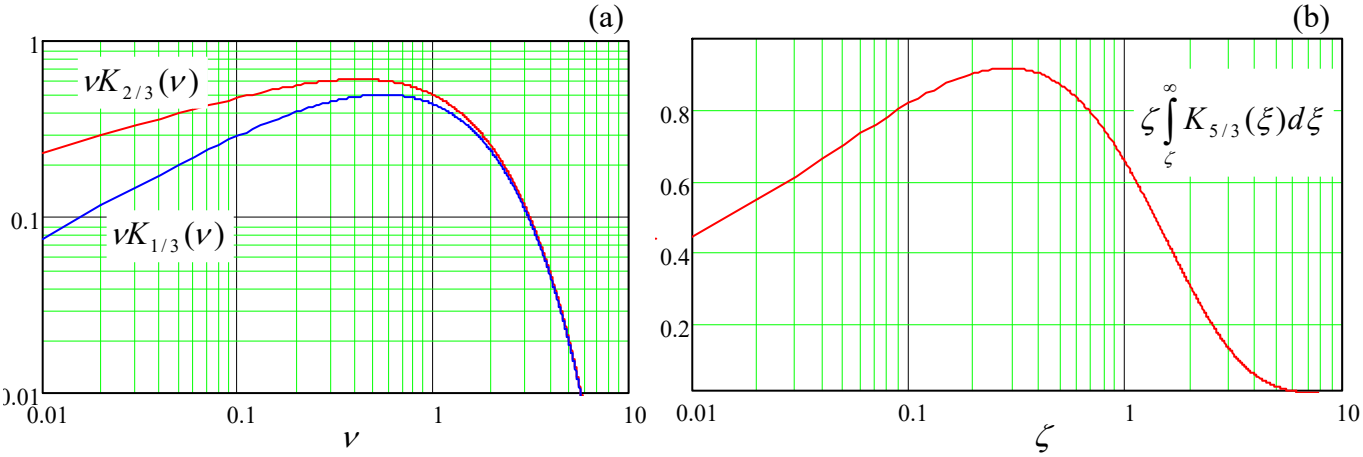


Fig. 10.9. The frequency spectra of: (a) two components of the synchrotron radiation, at a fixed angle θ_0, and (b) its total (polarization- and angle-averaged) intensity.

Naturally, the spectral density integrated over the full solid angle exhibits a similar frequency behavior. Without performing the integration,[23] let me just give the result (also valid for $\gamma >> 1$ only) for the reader's reference:

$$\oint_{4\pi} I(\omega) d\Omega = \frac{\sqrt{3}}{4\pi} q^2 \gamma \zeta \int_{\zeta}^{\infty} K_{5/3}(\xi) d\xi, \qquad \text{where } \zeta \equiv \frac{2}{3}\frac{\omega}{\omega_c \gamma^3}. \qquad (10.71)$$

Figure 9b shows the dependence of this integral on the normalized frequency ζ. (This plot is sometimes called the "universal flux curve".) In accordance with the estimate (50), it reaches the maximum at

[22] The law of the spectral density decrease at large ν may be readily obtained from the second of Eqs. (2.158), which is valid even for any (even non-integer) Bessel function index n: $a_x \propto a_y \propto \nu^{1/2}\exp\{-\nu\}$. Here the exponential factor is certainly the most important one.

[23] For that, and many other details, the interested reader may be referred, for example, to the fundamental review collection by E. Koch *et al*. (eds.) *Handbook on Synchrotron Radiation* (in 5 vols.), North-Holland, 1983-1991, or to a more concise monograph by A. Hofmann, *The Physics of Synchrotron Radiation*, Cambridge U. Press, 2007.

$$\zeta_{\max} \approx 0.3, \quad \text{i.e.}\, \omega_{\max} \approx \frac{\omega_c}{2}\gamma^3. \tag{10.72}$$

For example, in the National Synchrotron Light Source (NSLS-II) in the Brookhaven National Laboratory near our SBU campus, with its ring's circumference of 792 m, the electron revolution period $\mathcal{T}$ is 2.64 μs. With $\omega_c = 2\pi/\mathcal{T} \approx 2.4\times10^6$ s^{-1}, for the achieved $\gamma \approx 6\times10^3$ ($\mathscr{E} \approx 3$ GeV), we get $\omega_{\max} \sim 3\times10^{17}$ s^{-1}, i.e. the photon energy $\hbar\omega_{\max} \sim 200$ eV corresponding to soft X-rays. In light of this estimate, the reader may be surprised by Fig. 10, which shows the calculated spectra of the radiation that this facility was designed to produce, with the intensity maxima at photon energies up to a few keV.

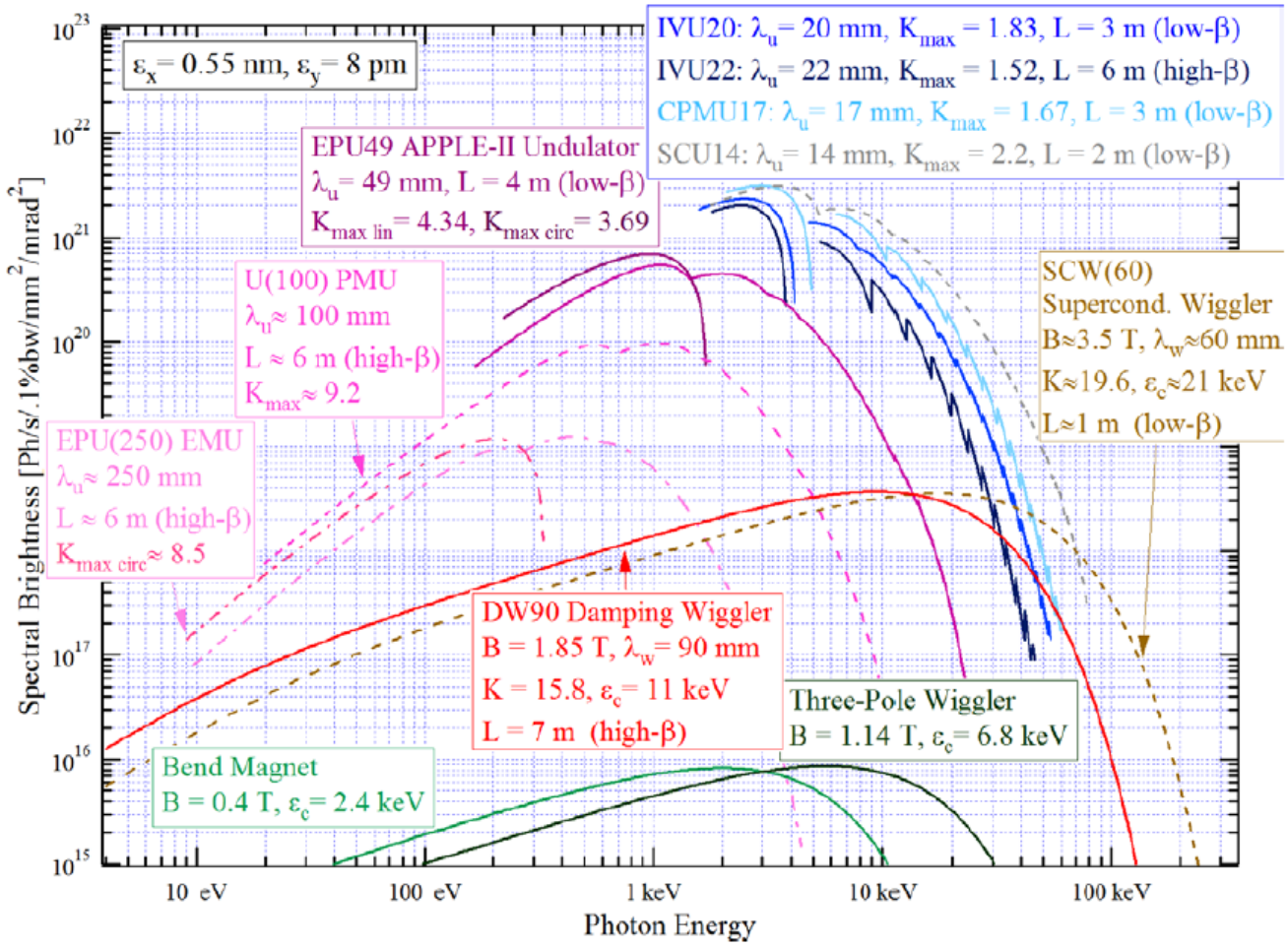


Fig. 10.10. Design brightness of various synchrotron radiation sources of the NSLS-II facility. For the bend magnets and wigglers, the "brightness" may be obtained by multiplication of the one-pulse spectral density $I(\omega)$ calculated above, by the number of electrons passing the source per second. (Note the non-SI units used by the synchrotron radiation community.) However, for undulators, there is an additional factor due to the partial coherence of radiation – see below. (Adapted from the document *NSLS-II Source Properties and Floor Layout* that was available online at https://www.bnl.gov/ps/docs/pdf/SourceProperties.pdf in 2011-2020.)

The reason for this discrepancy is that in the NLLS-II, and in all modern synchrotron light sources, most radiation is produced not by the circular orbit itself (which is, by the way, not exactly

circular, but consists of a series of straight and bend-magnet sections), but by such bend sections, and the devices called *wigglers* and *undulators*: strings of several strong magnets with alternating field direction (Fig. 11), that induce periodic bending (wiggling") of the electron's trajectory, with the synchrotron radiation emitted at each bend.

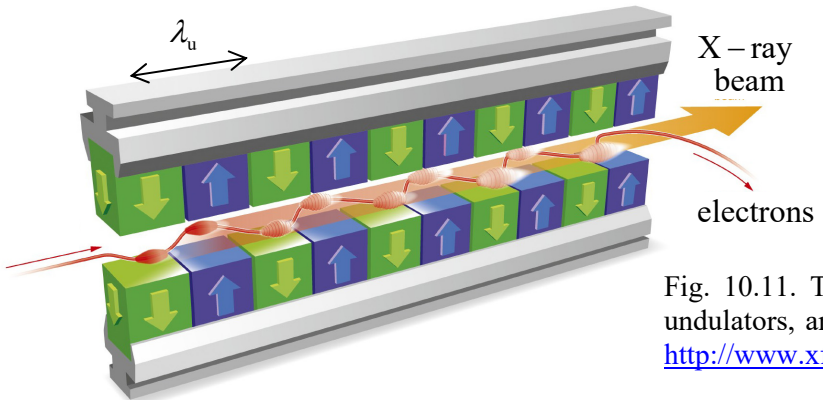


Fig. 10.11. The generic structure of the wigglers, undulators, and free-electron lasers. (Adapted from http://www.xfel.eu/overview/how_does_it_work/.)

The difference between the wigglers and the undulators is more quantitative than qualitative: the former devices have a larger spatial period λ_u (the distance between the adjacent magnets of the same polarity, see Fig. 11), giving enough space for the electron beam to bend by an angle larger than γ^{-1}, i.e. larger than the radiation cone's width. As a result, the radiation reaches an in-plane observer as a periodic sequence of individual pulses – see Fig. 12a.

