Chapter 8. Macroscopic media: response theory and the physics of $\varepsilon(\omega)$ (05 Nov 2020)

A. Overview:

The equations that we have used so far are microscopic, with each charge treated individually. At room temperature a cubic centimeter of crystalline Si contains 4.99 × 10^{22} atoms, each of which has 4 valence electrons and 10 core electrons. The valence electrons alone represent about 2 × 10^{23} charges per cm^3. There is obviously no way we can deal with such enormous detail, nor would it be useful to do so. To move forward, we follow the path laid out by thermodynamics. Rather than track the motion of each molecule in a finite volume of gas, we define macroscopic quantities such as pressure and temperature, and develop macroscopic equations such as $PV = nRT$ that describe molecular behavior at the macroscopic level.

Fortunately, in E&M the macroscopic equations are virtually identical to their microscopic counterparts. Hence in this case the transition from microscopic to macroscopic is not characterized by major formal differences. The new aspects are the constitutive relations

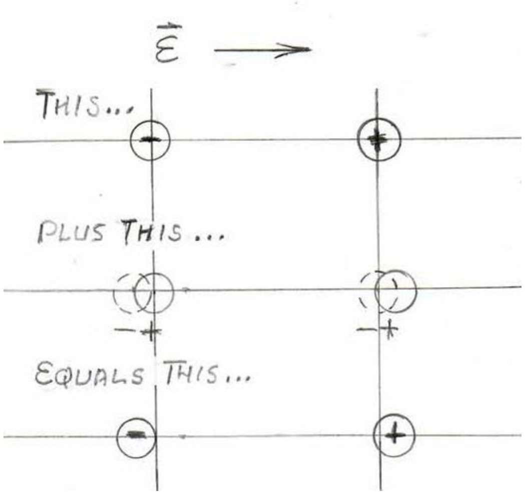
$$\vec{D} = \varepsilon\vec{E} = \vec{E} + 4\pi\vec{P} \; ; \tag{8.1a}$$

$$\vec{B} = \mu\vec{H} = \vec{H} + 4\pi\vec{M} \; ; \tag{8.1b}$$

where ε and μ are the macroscopic dielectric function and permeability, respectively, and $\vec{P}$ and $\vec{M}$ are the volume density of electric and magnetic dipoles, respectively. The $\varepsilon = \varepsilon(\omega)$ and $\mu = \mu(\omega)$ describe the steady-state response of a material to applied fields of harmonic time dependences ω. They can be treated entirely phenomenologically. They can be measured, and the values obtained used to describe functionality, for example energy dissipation, optical reflection, and ferromagnetism, among many other phenomena.

However, recognizing that physics takes place in real space, in real time, and on the atomic scale, we need to establish a connection to the microscopic world to ensure that our macroscopic results are firmly based in physics. For ε, where the dipoles are induced, this can be done with simple mechanical models based on force. We are assisted here because the maximum static fields that can be applied, of the order of 10^6 V/cm, are about 3 orders of magnitude smaller than the intrinsic fields holding materials together. Thus the equivalent of perturbation theory is satisfactory. For μ, the situation is qualitatively different. For ferromagnets, the materials that we consider here, the dipoles are pre-existing and are orders of magnitude larger than the induced dipoles of electrostatics. This calculation must be done by considering energy, a much less straightforward task. However, both cases highlight the importance of dipoles in the multipole expansions discussed in the preceding chapter. We treat the electric case in this chapter, boundary-condition problems involving ε in Ch. 9, and magnetics in Ch. 10.

To begin, consider what happens when an electric field $\vec{E}$ is applied to a hypothetical solid that consists of a lattice of positive and negative point charges $+q$ and $-q$, respectively. The field generates a force $\vec{F} = q\vec{E}$ causes the charges to shift out of their equilibrium positions, with the positive charges moving in the field direction and the negative charges in the opposite direction, as indicated in highly exaggerated form in the diagram. It is seen that in both cases the result can be described as the superposition of the original charge and an induced dipole. For the positive charge the negative end of the dipole cancels the original positive charge, leaving the charge in a new position. For the negative charge the positive end cancels the original negative charge, also leaving it in the new position. It can be appreciated that in both cases the dipole is oriented in the same direction, parallel to $\vec{E}$. Hence without loss of generality we can work exclusively with positive charges.



To set the scale of these shifts, static electric fields that can be applied externally rarely exceed 10^6 V/cm, a value at least 3 orders of magnitude less than the ca. 10^9 V/cm fields holding materials together. Thus the $\Delta\vec{r}$ displacements are effectively infinitesimal, and can be accurately described as perturbations. This greatly simplifies calculations in that we let Nature solve the crystal-structure problem (the heavy lifting), while we focus only on field-induced changes. Nevertheless, despite their seemingly mathematical definition, the effects of these dipoles are quite real, for example giving rise to the screening charge that occurs at the boundaries of dissimilar materials.

In contrast, owing to their quantum origin, magnetic dipoles are typically orders of magnitude larger than induced electric dipoles. This is evidenced by the fact that permeabilities of 10^5 to 10^6 are routine in ferromagnets, while dielectric functions of insulators are typically of the order of 10 or less. Magnetic dipoles typically organize into domains where the individual dipoles are locally parallel, and the domains orient quasi-randomly. Domain formation happens because aligning dipoles parallel to a field minimizes the energy $W = -\vec{m} \cdot \vec{B}$. Quasi-random domain orientation happens as a result of the system attempting to keep as much of the magnetic flux inside the material as possible. It is energetically far more costly to force high magnetic fields to exist outside ferromagnets than to create ~10 nm walls between domains to keep the fields in. When an external field intensity $\vec{H}$ is applied, net orientation is realized either by dipoles falling into line through domain-wall migration, or by the nearly instant reorientation of an entire domain. The latter give rise to "Barkhausen" steps, electrical impulses that are generated in a coil placed around the magnet. As can be appreciated, these processes are anything but perturbative, so the treatment of magnetics is more difficult.

Relaxation processes are likewise significantly different. When an electric field is removed, the material relaxes on a time scale of the order of the reciprocal of phonon frequencies, or about 10^{-13} to 10^{-14} s. In contrast, magnetic relaxation occurs on time scales of the order of 10^{10} times longer. It is clear that we will not expect a magnetic version of integrated-circuits technology any time soon.