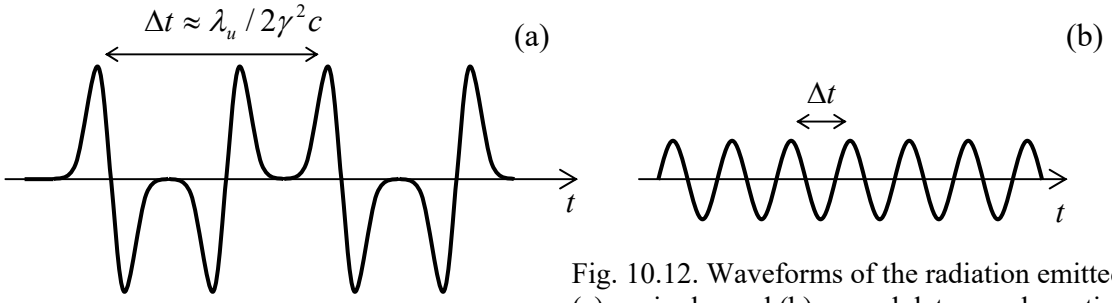


Fig. 10.12. Waveforms of the radiation emitted by (a) a wiggler and (b) an undulator – schematically.

The shape of each pulse, and hence its frequency spectrum, are essentially similar to those discussed above,[24] but with much higher local values of ω_c and hence ω_{max} – see Fig. 10. Another difference is a much higher frequency of the pulses. Indeed, the fundamental Eq. (16) allows us to calculate the time distance between them, for the observer, as

$$\Delta t \approx \frac{\partial t}{\partial t_{ret}} \Delta t_{ret} \approx (1-\beta)\frac{\lambda_u}{u} \approx \frac{1}{2\gamma^2}\frac{\lambda_u}{c} << \frac{\lambda_u}{c}, \qquad (10.73)$$

[24] Indeed, the period λ_u is typically a few centimeters (see the numbers in Fig. 10), i.e. is much larger than the interval $\Delta r' \sim R/\gamma^3$ estimated above. Hence the synchrotron radiation results may be applied locally, to each electron beam's bend. (In this context, a simple problem for the reader: use Eqs. (19) and (63) to explain the difference between shapes of the in-plane electric field pulses emitted at opposite magnetic poles of the wiggler, which is schematically shown in Fig. 12a.)

where the first two relations are valid at $\lambda_u << R$ (the relation typically satisfied very well, see the numbers in Fig. 10), and the last two relations assume the ultra-relativistic limit. As a result, the radiation intensity, which is proportional to the number of poles, is much higher than that from the bend magnets – see Fig. 10 again.

The situation is different in undulators – similar structures with a smaller spatial period λ_u, in which the electron's velocity vector oscillates with an angular amplitude smaller than γ^{-1}. As a result, the radiation pulses overlap (Fig. 12b), and the radiation waveform is closer to the sinusoidal one. As a result, the radiation spectrum narrows to the central frequency[25]

$$\omega_0 = \frac{2\pi}{\Delta t} \approx 2\gamma^2 \frac{2\pi c}{\lambda_u}. \tag{10.74}$$

For example, for the LSNL-II undulators with $\lambda_u = 2$ cm, this formula predicts a radiation peak at phonon energy $\hbar\omega_0 \approx 4$ keV, in reasonable agreement with the quantitative calculation results shown in Fig. 10.[26] Due to the spectrum narrowing, the undulator's radiation intensity is higher than that of wigglers using the same electron beam.

This spectrum-narrowing trend is brought to its logical conclusion in the so-called *free-electron lasers*[27] whose basic structure is the same as that of wigglers and undulators (Fig. 11), but the radiation at each beam bend is so intense and narrow-focused that it affects the electron motion downstream of the radiation cone. As a result, the radiation spectrum narrows around the central frequency (74), and its power grows as a square of the number N of electrons in the structure (rather than proportionately to N in wigglers and undulators).

Finally, note that wigglers, undulators, and free-electron lasers may be also used at the end of a linear electron accelerator (such as SLAC) which, as was noted above, may provide extremely high values of γ, and hence radiation frequencies, due to the smallness of radiation energy losses at the electron acceleration stage. Very unfortunately, I do not have time/space to discuss the (very interesting) physics of these devices in more detail.[28]

10.4. Bremsstrahlung and Coulomb losses

Surprisingly, a very similar mechanism of radiation by charged particles works on a much smaller spatial scale, namely at their scattering by charged particles of the propagation medium. This

[25] This important formula may be also derived in the following way. Due to the relativistic length contraction (9.20), the undulator structure period as perceived by beam electrons is $\lambda' = \lambda_u/\gamma$, so the central frequency of the radiation in the reference frame moving with the electrons is $\omega_0' = 2\pi c/\lambda' = 2\pi c\gamma/\lambda_u$. For the lab-frame observer, this frequency is Doppler-upshifted in accordance with Eq. (9.44): $\omega_0 = \omega_0'[(1 + \beta)/(1 - \beta)]^{1/2} \approx 2\gamma\omega_0'$, giving the same result as Eq. (74).

[26] Some of the difference is due to the fact that those plots show the spectral density of the number of *photons* $n = \mathcal{E}/\hbar\omega$ per second, which peaks at a frequency below that of the density of power, i.e. of the *energy* $\mathcal{E}$ per second.

[27] This name is somewhat misleading, because in contrast to the usual ("quantum") lasers, a free-electron laser is essentially a classical device, and the dynamics of electrons in it is very similar to that in vacuum-tube microwave generators, such as the magnetrons briefly discussed in Sec. 9.6.

[28] The interested reader may be referred, for example, to either P. Luchini and H. Motz, *Undulators and Free-electron Lasers*, Oxford U. Press, 1990; or E. Salin *et al*., *The Physics of Free Electron Lasers*, Springer, 2000.

effect, traditionally called by its German name *bremsstrahlung* ("brake radiation"), is responsible, in particular, for the continuous part of the frequency spectrum of the radiation produced in standard vacuum X-ray tubes, at the electron collisions with a metallic "anticathode". [29]

The bremsstrahlung in condensed matter is generally a rather complicated phenomenon because of the simultaneous involvement of many particles, and (frequently) some quantum electrodynamic effects. This is why I will give only a very brief glimpse at the theoretical description of this effect, for the simplest case when the scattering of incoming, relatively light charges (such as electrons, protons, α-particles, etc.) is produced by atomic nuclei, which remain virtually immobile during the scattering event (Fig. 13a). This is a reasonable approximation if the energy of incoming particles is not too low; otherwise, most scattering is produced by atomic electrons whose dynamics is substantially quantum – see below.

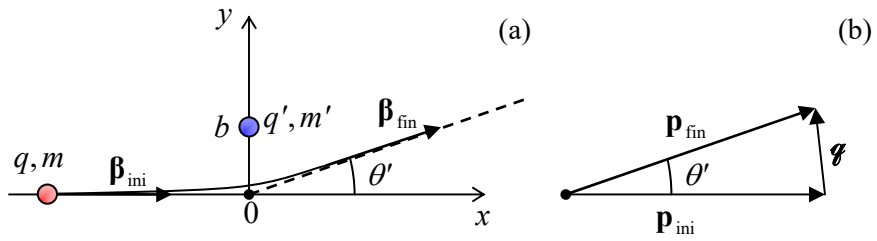


Fig. 10.13. The basic geometry of the bremsstrahlung and the Coulomb loss problems in the (a) direct and (b) reciprocal spaces.

To calculate the frequency spectrum of the radiation emitted during a single scattering event, it is convenient to use a byproduct of the last section's analysis, namely Eq. (59) with the replacement (60):[30]

$$I(\omega)=\frac{q^2}{4\pi\varepsilon_0}\frac{1}{4\pi^2 c}\left|\int_{-\infty}^{+\infty}\left[\frac{d}{dt}\frac{\mathbf{n}\times(\mathbf{n}\times\boldsymbol{\beta})}{1-\boldsymbol{\beta}\cdot\mathbf{n}}\exp\left\{i\omega\left(t-\frac{\mathbf{n}\cdot\mathbf{r}'}{c}\right)\right\}\right]_{\text{ret}}dt_{\text{ret}}\right|^2. \tag{10.75}$$

A typical duration τ of a single scattering event we are discussing is of the order of $\tau \equiv a_0/c \sim (10^{-10}$ m)/(3×10^8 m/s) $\sim 10^{-18}$ s in solids, and only an order of magnitude longer in gases at ambient conditions. This is why for most frequencies of interest, from zero all the way up to at least soft X-rays,[31] we can use the so-called *low-frequency approximation*, taking the exponent in Eq. (75) for 1 through the whole collision event, i.e. the integration interval. This approximation immediately yields

Brems-strahlung: single collision

$$I(\omega)=\frac{q^2}{4\pi\varepsilon_0}\frac{1}{4\pi^2 c}\left|\frac{\mathbf{n}\times(\mathbf{n}\times\boldsymbol{\beta}_{\text{fin}})}{1-\boldsymbol{\beta}_{\text{fin}}\cdot\mathbf{n}}-\frac{\mathbf{n}\times(\mathbf{n}\times\boldsymbol{\beta}_{\text{ini}})}{1-\boldsymbol{\beta}_{\text{ini}}\cdot\mathbf{n}}\right|^2. \tag{10.76}$$

[29] Such X-ray radiation had been first observed experimentally (though not correctly interpreted) by N. Tesla in 1887, i.e. before it was rediscovered and studied in detail by W. Röntgen.

[30] In publications on this topic (whose development peak was in the 1920s-1930s), the Gaussian units are more common, and the uppercase letter Z is usually reserved for expressing charges as multiples of the fundamental charge e, rather than for the wave impedance. This is why, in order to avoid confusion and facilitate the comparison with other texts, in this section I (while still staying with the SI units used throughout my series) will use the fraction $1/\varepsilon_0 c$, instead of its equivalent Z_0, for the free-space wave impedance, and write the coefficients in a form that makes the transfer to the Gaussian units elementary: it is sufficient to replace all $(qq'/4\pi\varepsilon_0)_{\text{SI}}$ with $(qq')_{\text{Gaussian}}$. In the (rare) cases when I spell out the charge values, I will use a different font: $q \equiv \mathfrak{Z}e$, $q' \equiv \mathfrak{Z}'e$.

[31] A more careful analysis shows that this approximation is actually quite reasonable up to much higher frequencies, of the order of γ^2/τ.

In the non-relativistic limit (β_{ini}, β_{fin} << 1), this formula is reduced to the following result:

$$I(\omega) = \frac{q^2}{4\pi\varepsilon_0}\frac{1}{4\pi^2 c}\frac{\mathcal{q}^2}{m^2c^2}\sin^2\theta \tag{10.77}$$

(which may be derived from Eq. (8.27) as well), where $\boldsymbol{\mathcal{q}}$ is the momentum transferred from the scattering center to the scattered charge (Fig. 13b):[32]

$$\boldsymbol{\mathcal{q}} \equiv \mathbf{p}_{fin} - \mathbf{p}_{ini} = m\Delta\mathbf{u} = mc\Delta\boldsymbol{\beta} = mc\left(\boldsymbol{\beta}_{fin} - \boldsymbol{\beta}_{ini}\right), \tag{10.78}$$

and θ (not to be confused with the particle scattering angle θ' shown in Fig. 13!) is the angle between the vector $\boldsymbol{\mathcal{q}}$ and the direction **n** toward the observer – at the collision moment.

The most important feature of the result (77)-(78) is the frequency-independent ("white") spectrum of the radiation, very typical for any rapid pulses that may be approximated as delta functions of time.[33] (Note, however, that Eq. (77) implies a fixed value of $\mathcal{q}$, so the statistics of this parameter, to be discussed in a minute, may "color" the radiation.)

Note also the "doughnut-shaped" angular distribution of the radiation, typical for non-relativistic systems, with the symmetry axis directed along the momentum transfer vector $\boldsymbol{\mathcal{q}}$. In particular, this means that in typical cases when $|\theta'| << 1$, i.e. $\mathcal{q} << p$, when the vector $\boldsymbol{\mathcal{q}}$ is nearly normal to the vector $\mathbf{p}_{ini}$ (see, e.g., the example shown in Fig. 13b), the bremsstrahlung produces a significant radiation flow in the direction back to the particle source – the fact significant for the operation of X-ray tubes.

Now integrating Eq. (77) over all wave propagation angles, just as we did for the instant radiation power in Sec. 8.2, we get the following spectral density of the particle energy loss,

$$-\frac{d\mathscr{E}}{d\omega} = \oint_{4\pi} I(\omega)\,d\Omega = \frac{2}{3\pi c}\frac{q^2}{4\pi^2\varepsilon_0}\frac{\mathcal{q}^2}{m^2c^2}. \tag{10.79}$$

In most applications of the bremsstrahlung theory (as in most scattering problems[34]), the impact parameter b (Fig. 13a), and hence the scattering angle θ' and the transferred momentum $\boldsymbol{\mathcal{q}}$, have to be

[32] Please note the font-marked difference between this variable ($\boldsymbol{\mathcal{q}}$) and the particle's electric charge (q).

[33] This is the basis, in particular, of the so-called *High-Harmonic Generation* (HHG) effect, discovered in 1977, which takes place at the irradiation of gases by intensive laser beams. The high electric field of the beam strips electrons from atoms, and accelerates them away from the remaining ions, just to slam them back into the same ions as the field's polarity changes in time. The electrons change their momentum sharply during their recombination with the ions, resulting in bremsstrahlung-like emission of short radiation pulses. The spectrum of radiation from each such pulse obeys Eq. (77), but since the ionization/acceleration/recombination cycles repeat periodically with the frequency ω_0 of the laser field, the final spectrum consists of many equidistant lines, with frequencies $n\omega_0$. The classical theory of the bremsstrahlung does not give a cutoff $\omega_{max} = n_{max}\omega_0$ of the spectrum; such a limit is imposed by quantum mechanics: $\hbar\omega_{max} \approx \hbar\omega_{max} + 3\mathscr{E}_p$, where the so-called *ponderomotive energy* $\mathscr{E}_p = (eE_0/\omega_0)^2/4m_e$ is the average kinetic energy given to a free electron by the periodic electric field of the laser beam, with amplitude E_0 – see, for example, M. Lewenstein *et al.*, *Phys. Rev. A* **49**, 2117 (1994). In practice, the HHG pulses may be shorter than 10^{-15} s, and n_{max} as high as ~100, enabling numerous applications of this effect.

[34] See, e.g., CM Sec. 3.5 and QM Sec. 3.3.

considered random. For elastic ($\beta_{ini} = \beta_{fin} \equiv \beta$) Coulomb collisions we can use the so-called *Rutherford formula* for the differential cross-section of scattering[35]

$$\frac{d\sigma}{d\Omega'} = \left(\frac{qq'}{4\pi\varepsilon_0}\right)^2 \left(\frac{1}{2pc\beta}\right)^2 \frac{1}{\sin^4(\theta'/2)}. \tag{10.80}$$

Here $d\sigma = 2\pi b db$ is the elementary area of the sample cross-section (as visible from the direction of the incident particles) corresponding to their scattering into an elementary body angle[36]

$$d\Omega' = 2\pi \sin\theta' \left| d\theta' \right|. \tag{10.81}$$

Differentiating the geometric relation, which is evident from Fig. 13b,

$$\hbar q = 2p \sin\frac{\theta'}{2}, \tag{10.82}$$

we may represent Eq. (80) in a more convenient form

$$\frac{d\sigma}{d(\hbar q)} = 8\pi \left(\frac{qq'}{4\pi\varepsilon_0}\right)^2 \frac{1}{u^2 (\hbar q)^3}. \tag{10.83}$$

Now combining Eqs. (79) and (83), we get

$$-\frac{d\mathcal{E}}{d\omega}\frac{d\sigma}{d(\hbar q)} = \frac{16}{3}\frac{q^2}{4\pi\varepsilon_0}\left(\frac{qq'}{4\pi\varepsilon_0 mc^2}\right)^2 \frac{1}{c\beta^2}\frac{1}{\hbar q}. \tag{10.84}$$

This product is called the *differential radiation cross-section*. When integrated over all values of $\hbar q$ (which is equivalent to averaging over all values of the impact parameter), it gives a convenient measure of the radiation intensity. Indeed, after the multiplication by the volume density n of independent scattering centers, such integral yields the particle's energy loss per unit bandwidth of radiation per unit path length, $-d^2\mathcal{E}/d\omega dx$. A minor problem here is that the integral of $1/\hbar q$ formally diverges at both infinite and zero values of $\hbar q$. However, these divergences are very weak (logarithmic), and the integral converges due to virtually any reason unaccounted for in our simple analysis. The standard (though slightly approximate) way to account for these effects is to write

Brems-strahlung: intensity

$$-\frac{d^2\mathcal{E}}{d\omega dx} \approx \frac{16}{3} n \frac{q^2}{4\pi\varepsilon_0}\left(\frac{qq'}{4\pi\varepsilon_0 mc^2}\right)^2 \frac{1}{c\beta^2} \ln\frac{\hbar q_{max}}{\hbar q_{min}}, \tag{10.85}$$

and then plug, instead of $\hbar q_{max}$ and $\hbar q_{min}$, the scales of the most important effects limiting the range of the transferred momentum's magnitude. In the classical-mechanics analysis, according to Eq. (82), $\hbar q_{max} = 2p \equiv 2mu$. To estimate $\hbar q_{min}$, let us note that the very small momentum transfer takes place when the impact parameter b is very large, and hence the effective scattering time $\tau \sim b/v$ is very long. Recalling the condition of the low-frequency approximation, we may associate $\hbar q_{min}$ with $\tau \sim 1/\omega$ and hence with $b \sim$

[35] See, e.g., CM Eq. (3.73) with $\alpha = qq'/4\pi\varepsilon_0$. In the form used in Eq. (80), the Rutherford formula is also valid for the small-angle scattering of relativistic particles, the criterion being $|\Delta\boldsymbol{\beta}| << 2/\gamma$.

[36] Again, the angle θ' and the differential $d\Omega'$, describing the scattered *particles* (see Fig. 13) should not be confused with the parameters θ and $d\Omega$ describing the *radiation* emitted at the scattering event.

$u\tau \sim v/\omega$. Since for the small scattering angles, $\mathcal{q}$ is close to the impulse $F\tau \sim (qq'/4\pi\varepsilon_0 b^2)\tau$ of the Coulomb force, we get the estimate $\mathcal{q}_{\min} \sim (qq'/4\pi\varepsilon_0)\omega/u^2$, and Eq. (85) should be used with

$$\ln\frac{\mathcal{q}_{\max}}{\mathcal{q}_{\min}} = \ln\left(\frac{2mu^3}{\omega}\Big/\frac{qq'}{4\pi\varepsilon_0}\right). \tag{10.86}$$

Classical brems-strahlung

This is *Bohr's formula* for what is called the *classical bremsstrahlung*. We see that the low momentum cutoff indeed makes the spectrum slightly colored, with more energy going to lower frequencies. There is even a formal divergence at $\omega \to 0$; however, this divergence is integrable, so it does not present a problem for finding the total energy radiative losses ($-d\mathcal{E}/dx$) as an integral of Eq. (86) over all radiated frequencies ω. A larger problem for this procedure is the upper integration limit, $\omega \to \infty$, at which the integral diverges. This means that our approximate description, which considers the collision as an elastic process, becomes invalid and needs to be amended by taking into account the difference between the initial and final kinetic energies of the particle due to radiation of the energy quantum $\hbar\omega$ of the emitted photon, so

$$\frac{p_{\text{ini}}^2}{2m} - \frac{p_{\text{fin}}^2}{2m} = \hbar\omega, \qquad \text{i.e. } \frac{p_{\text{ini}}^2}{2m} = \mathcal{E}, \quad \frac{p_{\text{fin}}^2}{2m} = \mathcal{E} - \hbar\omega, \qquad . \tag{10.87}$$

As a result, taking into account that the minimum and maximum values of $\mathcal{q}$ correspond to, respectively, the parallel and antiparallel alignments of the vectors $\mathbf{p}_{\text{ini}}$ and $\mathbf{p}_{\text{fin}}$, we get

$$\ln\frac{\mathcal{q}_{\max}}{\mathcal{q}_{\min}} = \ln\frac{p_{\text{ini}} + p_{\text{fin}}}{p_{\text{ini}} - p_{\text{fin}}} \equiv \ln\frac{\left(p_{\text{ini}} + p_{\text{fin}}\right)^2/2m}{\left(p_{\text{ini}}^2 - p_{\text{fin}}^2\right)/2m} = \ln\frac{\left[\mathcal{E}^{1/2} + \left(\mathcal{E} - \hbar\omega\right)^{1/2}\right]^2}{\hbar\omega}, \tag{10.88}$$

Quantum brems-strahlung

Plugged into Eq. (85), this expression yields the so-called *Bethe-Heitler formula* for *quantum bremsstrahlung*.[37] Note that in this approach, $\mathcal{q}_{\max}$ is close to that of the classical approximation, but $\mathcal{q}_{\min}$ is of the order of $\hbar\omega/u$, so

$$\frac{\mathcal{q}_{\min}\big|_{\text{classical}}}{\mathcal{q}_{\min}\big|_{\text{quantum}}} \sim \frac{\alpha\mathcal{Z}\mathcal{Z}'}{\beta}, \tag{10.89}$$

where $\mathcal{Z}$ and $\mathcal{Z}'$ are the particles' charges in the units of e, and α is the dimensionless *fine structure* ("Sommerfeld") *constant*,

$$\alpha \equiv \frac{e^2}{4\pi\varepsilon_0\hbar c}\Big|_{\text{SI}} = \frac{e^2}{\hbar c}\Big|_{\text{Gaussian}} \approx \frac{1}{137} << 1, \tag{10.90}$$

which is one of the basic notions of quantum mechanics.[38] Due to the smallness of the constant, the ratio (89) is below 1 for most cases of practical interest, and since the integral of (84) over $\mathcal{q}$ is limited by the largest of all possible cutoffs $\mathcal{q}_{\min}$, it is the Bethe-Heitler formula that should be used.