While emphasis in the following is on developing $\varepsilon = \varepsilon(\omega)$ as the response of particular materials to a steady-state harmonic field $\vec{E}(t) = \vec{E}_o e^{-i\omega t}$, we also cover response theory to highlight fundamental constraints on ε (and μ) imposed by reality, causality, and linearity. These include the relation $\varepsilon(\omega) = \varepsilon^*(-\omega)$, that all poles of $\varepsilon(\omega)$ lie in the lower half of the complex ω plane, that the real and imaginary parts of $\varepsilon(\omega)$ obey the Kramers-Kronig relations, that $\lim_{\omega\to\infty}(\varepsilon(\omega)) = 1 - \omega_p^2 / \omega^2$, where ω_p is the plasma frequency, that plasmonics is nothing more than the solution of $\varepsilon(\omega) = 0$, i.e., the appearance of a dipole density $\vec{P}$ in the absence of an applied field $\vec{E}$, and that the Fourier transform of $(\varepsilon(\omega) - 1)$ is the Green function $G(t,t')$, where

$$\vec{P}(t) = \int_{-\infty}^{\infty} dt' G(t,t')\vec{E}(t'). \tag{8.2}$$

Logical contradictions that occur when the steady-state assumption is ignored are also discussed. Examples are given, with the objective of illustrating the underlying points that are discussed. Although these are covered from the perspective of E&M, the results are far more general, being applicable to any causal system.

Section B of Ch. 8 deals with spatial averaging, which defines the macroscopic quantities that constitute Maxwell's Equations on the macroscopic level. Given these definitions, in Sec. C we develop the means of expressing the macroscopic dielectric function in terms of the microscopic properties of the relevant charges. Section D then takes these equations and generates expressions for ε where the inertial, dissipative, and restoring forces dominate. Section E provides examples of ε for materials of increasing complexity: metals, metals with energy loss, dielectrics, and semiconductors. The remainder of the chapter deals with general properties of ε, including plasmonics, response theory in the frequency domain, Kramers-Kronig relations, sum rules, nonlinear materials (nonlinear optics), and anisotropic materials (crystal optics). Chapter 8 concludes with a discussion of local-field effects, as encoded in the Clausius-Mossotti equation, and the connection between $\varepsilon(\omega)$ and the Green function of Eq. (8.2). Local-field effects are a class of phenomena that is generally misinterpreted in standard textbook treatments (Jackson is a partial exception.)

Upon finishing Ch. 8, you should be convinced that dielectric functions are not simply phenomenological parameters with no physical meaning, but are deterministic, based on atomic-scale properties, contain information about them, and are macroscopic quantities that you can derive under quite general circumstances, more so than discussed here.

B. Spatial averaging: the micro-to-macro transition.

The perturbation treatment connects the macroscopic properties of a material (its dielectric function ε) to the atomic-scale properties of its constituent charges, electrons and atoms (the microscopic perspective). The key to this connection is spatial averaging. Averaging is easily justified. Unless the measurement scale (wavelength) of a laboratory probe is of the order of atomic spacings, for example X-ray scattering or transmission electron microscopy, most laboratory measurements intrinsically perform this average over atomic-scale properties. The length scales of the fields and potentials with which we generally work, at least for frequencies from the static limit to the visible-near ultraviolet spectral range, are so much larger than the atomic scale that single atoms or even nanostructures cannot be resolved.

An average is a convolution of a weighting or averaging function with the microscopic quantity of interest. Representing the averaging function as $W(\vec{r}-\vec{r}\,')$, and using the microscopic electric field $\vec{E}(\vec{r})$ as an example, the macroscopic equivalent $<\vec{E}(\vec{r})>$ of $\vec{E}(\vec{r})$ is

$$<\vec{E}(\vec{r})> = \int d^3r' W(\vec{r}-\vec{r}\,')\vec{E}(\vec{r}\,') . \tag{8.3}$$

The weighting function $W(\vec{r}-\vec{r}\,')$ $W(\vec{r}-\vec{r}\,')$ need be defined only in very general terms:

(1) it is positive definite;

(2) it varies slowly on the atomic scale but “fast enough” on the laboratory scale;

(3) it is a function of $(\vec{r}-\vec{r}\,')$, where $\vec{r}$ is the location of the observer and $\vec{r}\,'$ traces the atomic-scale values of the quantity being averaged;

(4) it is differentiable; and

(5) for arbitrarily large volumes it integrates to 1:

$$\int d^3r' W(\vec{r}-\vec{r}\,') = 1 . \tag{8.4}$$

“Fast enough” means that $W(\vec{r}-\vec{r}\,')$ is not so broad that laboratory-scale details are washed out.

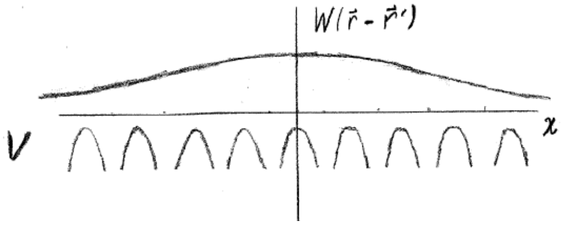
One-dimensional examples of $W(\vec{r}-\vec{r}\,')$ are

$$W_G(\vec{r}-\vec{r}\,') = \frac{1}{\Delta x\sqrt{\pi}} e^{-(x-x')^2/\Delta x^2} ; \tag{8.5a}$$

$$W_L(\vec{r}-\vec{r}\,') = \frac{\Delta x}{\pi}\frac{1}{(x-x')^2+\Delta x^2}\,; \tag{8.5b}$$

where in Eq. (8.5b) Δx is the half-width of the structure at half-maximum. These averaging functions have dimensions of 1/(length). Since $W(\vec{r}-\vec{r}\,')$ must integrate to unity, the two- and three-dimensional versions have dimensions of 1/(length)^2 and 1/(length)^3, respectively.

A qualitative illustration of the meaning of $W(\vec{r}-\vec{r}\,')$ is shown in the diagram. This illustrates schematically how it varies in space relative to atomic-scale potentials. In practice, many more lattice sites are involved instead of the relatively small number indicated here.



A reciprocal-space interpretation can also be given. The Fourier transform of a convolution of two functions in direct space, as in Eq. (7.8), is the product of the Fourier transforms of the individual functions in the reciprocal space. The spatial Fourier coefficients of a microscopic field $\vec{E}(\vec{r}\,')$ encode variations on an atomic scale. As can be appreciated from the diagram, such coefficients are expected to be large at high orders. In contrast, those of $W(\vec{r}-\vec{r}\,')$ will have a narrow distribution about zero. Hence in forming the product, the high-order coefficients of $\vec{E}(\vec{r})$ are greatly reduced or even eliminated, resulting in a smoothed output. The detailed atomic structure is obviously sacrificed in the process.

We next show that an operation such as ∇, $\nabla\cdot$, and $\nabla\times$ acting on a macroscopic average of a function is equal to the average of the operation acting on the function itself. Given that averaging is a linear operation, this is not surprising. The result follows because $W(\vec{r}-\vec{r}\,')$ depends on the difference $(\vec{r}-\vec{r}\,')$ of the coordinates and its volume integral is 1.

All such calculations follow the same pattern. Using $\vec{E} = -\nabla\phi$ as an example, consider

$$\nabla_{\vec{r}} < \phi(\vec{r}) > = \nabla_{\vec{r}}\int d^3r' W(\vec{r}-\vec{r}\,')\phi(\vec{r}\,') = \int d^3r'\{\nabla_{\vec{r}}[W(\vec{r}-\vec{r}\,')]\}\phi(\vec{r}\,')$$
$$= -\int d^3r'\{\nabla_{\vec{r}\,'}[W(\vec{r}-\vec{r}\,')]\}\phi(\vec{r}\,')\,. \tag{8.6}$$

The last step follows because W is a function of $(\vec{r}-\vec{r}\,')$. Next, do an independent calculation, where the operator works on the entire integrand:

$$\int_V d^3r'\nabla_{\vec{r}\,'}[W(\vec{r}-\vec{r}\,')\phi(\vec{r}\,')]$$
$$= \int_V d^3r'\big(\phi(\vec{r}\,')\,\nabla_{\vec{r}\,'}(W(\vec{r}-\vec{r}\,')) + W(\vec{r}-\vec{r}\,')\nabla_{\vec{r}\,'}\phi(\vec{r}\,')\big) \tag{8.7a}$$

$$= \int_S d^2r'\,\hat{n}W(\vec{r}-\vec{r}\,')\phi(\vec{r}\,') = 0\,. \tag{8.7b}$$

Equation (8.7b) follows because the integral of the gradient of a function over a volume *V* is equal to the integral of the same function over the surface *S* of *V*, with ∇ replaced by the normal vector $\hat{n}$ of the surface. Because the range of *W* is finite, a surface can always be found such that *W* vanishes on it. Finally, make the replacement

$$\int_V d^3r'\{\nabla_{\vec{r}\,'}[W(\vec{r}-\vec{r}\,')]\}\phi(\vec{r}\,') = -\int_V d^3r'W(\vec{r}-\vec{r}\,')\{\nabla_{\vec{r}\,'}\phi(\vec{r}\,')\}\,, \tag{8.8}$$

so

$$-\nabla<\phi> = -\int_V d^3r'W(\vec{r}-\vec{r}\,')\nabla_{\vec{r}\,'}\phi(\vec{r}\,') = \int_V d^3r'W(\vec{r}-\vec{r}\,')\vec{E}(\vec{r}\,') = <\vec{E}>\,, \tag{8.9}$$

which is the desired result.

A particularly important relation that can be proved similarly is the differential version of Coulomb's Law:

$$\nabla\cdot<\vec{E}> = 4\pi<\rho>\,. \tag{8.10}$$

In the next section, this will be used to calculate $\mathcal{E}$. Start with

$$\nabla_{\vec{r}}\cdot<\vec{E}(\vec{r})> = \nabla_{\vec{r}}\cdot\int_V d^3r'W(\vec{r}-\vec{r}\,')\vec{E}(\vec{r}\,') \tag{8.11a}$$

$$= \int_V d^3r'\{\nabla_{\vec{r}}W(\vec{r}-\vec{r}\,')\}\cdot\vec{E}(\vec{r}\,') + \int_V d^3r'W(\vec{r}-\vec{r}\,')\nabla_{\vec{r}}\cdot\vec{E}(\vec{r}\,')\,. \tag{8.11b}$$

The second integral in Eq. (8.11b) vanishes because $\vec{E}(\vec{r}\,')$ is a function of $\vec{r}\,'$, not $\vec{r}$. Now change the target of the gradient operating on $W(\vec{r}-\vec{r}\,')$ from $\vec{r}$ to $\vec{r}\,'$, which introduces a minus sign.

Next, start a new line:

$$\int_V d^3r'\nabla_{\vec{r}\,'}\cdot\{W(\vec{r}-\vec{r}\,')\vec{E}(\vec{r}\,')\} = \int_S d^2r'\hat{n}\cdot\left(W(\vec{r}-\vec{r}\,')\vec{E}(\vec{r}\,')\right) = 0 \tag{8.12a}$$

$$= \int_V d^3r'[\nabla_{\vec{r}\,'}W(\vec{r}-\vec{r}\,')\}\cdot\vec{E}(\vec{r}\,')\} + \int_V d^3r'W(\vec{r}-\vec{r}\,')\{\nabla_{\vec{r}\,'}\cdot\vec{E}(\vec{r}\,')\}\,. \tag{8.12b}$$

The top line vanishes because we can always make S large enough so $W(\vec{r}-\vec{r}\,')$ vanishes everywhere on it. Putting everything together, the result is

$$\nabla_{\vec{r}}\cdot<\vec{E}(\vec{r})> = \int_V d^3r'W(\vec{r}-\vec{r}\,')\{\nabla_{\vec{r}\,'}\cdot\vec{E}(\vec{r}\,')\}$$

$$= 4\pi\int_V d^3r'W(\vec{r}-\vec{r}\,')\rho(\vec{r}\,') = 4\pi<\rho>\,, \tag{8.13}$$

which was to be shown.