[37] The modifications of this formula necessary for the relativistic description are surprisingly minor – see, e.g., Chapter 15 in J. Jackson, Classical Electrodynamics, 3rd ed., Wiley 1999. For even more detail, the standard reference monograph on bremsstrahlung is W. Heitler, *The Quantum Theory of Radiation*, 3rd ed., Oxford U. Press 1954 (reprinted in 1984 and 2010 by Dover).

[38] See, e.g., QM Secs. 4.4, 6.3, 6.4, 9.3, 9.5, and 9.7.

Now nothing prevents us from calculating the total radiative losses of energy per unit length:

$$-\frac{d\mathscr{E}}{dx}=\int_0^{\infty}\left(-\frac{d^2\mathscr{E}}{d\omega dz}\right)d\omega=\frac{16}{3}n\frac{q^2}{4\pi\varepsilon_0 c}\left(\frac{qq'}{4\pi\varepsilon_0 mc^2}\right)^2\frac{1}{\beta^2}2\int_0^{\omega_{\max}}\ln\frac{\mathscr{E}^{1/2}-(\mathscr{E}-\hbar\omega)^{1/2}}{(\hbar\omega)^{1/2}}d\omega, \quad (10.91)$$

where $\hbar\omega_{\max}=\mathscr{E}$ is the maximum energy of the radiation quantum. By introducing the dimensionless integration variable $\xi\equiv\hbar\omega/\mathscr{E}=2\hbar\omega/(mu^2/2)$, this integral is reduced to a table one,[39] and we get

$$-\frac{d\mathscr{E}}{dx}=\frac{16}{3}n\frac{q^2}{4\pi\varepsilon_0 c}\left(\frac{qq'}{4\pi\varepsilon_0 mc^2}\right)^2\frac{1}{\beta^2}\frac{u^2}{\hbar}\equiv\frac{16}{3}n\left(\frac{q'^2}{4\pi\varepsilon_0\hbar c}\right)\left(\frac{q^2}{4\pi\varepsilon_0}\right)^2\frac{1}{mc^2}. \quad (10.92)$$

Following my usual style, at this point I would give you an estimate of the losses for a typical case; however, let me first discuss a parallel particle energy loss mechanism, the so-called *Coulomb losses*, due to the transfer of mechanical impulse from the scattered particle to the scattering centers. (This energy eventually goes into an increase of the thermal energy of the scattering medium, rather than to the electromagnetic radiation.)

Using Eqs. (9.139) for the electric field of a linearly moving charge q, we can readily find the momentum it transfers to the counterpart charge q':[40]

$$\Delta p'=\left|(\Delta p')_y\right|=\left|\int_{-\infty}^{+\infty}(\dot{p}')_y\,dt\right|=\left|\int_{-\infty}^{+\infty}q'E_y\,dt\right|=\frac{qq'}{4\pi\varepsilon_0}\int_{-\infty}^{+\infty}\frac{\gamma b}{\left(b^2+\gamma^2u^2t^2\right)^{3/2}}dt=\frac{qq'}{4\pi\varepsilon_0}\frac{2}{bu}. \quad (10.93)$$

Hence, the kinetic energy acquired by the scattering particle (and hence to the loss of the energy $\mathscr{E}$ of the incident particle) is

$$-\Delta\mathscr{E}=\frac{(\Delta p')^2}{2m'}=\left(\frac{qq'}{4\pi\varepsilon_0}\right)^2\frac{2}{m'u^2b^2}. \quad (10.94)$$

Such elementary energy losses have to be summed up over all collisions, with random values of the impact parameter b. At the scattering center density n, the number of collisions per small path length dx per small range db is $dN=n\,2\pi b db\,dx$, so

Coulomb losses

$$-\frac{d\mathscr{E}}{dx}=-\int\Delta\mathscr{E}dN=n\left(\frac{qq'}{4\pi\varepsilon_0}\right)^2\frac{2}{m'u^2}2\pi\int_{b_{\min}}^{b_{\max}}\frac{db}{b}=4\pi n\left(\frac{qq'}{4\pi\varepsilon_0}\right)^2\frac{\ln B}{m'u^2}, \quad \text{where } B\equiv\frac{b_{\max}}{b_{\min}}. \quad (10.95)$$

Here, at the last step, the logarithmic integral over b was treated similarly to that over q in the bremsstrahlung theory. This approximation is adequate because the ratio $b_{\max}/b_{\min}$ is much larger than 1. Indeed, $b_{\min}$ may be estimated from $(\Delta p')_{\max}\sim p=\gamma mu$. For this value, Eq. (93) with $q'\sim q$ gives $b_{\min}\sim r_c$ (see Eq. (8.41) and its discussion), which, for elementary particles, is of the order of 10^{-15}m. On the other hand, for the most important case when the Coulomb energy absorbers are electrons (which, according to Eq. (94), are the most efficient ones, due to their very low mass m'), $b_{\max}$ may be estimated from the condition $\tau=b/\gamma u\sim 1/\omega_{\min}$, where $\omega_{\min}\sim 10^{16}$ s^{-1} is the characteristic frequency of electron

[39] See, e.g., MA Eq. (6.14).

[40] According to Eq. (9.139), $E_z=0$, while the net impulse of the longitudinal force $q'E_x$ is zero, so Eq. (93) gives the full momentum transfer.

transitions in atoms. (Quantum mechanics forbids such energy transfer at lower frequencies.) From here, we have the estimate $b_{max} \sim \gamma u/\omega_{min}$, so

$$B \equiv \frac{b_{max}}{b_{min}} \sim \frac{\gamma u}{r_c \omega_{min}}, \tag{10.96}$$

for $\gamma \sim 1$ and $u \sim c \approx 3\times10^8$ m/s giving $b_{max} \sim 3\times10^{-8}$ m, so $B \sim 10^9$ (give or take a couple of orders of magnitude – this does not change the estimate $\ln B \approx 20$ too much). [41]

Now we can compare the non-radiative Coulomb losses (95) with the radiative losses due to the bremsstrahlung, given by Eq. (92):

$$\frac{-d\mathscr{E}\big|_{radiation}}{-d\mathscr{E}\big|_{Coulomb}} \sim \alpha \mathscr{Z}\mathscr{Z}' \frac{m'}{m}\beta^2 \frac{1}{\ln B}, \tag{10.97}$$

Since $\alpha \sim 10^{-2} << 1$, for non-relativistic particles ($\beta << 1$) the bremsstrahlung losses of energy are much lower (that is why I did not want to rush with their estimates), and only for ultra-relativistic particles, the relation may be opposite.

According to Eqs. (95)-(96), for electron-electron scattering ($q = q' = -e$, $m = m' = m_e$),[42] at the value $n = 6\times10^{26}$ m^{-3} typical for air at ambient conditions, the characteristic length of energy loss,

$$l_c \equiv \frac{\mathscr{E}}{\left(-d\mathscr{E}/dx\right)}, \tag{10.98}$$

for electrons with kinetic energy $\mathscr{E} = 6$ keV is close to 2×10^{-4} m ≡ 0.2 mm. (This is why we need high vacuum in electron microscope columns and other vacuum electron devices.) Since $l_c \propto \mathscr{E}^2$, more energetic particles penetrate to matter deeper, until the bremsstrahlung steps in, and limits this trend at very high energies.

10.5. Density effects and the Cherenkov radiation

For condensed matter, the Coulomb loss estimate made in the last section is not quite suitable, because it is based on the upper cutoff $b_{max} \sim \gamma u/\omega_{min}$. For the example given above, the incoming electron velocity u is close to 5×10^7 m/s, and for the typical value $\omega_{min} \sim 10^{16}$ s^{-1} ($\hbar\omega_{min} \sim 10$ eV), this cutoff b_{max} is of the order of $\sim5\times10^{-9}$ m = 5 nm. Even for air at ambient conditions, this is somewhat larger than the average distance (~ 2 nm) between the molecules, so at the high end of the impact parameter range, at $b \sim b_{max}$, the Coulomb loss events in adjacent molecules are not quite independent, and the theory needs some corrections. For condensed matter, with much higher particle density n, most collisions satisfy the following condition:

[41] A quantum analysis (carried out by Hans Bethe in 1940) replaces, in Eq. (95), $\ln B$ with $\ln(2\gamma^2 mu^2/\hbar\langle\omega\rangle) - \beta^2$, where $\langle\omega\rangle$ is the average frequency of the atomic quantum transitions weight by their oscillator strength. This refinement does not change the estimate given below. Note that both the classical and quantum formulas describe a fast increase (as $1/\beta$) of the energy loss rate ($-d\mathscr{E}/dx$) at $\gamma \to 1$, and its slow increase (as $\ln\gamma$) at $\gamma \to \infty$, so the losses have a minimum at $(\gamma - 1) \sim 1$.

[42] Actually, the above analysis has neglected the change of momentum of the incident particle. This is legitimate at $m' << m$, but for $m = m'$ the change approximately doubles the energy losses. Still, this does not change the order of magnitude of the estimate.

$$nb^3 >> 1, \tag{10.99}$$

and the treatment of Coulomb collisions as a set of independent events is inadequate. However, this condition enables the opposite approach: treating the medium as a continuum. In the time-domain formulation used in the previous sections of this chapter, this would be a very complex problem, because it would require an explicit description of the medium dynamics. Here the frequency-domain approach, based on the Fourier transform in both time and space, helps a lot, provided that the functions $\varepsilon(\omega)$ and $\mu(\omega)$ are considered known – either calculated or taken from experiment. Let us have a good look at this approach because it gives some interesting (and practically important) results.