The concept of averaging obviously breaks down at boundaries between different materials, so at interfaces we assume that fields are already bulk-averaged macroscopic values. The averaged fields are connected across the interface using standard boundary conditions.

C. The atomic-scale basis of $\varepsilon(\omega)$.

We now place ε on a deterministic foundation at the atomic level. The treatment is general, and with minor extensions discussed below describes nonlinear optics and the tensorial dielectric response of anisotropic materials, among other phenomena. These are left for homework assignments. A byproduct is the definition of the macroscopic displacement field $\vec{D} = < \vec{D} >$.

The calculation proceeds as follows. Suppose that the material is in static equilibrium before $\vec{E}$ is applied. The macroscopic version of Coulomb's Law is

$$\nabla_{\vec{r}} \cdot < \vec{E}(\vec{r}) > = 4\pi < \rho_o(\vec{r}) >, \tag{8.14}$$

where $< \rho_o(\vec{r}) >$ is the macroscopic average charge density of the *unperturbed* material. For the usual case of neutral materials, this is zero. From now on we drop the arguments $(\vec{r})$ and brackets $< >$ for notational simplicity except where necessary. After $\vec{E}$ is applied, Eq. (8.14) becomes

$$\nabla \cdot \vec{E} = 4\pi\rho, \tag{8.15}$$

Where ρ includes the dipoles induced by $\vec{E}$.

While Eq. (8.15) is correct, it is not particularly informative. However, we can fix this by breaking ρ into two parts:

$$\rho = \rho_o + (\rho - \rho_o), \tag{8.16}$$

The first term describes the unperturbed material, and the part in parentheses the response of the material to the field. With this substitution the divergence expression now reads

$$\nabla \cdot < \vec{E} > = 4\pi < \rho_o > + 4\pi \int d^3r' W(\vec{r} - \vec{r}\,')[\rho(\vec{r}\,') - \rho_o(\vec{r}\,')]. \tag{8.17}$$

To proceed further we need a model. Let the material consist of a set of point charges q_j (for example, electrons) at locations $\vec{r}_j$, and let $\vec{E}$ move each charge to a new position $\vec{r}_j + \Delta\vec{r}_j$, where as noted above the $\Delta\vec{r}_j$ are small. For mathematical simplicity let all $q_j = q$ and $\Delta\vec{r}_j = \Delta\vec{r}$. Although real materials contain charges of different types with necessarily different responses, the equations are linear, hence these can be added as needed. This is the big advantage of the dielectric function: different classes of charges typically contribute approximately independently, so if another mechanism is important, charges can be added as appropriate. If the charges interact the calculations are similar but terms that connect the different charges must be added. This case gives rise to more exotic phenomena such as electromagnetically induced transparency, which will be discussed next semester.

For now, assume that each charge is independent, and write

$$\rho_o(\vec r\,') = \sum_j q\delta(\vec r\,' - \vec r_j)\,, \tag{8.18a}$$

$$\rho(\vec r\,') = \sum_j q\delta(\vec r\,' - \vec r_j - \Delta\vec r)\,. \tag{8.18b}$$

With the charge densities given by Eqs. (8.18), the integration in Eq. (8.17) is trivial. The result is

$$\begin{aligned}
&4\pi\int d^3r' W(\vec r - \vec r\,')[\rho(\vec r\,') - \rho_o(\vec r\,')] \\
&\qquad = 4\pi\int d^3r' W(\vec r - \vec r\,')[\sum_j q\delta(\vec r\,' - \vec r_j - \Delta\vec r) - \sum_j q\delta(\vec r\,' - \vec r_j)] && (8.19a) \\
&\qquad = 4\pi\, q\sum_j [W(\vec r - \vec r_j - \Delta\vec r) - W(\vec r - \vec r_j)]\,. && (8.19b)
\end{aligned}$$

Now W is a slowly varying function of its argument. Hence we can expand $W(\vec r - \vec r\,' - \Delta\vec r\,')$ in a 3-dimensional Taylor series to first order in $\Delta\vec r$:

$$W(\vec r - \vec r\,' - \Delta\vec r) \approx W(\vec r - \vec r\,') - (\Delta\vec r\cdot\nabla_{\vec r})W(\vec r - \vec r\,')\,. \tag{8.20}$$

We choose ∇ to target $\vec r$ rather than $\vec r\,'$ to simplify what follows. With this substitution, the zero-order terms in Eq. (8.19b) cancel.

Next, convert the sum to an integral with the replacement $\sum_j \to \int_V d^3r' n(\vec r\,')$, where $n(\vec r\,')$ is the local density of states. This is the same transformation that converts Coulomb's Law for discrete charges to its continuum equivalent. The result is

$$4\pi\int_V d^3r' W(\vec r - \vec r\,')\left(\rho - \rho_o\right) = -4\pi\int_V d^3r' n(\vec r\,')\Delta\vec r(\vec r\,')\cdot\nabla_{\vec r}W(\vec r - \vec r\,') \tag{8.21}$$

where we note explicitly that n and $\Delta\vec r$ can be functions of $\vec r\,'$. Now consider

$$\nabla_{\vec r}\cdot\left(n(\vec r\,')\Delta\vec r(\vec r\,')W(\vec r - \vec r\,')\right) = \Delta\vec r\cdot\nabla_{\vec r}\left(nW\right) + nW\nabla_{\vec r}\cdot\Delta\vec r = n\Delta\vec r\cdot\nabla_{\vec r}W\,, \tag{8.22}$$

The other terms vanish because these targets of the gradient and divergence operation are functions of $\vec r\,'$ not $\vec r$. We are therefore left with

$$4\pi\int_V d^3r' W(\vec r - \vec r\,')\left(\rho - \rho_o\right) = -4\pi\nabla\cdot\vec P\,, \tag{8.23}$$

where

$$\vec P = \int_V d^3r'\, nq\Delta\vec r\, W(\vec r - \vec r\,') = \int_V d^3r'\, n\,\vec p\, W(\vec r - \vec r\,') \tag{8.24a,b}$$

is the macroscopic *dipole density*, that is, the macroscopic average of the number density *n* of the mathematical dipoles $\vec{p} = q\Delta\vec{r}$. If *n* and $\vec{p}$ are independent of $\vec{r}\,'$, which we assume in the following, then we can bring them out of the integral. Because the integral of $W(\vec{r} - \vec{r}\,')$ over all space is 1, the result is

$$\vec{P} = n\,\vec{p} = n\,q\,\Delta\vec{r}\,. \tag{8.25}$$

The macroscopic differential version of Coulomb's Law has therefore been reduced to

$$\nabla\cdot\vec{E} = 4\pi\rho_o - 4\pi\nabla\cdot\vec{P}\,, \tag{8.26a}$$

$$= 4\pi(\rho_o + \rho_P)\,, \tag{8.26b}$$

where $\rho_P = -\nabla\cdot\vec{P}$ is the *polarization charge density*. Equations (8.26) show that ρ_P is fully equivalent to ρ_o. Combining the two divergences in Eq. (8.26a) yields

$$\nabla\cdot(\vec{E} + 4\pi\vec{P}) = \nabla\cdot\vec{D}\,, \tag{8.27}$$

thereby defining $\vec{D}$, is the macroscopic *displacement field*. From Eq. (8.27), $\vec{D}$ satisfies

$$\nabla\cdot\vec{D} = 4\pi\rho_o\,, \tag{8.28}$$

which depends only on the (average) charge density of the material before the field is applied, and does not "see" the polarization charge. If $\rho_o = 0$, Gauss' Theorem leads in the usual way to the conclusion that the normal component of $\vec{D}$ is continuous at an interface between two different materials.

Next, suppose that the response is linear in $\vec{E}$, and write $\vec{P} = \chi_E\vec{E}$. This defines the electric susceptibility χ_E. The dielectric function ε follows as

$$\vec{D} = \vec{E} + 4\pi\,\vec{P} \tag{8.29a}$$

$$= (1 + 4\pi\chi_E)\,\vec{E} \tag{8.29b}$$

$$= \varepsilon\,\vec{E}\,, \tag{8.29c}$$

If *n* in Eqs. (8.24) is constant, and *q* and $\Delta\vec{r}$ assumed to be the same for all charges, then by Eq. (8.29c)

$$\varepsilon\,\vec{E} = \vec{E} + 4\pi\,n\,q\Delta\vec{r}. \tag{8.30}$$

For sufficiently small fields $\Delta\vec{r}$ is expected to be proportional to $\vec{E}$. In this case $\vec{E}$ can be eliminated as a common factor, leaving an expression that is field-independent. The dielectric function therefore provides a macroscopic summary of the atomic-scale response of a medium to an applied field.

Up to now we have assumed electrostatics. But Eq. (8.30) shows that cancellation also occurs if $\vec{E}$ and $\Delta\vec{r}$ share the *same* time dependence $e^{-i\omega t}$. Thus the above development allows time to be introduced as a new coordinate *if* calculations are restricted to the steady state (no transients). More general time dependences require $\vec{P}$ to be calculated using Green functions, as discussed in Sec. G. Also, with fields and responses having time dependences $e^{-i\omega t}$, ε becomes a function of frequency, or $\varepsilon = \varepsilon(\omega)$. Both results are huge steps forward, because we have at the same time developed the machinery to treat frequency-dependent phenomena, for example the optical properties of materials.

Finally, we have reduced the calculation of the response of a material to an electric field to a famous equation from another branch of physics, specifically $\vec{F} = m\vec{a}$. In the next section we model $\Delta\vec{r}$ in terms of atomic-scale parameters such as carrier mass, dissipation, and a Hooke's-Law restoring force. Elaborations include anharmonic and anisotropic restoring forces, which are the keys to nonlinear and crystal optics, respectively. For the moment these are left as homework assignments.

Because the results are more profound than might be apparent at first sight, it is worth summarizing what happened. We have

(1) Derived the differential version of Coulomb's Law for a macroscopic material;

(2) Obtained a recipe for calculating macroscopic quantities in terms of their microscopic equivalents;

(3) Showed that the appropriate description of the response of a material to an electric field is its dielectric function ε ;

(4) Showed that a dielectric function $\varepsilon = \varepsilon(\omega)$ can be defined for any system where the time dependence of the drive $\vec{E}$ and response $\Delta\vec{r}$ are the same, which applies not only to the static limit but especially to the harmonic time dependence $e^{-i\omega t}$;

(5) Provided the recipe for calculating ε in terms of atomic-scale parameters, and in so doing have

(6) Laid the foundation for describing the response of systems involving more general configurations including for example coupled charges, anisotropic and nonlinear restoring forces, magnetic-field effects, etc. The associated phenomena include electromagnetically induced transparency, and crystal and nonlinear optics.

(7) Indirectly introduced the concept of a local field, because the force acting on q is not necessarily the macroscopic field but the field at the charge site itself.

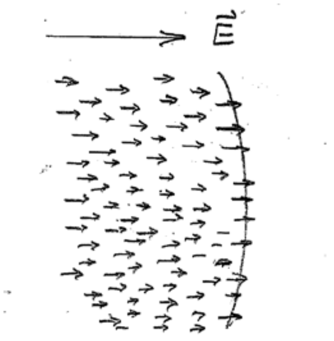
It is worth repeating that for harmonic time dependences $\varepsilon = \varepsilon(\omega)$. As shown below, to within a scaling factor $\left(\varepsilon(\omega)-1\right)$ is the Fourier transform of the Green function of the force equation.

We conclude this section by considering the physics of polarization charge and the displacement field. The relevant equations are

$$\nabla \cdot \vec{D} = 4\pi\rho_o ; \qquad (8.31a)$$

$$\nabla \cdot \vec{E} = 4\pi(\rho_o + \rho_P) = 4\pi\rho_o + 4\pi(-\nabla \cdot \vec{P}) . \qquad (8.31b)$$

Drawing a Gaussian pillbox straddling the interface and applying Gauss' Theorem to Eq. (8.31a) shows that $\nabla \cdot \vec{D}$ is sensitive only to the original unperturbed charge density. On the other hand, the same calculation applied to Eq. (8.31b) shows that $\nabla \cdot \vec{E}$ picks up not only this charge density but also the density resulting from the induced polarization. In practical terms this means that if ε is different on two sides of a boundary, then the normal component of $\vec{D}$ will be continuous but the normal component of $\vec{E}$ discontinuous as a result of the induced polarization, with the smaller normal component of $\vec{E}$ appearing in the material with the larger ε. This behavior is termed screening.