In Chapter 6, we have used the macroscopic Maxwell equations to derive Eqs. (6.118), which describe the time evolution of electrodynamic potentials in a linear medium with frequency-independent ε and μ. Looking for all functions participating in Eqs. (6.118) in the plane-wave expansion form[43]

$$f(\mathbf{r},t) = \int d^3k \int d\omega f_{\mathbf{k},\omega} e^{i(\mathbf{k}\cdot\mathbf{r}-\omega t)}, \tag{10.100}$$

and requiring all coefficients at similar exponents to be balanced, we get their Fourier images: [44]

$$\left(k^2 - \omega^2\varepsilon\mu\right)\phi_{\mathbf{k},\omega} = \frac{\rho_{\mathbf{k},\omega}}{\varepsilon}, \qquad \left(k^2 - \omega^2\varepsilon\mu\right)\mathbf{A}_{\mathbf{k},\omega} = \mu\mathbf{j}_{\mathbf{k},\omega}. \tag{10.101}$$

As was discussed in Chapter 7, in such a Fourier form, the macroscopic Maxwell theory remains valid even for dispersive (but isotropic and linear!) media, so Eqs. (101) may be generalized as

$$\left[k^2 - \omega^2\varepsilon(\omega)\mu(\omega)\right]\phi_{\mathbf{k},\omega} = \frac{\rho_{\mathbf{k},\omega}}{\varepsilon(\omega)}, \qquad \left[k^2 - \omega^2\varepsilon(\omega)\mu(\omega)\right]\mathbf{A}_{\mathbf{k},\omega} = \mu(\omega)\,\mathbf{j}_{\mathbf{k},\omega}, \tag{10.102}$$

An evident advantage of these equations is that their formal solution is elementary:

Field potentials in a linear medium

$$\phi_{\mathbf{k},\omega} = \frac{\rho_{\mathbf{k},\omega}}{\varepsilon(\omega)\left[k^2 - \omega^2\varepsilon(\omega)\mu(\omega)\right]}, \qquad \mathbf{A}_{\mathbf{k},\omega} = \frac{\mu(\omega)\,\mathbf{j}_{\mathbf{k},\omega}}{\left[k^2 - \omega^2\varepsilon(\omega)\mu(\omega)\right]}, \tag{10.103}$$

so the "only" remaining things to do is, first, to calculate the Fourier transforms of the functions $\rho(\mathbf{r}, t)$ and $\mathbf{j}(\mathbf{r}, t)$, describing stand-alone charges and currents, using the transform reciprocal to Eq. (100), with one factor $1/2\pi$ per each scalar dimension,

$$f_{\mathbf{k},\omega} = \frac{1}{(2\pi)^4}\int d^3r \int dt f(\mathbf{r},t) e^{-i(\mathbf{k}\cdot\mathbf{r}-\omega t)}, \tag{10.104}$$

and then to carry out the integration (100) of Eqs. (103).

For our problem of a single charge q uniformly moving through a medium with velocity $\mathbf{u}$,

$$\rho(\mathbf{r},t) = q\delta(\mathbf{r} - \mathbf{u}t), \qquad \mathbf{j}(\mathbf{r},t) = q\mathbf{u}\delta(\mathbf{r} - \mathbf{u}t), \tag{10.105}$$

[43] All integrals here and below are in infinite limits unless specified otherwise.

[44] As was discussed in Sec. 7.2, the Ohmic conductivity of the medium (generally, also a function of frequency) may be readily incorporated into the dielectric permittivity: $\varepsilon(\omega) \to \varepsilon_{\text{ef}}(\omega) + i\sigma(\omega)/\omega$. In this section, I will assume that such incorporation, which is especially natural for high frequencies, has been performed, so the current density $\mathbf{j}(\mathbf{r}, t)$ describes only stand-alone currents – for example, the current (105) of the incident particle.

the first task is easy:

$$\rho_{\mathbf{k},\omega} = \frac{q}{(2\pi)^4}\int d^3r \int dt\; q\delta(\mathbf{r}-\mathbf{u}t)e^{-i(\mathbf{k}\cdot\mathbf{r}-\omega t)} = \frac{q}{(2\pi)^4}\int e^{i(\omega t-\mathbf{k}\cdot\mathbf{u}t)}dt = \frac{q}{(2\pi)^3}\delta(\omega-\mathbf{k}\cdot\mathbf{u}). \quad (10.106)$$

Since the expressions (105) for $\rho(\mathbf{r}, t)$ and $\mathbf{j}(\mathbf{r}, t)$ differ only by a constant factor $\mathbf{u}$, it is clear that the absolutely similar calculation for the current gives

$$\mathbf{j}_{\mathbf{k},\omega} = \frac{q\mathbf{u}}{(2\pi)^3}\delta(\omega-\mathbf{k}\cdot\mathbf{u}). \quad (10.107)$$

Let us summarize what we have got by now, by plugging Eqs. (106)-(107) into Eqs. (103):

$$\phi_{\mathbf{k},\omega} = \frac{1}{(2\pi)^3}\frac{q\delta(\omega-\mathbf{k}\cdot\mathbf{u})}{\varepsilon(\omega)\left[k^2-\omega^2\varepsilon(\omega)\mu(\omega)\right]}, \quad \mathbf{A}_{\mathbf{k},\omega} = \frac{1}{(2\pi)^3}\frac{\mu(\omega)q\mathbf{u}\delta(\omega-\mathbf{k}\cdot\mathbf{u})}{\left[k^2-\omega^2\varepsilon(\omega)\mu(\omega)\right]} \equiv \varepsilon(\omega)\mu(\omega)\mathbf{u}\phi_{\mathbf{k},\omega}. \quad (10.108)$$

Now, at the last calculation step, namely the integration (100), we are starting to pay a heavy price for the easiness of the first steps. This is why let us think well about what exactly we need from it. First of all, for the calculation of power losses, the electric field is more convenient to use than the potentials, so let us calculate the Fourier images of **E** and **B**. Plugging the expansion (100) into the basic relations (6.7), and again requiring the balance of exponent's coefficients, we get

$$\mathbf{E}_{\mathbf{k},\omega} = -i\mathbf{k}\phi_{k,\omega} + i\omega\mathbf{A}_{\mathbf{k},\omega} = i\left[\omega\varepsilon(\omega)\mu(\omega)\mathbf{u}-\mathbf{k}\right]\phi_{\mathbf{k},\omega}, \quad \mathbf{B}_{\mathbf{k},\omega} = i\mathbf{k}\times\mathbf{A}_{\mathbf{k},\omega} = i\varepsilon(\omega)\mu(\omega)\mathbf{k}\times\mathbf{u}\phi_{\mathbf{k},\omega}, \quad (10.109)$$

so Eqs. (100) and (108) yield

$$\mathbf{E}(\mathbf{r},t) = \int d^3k\int d\omega\;\mathbf{E}_{\mathbf{k},\omega}e^{i(\mathbf{k}\cdot\mathbf{r}-\omega t)} = \frac{iq}{(2\pi)^3}\int d^3k\int d\omega\;\frac{\left[\omega\varepsilon(\omega)\mu(\omega)\mathbf{u}-\mathbf{k}\right]\delta(\omega-\mathbf{k}\cdot\mathbf{u})}{\varepsilon(\omega)\left[k^2-\omega^2\varepsilon(\omega)\mu(\omega)\right]}e^{i(\mathbf{k}\cdot\mathbf{r}-\omega t)}. \quad (10.110)$$

This formula may be rewritten as the temporal Fourier integral (51), with the following **r**-dependent complex amplitude:

$$\mathbf{E}_\omega(\mathbf{r}) = \int \mathbf{E}_{\mathbf{k},\omega}e^{i\mathbf{k}\cdot\mathbf{r}}d^3k = \frac{iq}{(2\pi)^3}\int\frac{\left[\omega\varepsilon(\omega)\mu(\omega)\mathbf{u}-\mathbf{k}\right]\delta(\omega-\mathbf{k}\cdot\mathbf{u})}{\varepsilon(\omega)\left[k^2-\omega^2\varepsilon(\omega)\mu(\omega)\right]}e^{i\mathbf{k}\cdot\mathbf{r}}d^3k. \quad (10.111)$$

Let us calculate the Cartesian components of this partial Fourier image $\mathbf{E}_\omega$, at a point separated by distance b from the particle's trajectory. Selecting the coordinates and time origin as shown in Fig. 3, we have $\mathbf{r} = \{0, b, 0\}$ and $\mathbf{u} = \{u, 0, 0\}$, so only E_x and E_y are different from zero. In particular, according to Eq. (111),

$$(E_x)_\omega = \frac{iq}{(2\pi)^3\varepsilon(\omega)}\int dk_x\int dk_y\int dk_z\;\frac{\omega\varepsilon(\omega)\mu(\omega)u-k_x}{k^2-\omega^2\varepsilon(\omega)\mu(\omega)}\delta(\omega-k_xu)\exp\{ik_yb\}. \quad (10.112)$$

The delta function kills one integral (over k_x) of the three, and we get

$$(E_x)_\omega = \frac{iq}{(2\pi)^3\varepsilon(\omega)u}\left[\omega\varepsilon(\omega)\mu(\omega)u-\frac{\omega}{u}\right]\int\exp\{ik_yb\}dk_y\int\frac{dk_z}{\omega^2/u^2+k_y^2+k_z^2-\omega^2\varepsilon(\omega)\mu(\omega)}. \quad (10.113)$$

The internal integral (over k_z) may be readily reduced to the table integral $\int d\xi/(1+\xi^2)$ in infinite limits, equal to π,[45] and the result represented as

$$(E_x)_\omega = -\frac{i\pi q\kappa^2}{(2\pi)^3\omega\varepsilon(\omega)}\int \frac{\exp\{ik_y b\}}{\left(k_y^2+\kappa^2\right)^{1/2}}dk_y, \tag{10.114}$$

where the parameter κ (generally, a complex function of frequency) is defined as[46]

Function $\kappa(\omega)$

$$\kappa^2(\omega) \equiv \omega^2\left[\frac{1}{u^2} - \varepsilon(\omega)\mu(\omega)\right]. \tag{10.115}$$

The last integral may be expressed via the modified Bessel function of the second kind:[47]

$$(E_x)_\omega = -\frac{iqu\kappa^2}{(2\pi)^2\omega\varepsilon(\omega)}K_0(\kappa b). \tag{10.116}$$

A very similar calculation yields

$$(E_y)_\omega = \frac{q\kappa}{(2\pi)^2\varepsilon(\omega)}K_1(\kappa b). \tag{10.117}$$

Now, instead of rushing to make the final integration (51) over ω to calculate $\mathbf{E}(t)$, let us realize that what we need most is the total energy loss through the whole time of the particle's passage over an elementary distance dx. According to Eq. (4.38), the energy loss per unit volume is

$$-\frac{d\mathscr{E}}{dV} = \int \mathbf{j}\cdot\mathbf{E}dt, \tag{10.118}$$

where $\mathbf{j}$ is the current of the bound charges in the medium, and should not be confused with the stand-alone incident-particle current (105). This integral may be readily expressed via the partial Fourier image $\mathbf{E}_\omega$ and the similarly defined image $\mathbf{j}_\omega$, just as it was done at the derivation of Eq. (54):

$$-\frac{d\mathscr{E}}{dV} = \int dt\int d\omega e^{-i\omega t}\int d\omega' e^{-i\omega' t}\mathbf{j}_\omega\cdot\mathbf{E}_{\omega'} = 2\pi\int d\omega\int d\omega'\,\mathbf{j}_\omega\cdot\mathbf{E}_{\omega'}\delta(\omega+\omega') = 2\pi\int \mathbf{j}_\omega\cdot\mathbf{E}_{-\omega}d\omega. \tag{10.119}$$

Let us incorporate the effective Ohmic conductivity $\sigma_{ef}(\omega)$ into the complex permittivity $\varepsilon(\omega)$ just as this was discussed in Sec. 7.2, using Eq. (7.46) to write

$$\mathbf{j}_\omega = \sigma_{ef}(\omega)\mathbf{E}_\omega = -i\omega\varepsilon(\omega)\mathbf{E}_\omega. \tag{10.120}$$

As a result, Eq. (119) yields

$$-\frac{d\mathscr{E}}{dV} = -2\pi i\int \varepsilon(\omega)\mathbf{E}_\omega\cdot\mathbf{E}_{-\omega}\omega d\omega = 4\pi\,\mathrm{Im}\int_0^\infty \varepsilon(\omega)\left|E_\omega\right|^2\omega d\omega. \tag{10.121}$$

(The last step was possible due to the property $\varepsilon(-\omega) = \varepsilon^*(\omega)$, which was discussed in Sec. 7.2.)