The figure at the right illustrates the physics taking place at an interface between a material and a vacuum. If the material is uniform, the resulting dipole density in the bulk is also uniform and $(-\nabla \cdot \vec{P}) = 0$ there as well. This can be viewed as the tails of dipoles being cancelled by the heads of adjacent dipoles. This cancellation clearly stops at an interface. Doing the pillbox derivation at the front interface, we find

$$\int_V d^3 r' (-4\pi \nabla \cdot \hat{P}) = \int_S d^2 r' \hat{n} \cdot (-4\pi \vec{P}) = +\sigma_P \Delta A \tag{8.32a}$$

$$= (E_2 - E_1)\Delta A, \tag{8.32b}$$

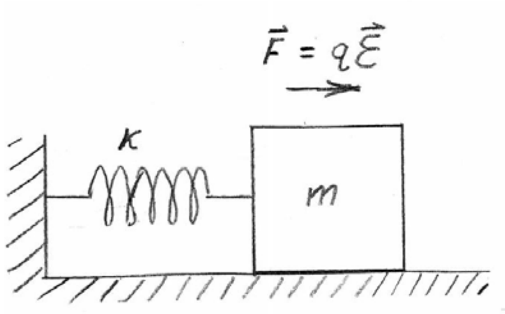
where σ_P is the interface polarization charge density and ΔA is the area of the relevant side. Thus $E_1 < E_2$, consistent with $\varepsilon E_1 = E_2$ for $\varepsilon > 1$. The technical term for this reduction of the field inside the material is screening. In the static limit, screening in metals is perfect, as we saw in Ch. 3.

If the material were not uniform, then $(-\nabla \cdot \vec{P})$ would leave behind a trail of charge, because the positive ends of the dipoles would not be cancelled exactly by the negative ends of the adjacent dipoles. An example of a nonuniform material is an epitaxial layer of the alloy $Al_xGa_{1-x}As$ where the composition *x* is graded across the layer. Because ε is a function of composition, it is therefore also a function of position. However, such situations are less common.

D. The mechanical model: general properties of $\varepsilon(\omega)$.

A mechanical analog that includes all immediately relevant interactions is that of a mass connected by a spring to a rigid post, with the mass sliding on a horizontal surface and experiencing viscous friction (see figure). With $\vec{E}$ applied, the equation of motion is

$$m\frac{d^2\vec{r}}{dt^2} = q\vec{E} - b\frac{d\vec{r}}{dt} - \kappa(\vec{r} - \vec{r}_o) \tag{8.33}$$



where the left side describes the inertial response $m\vec{a}$, and the right side the driving force $q\vec{E}$, viscous friction $b\, d\vec{r}/dt$, and a linear (Hooke's-Law) restoring force of spring constant κ and equilibrium position $\vec{r}_o$. For passive systems, the type that we consider here, *b* is positive definite because friction always acts in the direction opposing the velocity; passive systems can only dissipate energy. As noted above, Eq. (8.33) is easily generalized to describe more complex situations. For example, we can add an anharmonic restoring-force term to describe nonlinear optics, and make κ tensorial to describe crystal optics. Both extensions are discussed in Sec. F.

For now, we consider Eq. (8.33) as-is, with $\vec{E}$ having the time dependence $\vec{E}(t) = \vec{E}e^{-i\omega t}$. This is relevant for several reasons. First, plane electromagnetic waves have this time dependence, so the results describe optics in general. Short electromagnetic pulses can also be analyzed this way by Fourier transforming their time dependences and treating the result in reciprocal space. Second, harmonic functions are solutions of the homogeneous ($\vec{E} = 0$) version of Eq. (8.33), and hence will be used in Sec. G to construct the Green function of Eq. (8.33). This gives the general response of a system to an electric field with an arbitrary time dependence $\vec{E}(t)$. Third, the dielectric function $\varepsilon(\omega)$ that results is found to embody a substantial amount of physics, including reality, causality, the Kramers-Kronig relations, and sum rules, among other considerations.

Accordingly, we proceed. For $\vec{E}(t) = \vec{E}e^{-i\omega t}$, the obvious solution of Eq. (8.33) is based on

$$\vec{r} = \vec{r}(t) = \vec{r}_o + \Delta\vec{r}e^{-i\omega t}, \tag{8.34}$$

since this reduces Eq. (8.33) to algebra. It follows that

$$-m\omega^2\Delta\vec{r} = q\vec{E} + i\omega b\Delta\vec{r} - \kappa\Delta\vec{r} . \tag{8.35a}$$

or

$$\Delta\vec{r} = \frac{q\vec{E}}{\kappa - m\omega^2 - i\omega b} . \tag{8.35b}$$

Therefore, from Eq. (8.30)

$$\varepsilon(\omega) = 1 + \frac{4\pi nq^2}{\kappa - m\omega^2 - ib\omega} . \tag{8.36}$$

We can write this in more convenient form by dividing numerator and denominator by *m*, noting that $\kappa/m = \omega_o^2$ is the resonant frequency of the spring/mass system in the absence of friction. Then

$$\varepsilon(\omega) = 1 + \frac{4\pi nq^2 / m}{\omega_o^2 - \omega^2 - 2i\omega\Gamma}, \tag{8.37}$$

where we have defined $b/m = 2\Gamma$. For passive systems Γ is positive definite. It describes decay rates for transients, as well be described shortly.

Noting further that $\varepsilon(\omega)$ is dimensionless, the numerator must also be a frequency squared. Accordingly, define the plasma frequency ω_p as

$$\omega_p^2 = \frac{4\pi nq^2}{m}, \tag{8.38}$$

which is also physically significant. Accordingly, in the form that we will use it,

$$\varepsilon(\omega) = \varepsilon_1(\omega) + i\varepsilon_2(\omega) = 1 + \frac{\omega_p^2}{\omega_o^2 - \omega^2 - 2i\Gamma\omega}. \tag{8.39}$$

The notation $\varepsilon_1(\omega)$ and $\varepsilon_2(\omega)$ for the real and imaginary parts of $\varepsilon(\omega)$ is conventional, and is used extensively in the following.

The terms ω_o^2, ω^2, and $2\Gamma\omega$ in the denominator can be identified as originating from the restoring force, inertia, and friction, respectively. Because all materials contain several different classes of charge, for example ions of a lattice as well as free electrons in a metal or electrons in bonding orbitals in an insulator, a complete picture requires additional terms with different values of ω_o, Γ, and ω_p. Because Eq. (8.33) is linear, in this approximation these terms can be added with no complications. When a multiterm approach such as this is used to describe dielectric responses of materials, it is called the *spectral representation*. Because the parameters in these terms are material- and structure-dependent, uses of $\varepsilon(\omega)$ include identifying materials and analyzing structures through their dielectric-function spectra. Examples are given in the next section.

Equation (8.39) provides the basis not only for understanding dielectric functions of materials but also causal response in general. As a result of its origin in mechanics, it incorporates not only linearity but also causality, which in turn gives rise to the Kramers-Kronig relations between its real and imaginary parts. These generate sum rules, as discussed in following sections. As noted above, extensions of the force equation that result from adding anharmonic terms or making restoring forces directional yield more elaborate versions of Eq. (8.39) that describe nonlinear and anisotropic behavior, among other phenomena. These are discussed in Sec. F.

Before considering specific materials, we examine one general characteristic of $\varepsilon(\omega)$ known as the reality condition. When working with complex functions, reality conditions are necessary because measurable quantities such as dipole densities $\vec{P}(t)$ are real. To go from the frequency domain above to the time domain we Fourier-transform $\vec{P}(\omega)$ according to

$$\vec{P}(t) = \int_{-\infty}^{\infty} d\omega \vec{P}(\omega) e^{-i\omega t} \ . \tag{8.40a}$$

Whereas transformations from reciprocal to real space typically involve $e^{i\omega t}$ instead of $e^{-i\omega t}$, we use $e^{-i\omega t}$ to be consistent with the physics convention, which is used in the derivation of $\varepsilon(\omega)$. We assign the mandatory prefactor $(2\pi)^{-1}$ to the complementary transform

$$\vec{P}(\omega) = \frac{1}{2\pi} \int_{-\infty}^{\infty} dt \vec{P}(t) e^{i\omega t} \ . \tag{8.40b}$$

Now because $\vec{P}(t)$ is real, $\vec{P}^*(t) = \vec{P}(t)$. Hence

$$\vec{P}^*(t) = \int_{-\infty}^{\infty} d\omega \vec{P}^*(\omega) e^{i\omega t} = \int_{\infty}^{-\infty} d(-\omega) \vec{P}^*(-\omega) e^{-i\omega t} \tag{8.41a}$$

$$= \int_{-\infty}^{\infty} d\omega \vec{P}^*(-\omega) e^{-i\omega t} = \vec{P}(t) \, . \tag{8.41b}$$

Therefore

$$\vec{P}^*(-\omega) = \vec{P}(\omega) \, . \tag{8.42}$$

This is the reality condition for $\vec{P}(\omega)$.

Repeating the calculation for $\vec{E}(t)$ leads to the conclusion that $\vec{E}^*(-\omega) = \vec{E}(\omega)$. Now since

$$\vec{P}(\omega) = \frac{1}{4\pi} \left(\varepsilon(\omega) - 1 \right) \vec{E}(\omega) \, , \tag{8.43}$$

taking the complex conjugate of Eq. (8.43), and using Eq. (8.42) and the corresponding condition $\vec{E}^*(-\omega) = \vec{E}(\omega)$ shows that

$$\varepsilon^*(-\omega) = \varepsilon(\omega) \, . \tag{8.44}$$

This is the reality condition for $\varepsilon(\omega)$. Equation (8.39) clearly obeys this condition. Consequently, its real and imaginary parts are even and odd in ω:

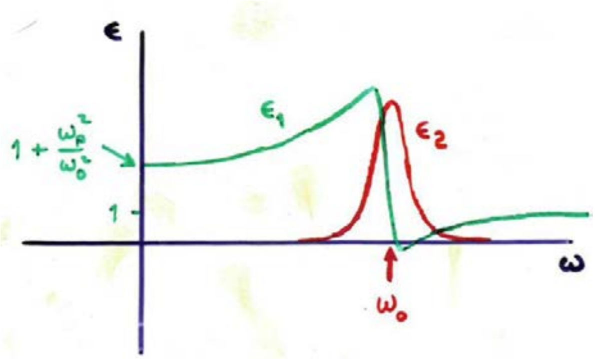
$$\varepsilon_1^*(\omega) = \varepsilon_1(-\omega) \, ; \ \ \varepsilon_2^*(\omega) = -\varepsilon_2(-\omega) \, . \tag{8.45a,b}$$

These conditions will be used in Sec. I to obtain compact versions of the Kramers-Kronig relations.

In spectroscopic analysis $\varepsilon_2(\omega)$ is more important than $\varepsilon_1(\omega)$ because $\varepsilon_2(\omega)$ is localized in frequency. To show this, let $\omega \sim \omega_o >> \Gamma$. Then to a term of order $(\Gamma/\omega_o)^2$ in the denominator, Eq. (8.39) can be approximated as

$$\varepsilon = \varepsilon(\omega) = 1 + \frac{\omega_p^2}{\omega_o^2 - \omega^2 - 2i\Gamma\omega} \approx 1 + \frac{\omega_p^2}{(\omega_o + \omega + i\Gamma)(\omega_o - \omega + i\Gamma)} \qquad (8.46a,b)$$

$$\approx 1 + \frac{\omega_p^2(\omega_o - \omega + i\Gamma)}{2\omega_o(\omega_o - \omega + i\Gamma)(\omega_o - \omega + i\Gamma)} = 1 + \frac{\omega_p^2}{2\omega_o}\frac{(\omega_o - \omega) + i\Gamma}{\left((\omega - \omega_o)^2 + \Gamma^2\right)} \qquad (8.46c,d)$$



The Lorentz form of the $\varepsilon_2(\omega)$ is evident. $\varepsilon_1(\omega)$ follows what is called the dispersion curve. The figure illustrates their behaviors.