[45] See, e.g., MA Eq. (6.5a).

[46] The frequency-dependent parameter $\kappa(\omega)$ should not be confused with the dc low-frequency dielectric constant $\kappa \equiv \varepsilon(0)/\varepsilon_0$ that was discussed in Chapter 3.

[47] As a reminder, the main properties of these functions are listed in Sec. 2.7 – see, in particular, Fig. 2.22 and Eqs. (2.157)-(2.158).

Finally, just as in the last section, we have to average the energy loss rate over random values of the impact parameter b:

$$-\frac{d\mathscr{E}}{dx}=\int\left(-\frac{d\mathscr{E}}{dV}\right)d^2b\approx 2\pi\int_{b_{\min}}^{\infty}\left(-\frac{d\mathscr{E}}{dV}\right)bdb=8\pi^2\int_{b_{\min}}^{\infty}bdb\int_0^{\infty}\left(\left|E_x\right|_\omega^2+\left|E_y\right|_\omega^2\right)\mathrm{Im}\,\varepsilon(\omega)\omega d\omega\,. \quad (10.122)$$

Due to the (weak) divergence of the functions $K_0(\xi)$ and $K_1(\xi)$ at $\xi\to 0$, we have to cut the resulting integral over b at some $b_{\min}$ where our theory loses legitimacy. (On that limit, we are not doing much better than in the past section). Plugging in the calculated expressions (116) and (117) for the field components, swapping the integrals over ω and b, and using the recurrence relations (2.142), which are valid for all Bessel functions, we finally get:

$$-\frac{d\mathscr{E}}{dx}=\frac{2}{\pi}q^2\,\mathrm{Im}\int_0^{\infty}(\kappa^* b_{\min})K_1(\kappa^* b_{\min})K_0(\kappa^* b_{\min})\frac{d\omega}{\omega\varepsilon(\omega)}\,. \quad (10.123)$$

Radiation intensity

This general result is valid for a linear medium with arbitrary dispersion relations $\varepsilon(\omega)$ and $\mu(\omega)$. (The last function participates in Eq. (123) only via Eq. (115) that defines the parameter κ.) To get more concrete results, some particular model of the medium should be used. Let us explore the Lorentz-oscillator model that was discussed in Sec. 7.2, in its form (7.33) suitable for the transition to the quantum-mechanical description of atoms:

$$\varepsilon(\omega)=\varepsilon_0+\frac{nq'^2}{m}\sum_j\frac{f_j}{(\omega_j^2-\omega^2)-2i\omega\delta_j},\quad \text{with}\ \sum_j f_j=1;\qquad \mu(\omega)=\mu_0\,. \quad (10.124)$$

If the damping of the effective atomic oscillators is low, $\delta_j << \omega_j$, as it typically is, and the particle's speed u is much lower than the typical wave's phase velocity v (and hence than c!), then for most frequencies Eq. (115) gives

$$\kappa^2(\omega)\equiv\omega^2\left[\frac{1}{u^2}-\frac{1}{v^2(\omega)}\right]\approx\frac{\omega^2}{u^2}\,, \quad (10.125)$$

i.e. $\kappa\approx\kappa^*\approx\omega/u$ is virtually real. In this case, Eq. (123) may be reduced to Eq. (95) with

$$b_{\max}=\frac{1.123u}{\langle\omega\rangle}\,. \quad (10.126)$$

The good news here is that both approaches (the microscopic analysis of Sec. 4 and the macroscopic analysis of this section) give essentially the same result. The same fact may be also perceived as bad news: the treatment of the medium as a continuum does not give any new results here. The situation somewhat changes at relativistic velocities, at which such treatment provides noticeable corrections (called *density effects*), in particular reducing the energy loss estimates.

Let me, however, leave these details for special topic courses and focus on a much more important effect described by our formulas. Consider the dependence of the electric field components on the impact parameter b, i.e. on the closest distance between the particle's trajectory and the field observation point. At $b\to\infty$, we can use, in Eqs. (116)-(117), the asymptotic formula (2.158),

$$K_n(\xi)\to\left(\frac{\pi}{2\xi}\right)^{1/2}e^{-\xi},\ \text{at}\ \xi\to\infty\,, \quad (10.127)$$

to conclude that if $\kappa^2 > 0$, i.e. if κ is real, the complex amplitudes E_ω of both components E_x and E_y of the electric field decrease with b exponentially. However, let us consider what happens at frequencies where $\kappa^2(\omega) < 0$,[48] i.e.

$$\varepsilon(\omega)\mu(\omega) \equiv \frac{1}{v^2(\omega)} < \frac{1}{u^2} < \frac{1}{c^2} \equiv \varepsilon_0\mu_0 . \tag{10.128}$$

(This condition means that the particle's velocity is larger than the phase velocity of the waves at this particular frequency.) In this case, the parameter $\kappa(\omega)$ is purely imaginary, so the functions $\exp\{\kappa b\}$ in the asymptotes (127) of Eqs. (116)-(117) become just phase factors, and the field component amplitudes fall very slowly:

$$\left|E_x(\omega)\right| \propto \left|E_y(\omega)\right| \propto \frac{1}{b^{1/2}} . \tag{10.129}$$

This means that the Poynting vector drops as $1/b$, so its flux through a surface of a round cylinder of radius b, with its axis on the particle trajectory (i.e. the power flow from the particle), does not depend on b at all. This is an electromagnetic wave emission – the famous *Cherenkov radiation*.[49]

The direction **n** of its propagation may be readily found taking into account that at large distances from the particle's trajectory, the emitted wave has to be locally planar and transverse (**n**⊥**E**), so the so-called *Cherenkov angle* θ between the vector **n** and the particle's velocity **u** may be simply found from the ratio of the electric field components – see Fig. 14a:

$$\tan\theta = -\frac{E_x}{E_y} . \tag{10.130}$$

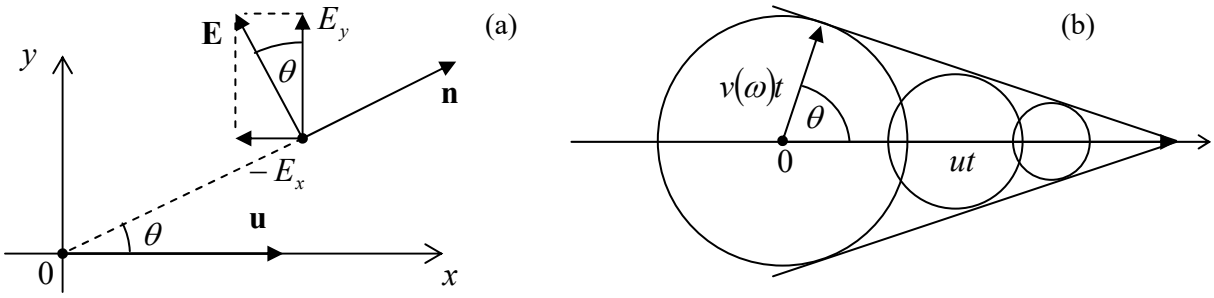


Fig. 10.14. (a) The Cherenkov radiation's propagation angle θ, and (b) its interpretation.

The ratio on the right-hand side of this relation may be calculated by plugging the asymptotic formula (127) into Eqs. (116) and (117) and calculating their ratio:

[48] Strictly speaking, the inequality $\kappa^2(\omega) < 0$ does not make sense for a medium with a complex product $\varepsilon(\omega)\mu(\omega)$, and hence complex $\kappa^2(\omega)$. However, in a typical medium where particles can propagate over substantial distances, the imaginary part of the product $\varepsilon(\omega)\mu(\omega)$ does not vanish only in very limited frequency intervals, much more narrow than the intervals that we are discussing now – please have one more look at Fig. 7.5.

[49] This radiation was observed experimentally by Pavel Alekseevich Cherenkov (in older Western texts, "Čerenkov") in 1934, with the observations explained by Ilya Mikhailovich Frank and Igor Yevgenyevich Tamm in 1937. Note, however, that the effect had been predicted theoretically as early as 1889 by the same Oliver Heaviside whose name was mentioned in this course so many times – and whose genius I believe is still underappreciated.

$$\tan\theta = -\frac{E_x}{E_y} = \frac{i\kappa u}{\omega} = \left[\varepsilon(\omega)\mu(\omega)u^2 - 1\right]^{1/2} \equiv \left[\frac{u^2}{v^2(\omega)} - 1\right]^{1/2}, \tag{10.131a}$$

so

$$\boxed{\cos\theta = \frac{v(\omega)}{u} < 1.} \tag{10.131b}$$

Cherenkov radiation: angle

Remarkably, this direction does not depend on the emission time t_{ret}, so the radiation of frequency ω, at each instant, forms a hollow cone led by the particle. This simple result allows an evident interpretation (Fig. 14b): the cone's interior is just the set of all observation points that have already been reached by the radiation, propagating with the speed $v(\omega) < u$, emitted from all previous points of the particle's trajectory by the given time t. This phenomenon is an analog of the so-called *Mach cone* in fluid dynamics,[50] besides that in the Cherenkov radiation, there is a separate cone for each frequency (of the range in which $v(\omega) < u$): the smaller is the $\varepsilon(\omega)\mu(\omega)$ product, i.e. the higher is the wave velocity $v(\omega) = 1/[\varepsilon(\omega)\mu(\omega)]^{1/2}$, the broader is the cone, so the earlier the corresponding "shock wave" arrives to an observer. Please note that the Cherenkov radiation is a unique radiative phenomenon: it takes place even if a particle moves without acceleration, and (in agreement with our analysis in Sec. 2), is impossible in free space, where $v(\omega) = c = \text{const}$ is larger than u for any particle.

The Cherenkov radiation's intensity may be also readily found by plugging the asymptotic expression (127), with imaginary κ, into Eq. (123). The result is

$$\boxed{-\frac{d\mathcal{E}}{dx} \approx \left(\frac{\mathcal{Z}e}{4\pi}\right)^2 \int_{v(\omega)<u} \omega\left[1 - \frac{v^2(\omega)}{u^2}\right] d\omega.} \tag{10.132}$$

Cherenkov radiation: intensity

For non-relativistic particles ($u << c$), the Cherenkov radiation condition $u > v(\omega)$ is fulfilled only in relatively narrow frequency intervals where the product $\varepsilon(\omega)\mu(\omega)$ is very large (usually, due to optical resonance peaks of the electric permittivity – see Fig. 7.5 and its discussion). In this case, the emitted light consists of a few nearly-monochromatic components. On the contrary, if the condition $u > v(\omega)$, i.e. $u^2/\varepsilon(\omega)\mu(\omega) > 1$ is fulfilled in a broad frequency range, as it is for ultra-relativistic particles in condensed media, then the radiated power, according to Eq. (132), is dominated by higher frequencies of the range – hence the famous bluish color of the Cherenkov radiation glow from water-filled nuclear reactors– see Fig. 15.

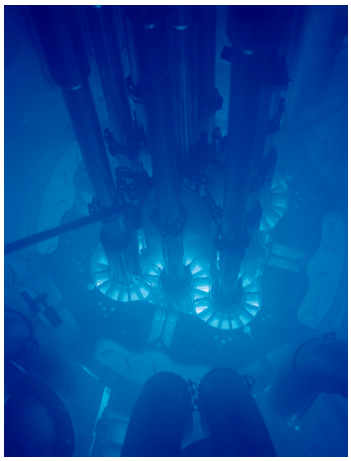


Fig. 10.15. The Cherenkov radiation glow in the Advanced Test Reactor of the Idaho National Laboratory in Arco, ID. (Adapted from http://en.wikipedia.org/wiki/Cherenkov_radiation under the Creative Commons CC-BY-SA-2.0 license.)

[50] Its brief discussion may be found in CM Sec. 8.6.

The Cherenkov radiation is broadly used in high-energy experiments for particle identification and speed measurement (since it is easy to pass the particles through layers of different densities and hence with different dielectric constants) – for example, in the so-called Ring Imaging Cherenkov (RICH) detectors that have been designed for the DELPHI experiment[51] at the Large Electron-Positron Collider (LEP) in CERN.

A little bit counter-intuitively, the formalism described in this section is also very useful for the description of an apparently rather different effect – the so-called *transition radiation* that takes place when a charged particle crosses a border between two media.[52] The effect may be interpreted as the result of the time dependence of the electric dipole formed by the moving charge q and its mirror image q' in the counterpart medium – see Fig. 16.

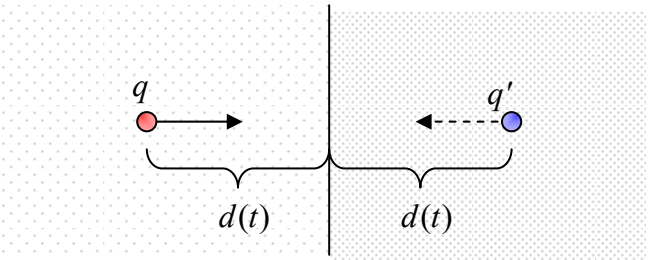


Fig. 10.16. The transition radiation's physics.

In the non-relativistic limit, this effect allows a straightforward description combining the electrostatics picture of Sec. 3.4 (see Fig. 3.9 and its discussion), and Eq. (8.27), corrected for the media polarization effects. However, if the particle's velocity u is comparable with the phase velocity of waves in either medium, the adequate theory of the transition radiation becomes very close to that of the Cherenkov radiation.

In comparison with the Cherenkov radiation, the transition radiation is rather weak, and its practical use (mostly for the measurement of the Lorentz factor γ, to which the radiation intensity is nearly proportional) requires multi-layered stacks.[53] In these systems, the radiation emitted at sequential borders may be coherent, and the system's physics may become close to that of the free-electron lasers mentioned in Sec. 4.

10.6. Radiation's back-action

An attentive and critically-minded reader could notice that so far our treatment of charged particle dynamics has never been fully self-consistent. Indeed, in Sec. 9.6 we have analyzed particle's motion in various external fields, ignoring those radiated by the particle itself, while in Sec. 8.2 and earlier in this chapter these fields have been calculated (admittedly, just for a few simple cases), but, again, their back-action on the emitting particle has been ignored. Only in very few cases we have taken

[51] See, e.g., http://delphiwww.cern.ch/offline/physics/delphi-detector.html. For an in-depth review of radiation detectors (including the Cherenkov ones), the reader may be referred, for example, to the classical text by G. F. Knoll, *Radiation Detection and Measurement*, 4th ed., Wiley, 2010, and a newer treatment by K. Kleinknecht, *Detectors for Particle Radiation*, Cambridge U. Press, 1999.

[52] The effect was predicted theoretically in 1946 by V. Ginzburg and I. Frank, and only later observed experimentally.

[53] See, e.g., Sec. 5.3 in K. Kleinknecht's monograph cited above.

the back effects of the radiation implicitly, via the energy conservation arguments. However, even in these cases, the near-field effects, such as the first term in Eq. (19), which affect the moving particle most, have been ignored.

At the same time, it is clear that in sharp contrast with electrostatics, the interaction of a moving point charge with its own field cannot be always ignored. As the simplest example, if an electron is made to fly through a resonant cavity, thus inducing electromagnetic oscillations in it, and then is forced (say, by an appropriate static field) to return into the cavity before the oscillations have decayed, its motion will certainly be affected by the oscillating fields, just as if they had been induced by another source. There is no conceptual problem with applying the Maxwell theory to such "field-particle rendezvous" effects; moreover, it is the basis of the engineering design of such vacuum electron devices as klystrons, magnetrons, and free-electron lasers.

A problem arises only when no clear "rendezvous" points are enforced by boundary conditions, so the most important self-field effects are at $R \equiv |\mathbf{r} - \mathbf{r}'| \to 0$, the most evident example being the charged particle's radiation into free space, described earlier in this chapter. We already know that such radiation takes away a part of the charge's kinetic energy, i.e. has to cause its deceleration. One should wonder, however, whether such self-action effects might be described in a more direct, non-perturbative way.

As the first attempt, let us try a phenomenological approach based on the already derived formulas for the radiation power $\mathscr{P}$. For the sake of simplicity, let us consider a non-relativistic point charge q in free space, so $\mathscr{P}$ is described by Eq. (8.27), with the electric dipole moment's derivative over time equal to $q\mathbf{u}$:

$$\mathscr{P} = \frac{Z_0 q^2}{6\pi c^2}\dot{u}^2 \equiv \frac{2}{3c^3}\frac{q^2}{4\pi\varepsilon_0}\dot{u}^2 . \tag{10.133}$$

The most naïve approach would be to write the equation of the particle's motion in the form

$$m\dot{\mathbf{u}} = \mathbf{F}_{\text{ext}} + \mathbf{F}_{\text{self}} , \tag{10.134}$$

and try to calculate the radiation back-action force $\mathbf{F}_{\text{self}}$ by requiring its instant power, $-\mathbf{F}_{\text{self}}\cdot\mathbf{u}$, to be equal to $\mathscr{P}$. However, this approach (say, for a 1D motion) would give a very unnatural result,

$$F_{\text{self}} \propto \frac{\dot{u}^2}{u} , \tag{10.135}$$

that might diverge at some points of the particle's trajectory. This failure is clearly due to the retardation effect: as the reader may recall, Eq. (133) results from the analysis of radiation fields in the far-field zone, i.e. at *large* distances R from the particle, e.g., from the second term in Eq. (19), i.e. when the non-radiative first term (which is much larger at *small* distances, $R \to 0$) is ignored.

Before exploring the effects of this term, let us, however, make one more attempt at Eq. (133), considering its *average* effect on some periodic motion of the particle. (A possible argument for this step is that at the periodic motion, the retardation effects should be averaged out – just as at the transfer from Eq. (8.27) to Eq. (8.28).) To calculate the average, let us write the identity

$$\overline{\dot{u}^2} \equiv \frac{1}{\mathscr{T}}\int_0^{\mathscr{T}} \dot{\mathbf{u}}\cdot\dot{\mathbf{u}}\,dt , \tag{10.136}$$

and carry out the integration on the right-hand side of Eq. (133) by parts over the motion period $\mathcal{T}$:

$$\overline{\mathscr{P}} = \frac{2}{3c^3}\frac{q^2}{4\pi\varepsilon_0}\overline{(\dot{\mathbf{u}})^2} = \frac{2}{3c^3}\frac{q^2}{4\pi\varepsilon_0}\frac{1}{\mathcal{T}}\left(\dot{\mathbf{u}}\cdot\mathbf{u}\Big|_0^{\mathcal{T}} - \int_0^{\mathcal{T}}\ddot{\mathbf{u}}\cdot\mathbf{u}\,dt\right) = -\frac{1}{\mathcal{T}}\int_0^{\mathcal{T}}\frac{2}{3c^3}\frac{q^2}{4\pi\varepsilon_0}\ddot{\mathbf{u}}\cdot\mathbf{u}dt\,. \quad (10.137)$$

On the other hand, the back-action force should give

$$\overline{\mathscr{P}} = -\frac{1}{\mathcal{T}}\int_0^{\mathcal{T}}\mathbf{F}_{\text{self}}\cdot\mathbf{u}\,dt\,. \quad (10.138)$$

These two averages coincide if[54]

Abraham-Lorentz force

$$\boxed{\mathbf{F}_{\text{self}} = \frac{2}{3c^3}\frac{q^2}{4\pi\varepsilon_0}\ddot{\mathbf{u}}\,.} \quad (10.139)$$

This is the so-called *Abraham-Lorentz force* of back-action. Before going after a more serious derivation of this formula, let us estimate its scale, representing Eq. (139) as

$$\mathbf{F}_{\text{self}} = m\tau\ddot{\mathbf{u}}, \qquad \text{with } \tau \equiv \frac{2}{3mc^3}\frac{q^2}{4\pi\varepsilon_0}, \quad (10.140)$$

where the constant τ evidently has the dimension of time. Recalling the definition (8.41) of the classical radius r_c of the particle, Eq. (140) for τ may be rewritten as

$$\tau = \frac{2}{3}\frac{r_c}{c}\,. \quad (10.141)$$

For the electron, τ is of the order of 10^{-23} s, so the right-hand side of Eq. (140) is very small. This means that in most cases the Abrahams-Lorentz force is either negligible or leads to the same results as the perturbative treatments of energy loss we have used earlier in this chapter.

However, Eq. (140) brings some unpleasant surprises. For example, let us consider a 1D oscillator with frequency ω_0. For it, Eq. (134), with the back-action force given by Eq. (140), takes the form

$$m\ddot{x} + m\omega_0^2 x = m\tau\dddot{x}\,. \quad (10.142)$$

Looking for the solution of this linear differential equation in the usual exponential form, $x(t) \propto \exp\{\lambda t\}$, we get the following characteristic equation,

$$\lambda^2 + \omega_0^2 = \tau\lambda^3\,. \quad (10.143)$$

It may look like that for any "reasonable" value of $\omega_0 << 1/\tau \sim 10^{23}$ s^{-1}, the right-hand side of this nonlinear algebraic equation may be treated as a perturbation. Indeed, looking for its solutions in the

[54] Just for the reader's reference, this formula may be readily generalized to the relativistic case, in the 4-form:

$$F^{\alpha}_{\text{self}} = \frac{2}{3mc^3}\frac{q^2}{4\pi\varepsilon_0}\left[\frac{d^2p^{\alpha}}{d\tau^2} + \frac{p^{\alpha}}{(mc)^2}\left(\frac{dp_{\beta}}{d\tau}\frac{dp^{\beta}}{d\tau}\right)\right],$$

– the so-called *Abraham-Lorentz-Dirac force*.

natural form $\lambda_{\pm} = \pm i\omega_0 + \lambda'$, with $|\lambda'| << \omega_0$, expanding both parts of Eq. (143) in the Taylor series in the small parameter λ', and keeping only the terms linear in λ', we get

$$\lambda' \approx -\frac{\omega_0^2 \tau}{2}. \tag{10.144}$$

This means that the energy of free oscillations decreases in time as $\exp\{2\lambda' t\} = \exp\{-\omega_0^2 \tau\, t\}$; this is exactly the radiative damping analyzed earlier. However, Eq. (143) is deceiving; it has the third root corresponding to unphysical, exponentially growing (so-called *run-away*) solutions. It is easiest to see this for a free particle, with $\omega_0 = 0$. Then Eq. (143) becomes very simple,

$$\lambda^2 = \tau\lambda^3, \tag{10.145}$$

and it is easy to find all its three roots explicitly: $\lambda_1 = \lambda_2 = 0$ and $\lambda_3 = 1/\tau$. While the first two roots correspond to the values $\lambda_{\pm}$ found earlier, the last one describes an exponential (and extremely rapid!) acceleration.

In order to remove this artifact, let us try to develop a self-consistent approach to the back-action effects, taking into account the near-field terms of particle fields. For that, we need to somehow overcome the divergence of Eqs. (10) and (19) at $R \to 0$. The most reasonable way to do this is to spread the particle's charge over a ball of radius a, with a spherically symmetric (but not necessarily constant) density $\rho(r)$, and at the end of the calculations trace the limit $a \to 0$.[55] Again sticking to the non-relativistic case (so the magnetic component of the Lorentz force is not important), we should calculate

$$\mathbf{F}_{\text{self}} = \int_V \rho(\mathbf{r})\mathbf{E}(\mathbf{r},t)d^3r, \tag{10.146}$$

where the electric field is that of the charge itself, with the field of any elementary charge $dq = \rho(r)d^3r$ described by Eq. (19).

To enable an analytical calculation of the force, we need to make the assumption $a << r_c$, treat the ratio $R/r_c \sim a/r_c$ as a small parameter, and expand the resulting right-hand side of Eq. (146) into the Taylor series in small R. This procedure yields

$$\mathbf{F}_{\text{self}} = -\frac{2}{3}\frac{1}{4\pi\varepsilon_0}\sum_{n=0}^{\infty}\frac{(-1)^n}{c^{n+2}n!}\frac{d^{n+1}\mathbf{u}}{dt^{n+1}}\int_V d^3r\int_V d^3r'\rho(r)R^{n-1}\rho(r'). \tag{10.147}$$

The distance R cancels only in the term with $n = 1$,

$$\mathbf{F}_1 = \frac{2}{3c^3}\frac{\ddot{\mathbf{u}}}{4\pi\varepsilon_0}\int_V d^3r\int_V d^3r'\rho(r)\rho(r') \equiv \frac{2}{3c^3}\frac{q^2}{4\pi\varepsilon_0}\ddot{\mathbf{u}}, \tag{10.148}$$

showing that we have recovered (now in an *apparently* legitimate fashion) Eq. (139) for the Abrahams-Lorentz force. One could argue that in the limit $a \to 0$ the terms higher in $R \sim a$ (with $n > 1$) could be ignored. However, we have to notice that the main contribution to the series (147) is *not* described by Eq. (148) for $n = 1$, but is given by the much larger term with $n = 0$:

[55] Note: this operation cannot be interpreted as describing a quantum spread due to the finite extent of the point particle's wavefunction. In quantum mechanics, different parts of the wavefunction of the same charged particle do *not* interact with each other!

$$\mathbf{F}_0 = -\frac{2}{3}\frac{1}{4\pi\varepsilon_0}\frac{\dot{\mathbf{u}}}{c^2}\int_V d^3r\int_V d^3r'\frac{\rho(r)\rho(r')}{R} \equiv -\frac{4}{3}\frac{\dot{\mathbf{u}}}{c^2}\frac{1}{4\pi\varepsilon_0}\frac{1}{2}\int_V d^3r\int_V d^3r'\frac{\rho(r)\rho(r')}{R} \equiv -\frac{4}{3c^2}\dot{\mathbf{u}}U, \quad (10.149)$$

where U is the electrostatic energy (1.59) of the static charge's self–interaction. This term may be interpreted as the inertial "force" [56] ($-m_{\text{ef}}\mathbf{a}$) with the following effective *electromagnetic mass*:

Electro-magnetic mass

$$m_{\text{ef}} = \frac{4}{3}\frac{U}{c^2}, \quad (10.150)$$

which is a factor of 4/3 larger than it should be according to Einstein's formula (9.73). This is the famous (or rather infamous :-) *4/3 problem* that does not allow one to interpret the electron's mass as that of its electric field. Some (admittedly, rather formal) resolution of this paradox is possible only in quantum electrodynamics with its renormalization techniques – beyond the framework of this course.

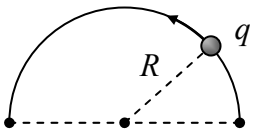
Note that all these issues are only important for motions with frequencies of the order of $1/\tau \sim 10^{23}$ s^{-1}, i.e. at energies as high as $\sim\hbar/\tau \sim 10^8$ eV, while other quantum electrodynamics effects may be observed at much lower frequencies, starting from $\sim 10^{10}$ s^{-1}. Hence the 4/3 problem is by no means the only or the most significant motivation for the transfer from classical to quantum electrodynamics. However, the reader should not think that their time spent on this course has been lost: quantum electrodynamics it heavily based on classical electrodynamics, incorporates virtually all its results, and the basic transition between them is surprisingly straightforward.[57] So, I look forward to welcoming the reader to the next, quantum-mechanics part of this series.

10.6. Exercise problems

10.1. Derive Eqs. (10) from Eqs. (1) by a direct (but careful!) integration.

10.2. Derive the radiation-related parts of Eqs. (19)-(20) from the Liénard-Wiechert potentials (10) by direct differentiation.

10.3. A point charge q that was in a stationary position on a circle of radius R is carried over, along the circle, to the opposite position on the same diameter (see the figure on the right) as fast as only physically possible, and then is kept steady at this new position. Calculate and sketch the time dependence of its electric field **E** at the center of the circle.



10.4. Express the instantaneous power of electromagnetic radiation by a relativistic particle with electric charge q and rest mass m, moving with velocity **u**, via the Lorentz force **F** providing its acceleration.

10.5. A relativistic particle with rest mass m and electric charge q, initially at rest, is accelerated by a constant force **F** until it reaches a certain velocity u and then is left to move by inertia. Calculate the total energy radiated during the acceleration.

[56] See, e.g., CM Sec. 4.6.
[57] See, e.g., QM Secs. 9.1-9.4.

10.6. A charged relativistic particle with an initial momentum $\mathbf{p}_0$ flies ballistically from a free-space region into a region of a constant, uniform electric field $\mathbf{E}$, whose force is directed opposite to $\mathbf{p}_0$. Calculate the energy radiated by the particle during its motion in the field, assuming that it is small in comparison with the particle's initial kinetic energy.

10.7. Calculate

(i) the instantaneous power, and
(ii)* the power spectrum

of the radiation emitted, into a unit solid angle, by a relativistic particle with charge q, performing 1D harmonic oscillations with frequency ω_0 and displacement amplitude a.

10.8. Calculate and analyze the time dependence of the energy of a charged relativistic particle rotating in a constant and uniform magnetic field $\mathbf{B}$ and, as a result, emitting the synchrotron radiation. Qualitatively, what is the particle's trajectory?

Hint: You may assume that the energy loss is relatively slow ($-d\mathscr{E}/dt \ll \omega_c\mathscr{E}$), but should spell out the condition of validity of this assumption.

10.9. Analyze the polarization of the synchrotron radiation propagating within the particle's rotation plane.

10.10. Analyze the polarization and the spectral contents of the synchrotron radiation propagating in the direction normal to the particle's rotation plane. How do the results change if not one, but $N > 1$ similar particles move around the circle, at equal angular distances?

10.11.* The basic quantum theory of radiation shows that the electric dipole radiation by a particle is allowed only if the change of its angular momentum's magnitude L at the transition is of the order of Planck's constant $\hbar$.

(i) Estimate the change of L of an ultra-relativistic particle due to its emission of a typical single photon of the synchrotron radiation.
(ii) Do you think quantum mechanics forbid such radiation? If not, why?

10.12. A relativistic particle moves along the z-axis, with velocity u_z, through an undulator – a system of permanent magnets providing (in the simplest model) a perpendicular magnetic field, whose distribution near the axis is sinusoidal:[58]

$$\mathbf{B} = \mathbf{n}_y B_0 \cos k_0 z .$$

Assuming that the field is so weak that it causes negligible deviations of the particle's trajectory from the straight line, calculate the angular distribution of the resulting radiation. What condition does the above assumption impose on the system's parameters?

[58] As the Maxwell equation for $\nabla\times\mathbf{H}$ shows, this field distribution cannot be created in any non-zero volume of free space. However, it may be created on a line – e.g., on the particle's trajectory.

10.13. Discuss possible effects of the interference of the undulator radiation from different periods of its static field distribution. In particular, calculate the angular positions of the power density maxima.

10.14. An electron launched directly toward a plane surface of a perfect conductor is instantly absorbed by it at the impact. Calculate the angular distribution and the frequency spectrum of the electromagnetic waves radiated at this event, provided that the initial kinetic energy T of the particle is much larger than the conductor's workfunction ψ.[59] Is your result valid near the conductor's surface?

10.15. A relativistic particle, with a rest mass m and an electric charge q, flies ballistically, with velocity u, by an immobile point charge q', with an impact parameter b so large that the deviations of its trajectory from the straight line are negligible. Calculate the total energy loss due to the electromagnetic radiation during the passage. Quantify the conditions of validity of your result.

[59] See Sec. 2.9, in particular Fig. 2.27a.

This page is intentionally left blank