$\varepsilon_2(\omega)$ is also useful because it describes the dissipation of energy in a medium. While this is usually discussed in connection with plane-wave propagation, a general demonstration is also possible. Start with the rate of change of the energy density from Ch. 1:

$$\frac{dU}{dt} = -\frac{1}{4\pi}\vec{E} \cdot \frac{d\vec{D}}{dt}\,. \qquad (1.45)$$

Because Eq. (1.45) is nonlinear, the fields that are used in the evaluation must be real. Accordingly, write

$$\vec{E}(t) = \mathrm{Re}\left(\vec{E}_o e^{-i\omega t}\right) = \vec{E}_o \cos\omega t\,; \qquad (8.47)$$

$$\frac{d(\vec{D}(t))}{dt} = \mathrm{Re}\left(-i\omega\,\varepsilon(\omega)\vec{E}_o e^{-i\omega t}\right) \qquad (8.48a)$$

$$= \mathrm{Re}\left(-i\omega(\varepsilon_1 + i\varepsilon_2)\vec{E}_o(\cos\omega t + i\sin\omega t)\right) \qquad (8.48b)$$

$$= \omega\vec{E}_o\left(\varepsilon_1 \sin\omega t + \varepsilon_2 \cos\omega t\right). \qquad (8.48c)$$

Then

$$\frac{dU}{dt} = -\frac{\vec{E}_o^2}{4\pi}\left(\varepsilon_1 \sin\omega t\cos\omega t + \varepsilon_2 \cos^2\omega t\right). \qquad (8.49)$$

The first term represents reactive transfer of power between the electromagnetic field and the energy stored in the electromagnetic field and the mechanical strain of the material, and time-averages to zero. The second term represent net transfer of energy from the field to the material, and time-averages to $1/2$. Thus ε_2 determines absorption of electromagnetic energy by a material.

E. Dielectric properties of real materials.

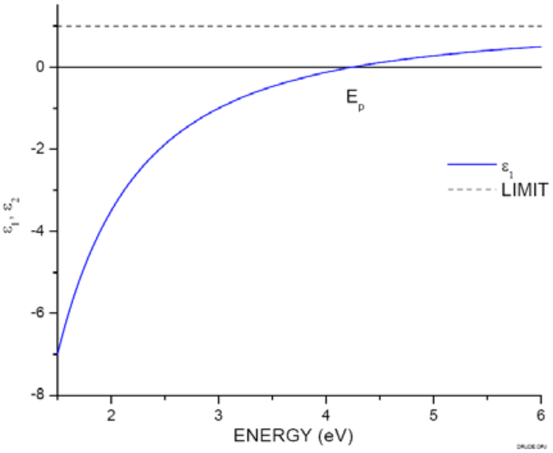
Having developed the basic form of $\varepsilon(\omega)$, the next step is to verify that the approach works. This section provides representative examples of these data, which also indicate useful limits to the general expression Eq. (8.39).

E1. Metals: no restoring force, $\kappa = 0$.

For free carriers such as electrons in a metal, and electrons or holes in a doped semiconductor, the spring constant κ and therefore ω_o are zero. Ignoring loss, Eq. (8.39) reduces to

$$\varepsilon(\omega) = 1 - \frac{\omega_p^2}{\omega^2} . \tag{8.50}$$

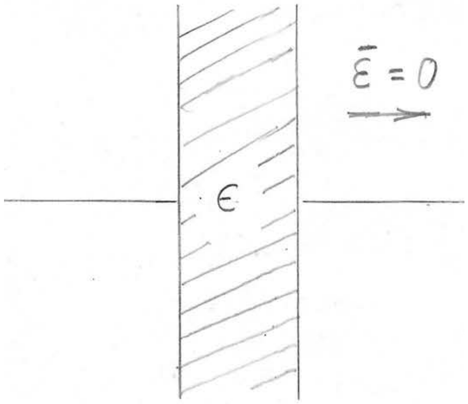
This is the Drude model of a metal. The behavior of $\varepsilon(\omega)$ in this case for a plasma energy $E_p = \hbar\omega_p = 4.2$ eV is shown in the figure. The significance of the zero crossing and the plasma terminology is discussed below. As $\omega \to 0$, $\varepsilon(\omega)$ diverges as $1/\omega^2$. As $\omega \to \infty$, $\varepsilon(\omega) \to 1$ from below.



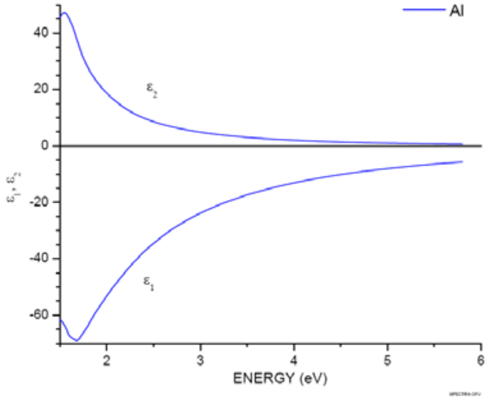
In the limit of very high frequencies the Drude model is again significant because at sufficiently high frequencies the inertial term dominates completely and the charge simply stops moving. In this case dissipation and restoring forces become irrelevant and the associated terms in the denominator of Eq. (8.39) can be neglected. The result is Eq. (8.50). Quantum mechanics views this as the situation where the energy of a photon far exceeds the binding energy of an electron with which it is interacting, so once again the binding energy can be ignored. These considerations make X-ray optics relatively straightforward: if the energy of a core-level electron is less than that of the photon it's a free electron; if it's greater, the electron can be ignored.

The use of the term "plasma frequency" for ω_p can be clarified by considering the following. Suppose a parallel-plate capacitor is filled with a material of dielectric function ε, as indicated in the diagram. The electric field $\vec{E}$ outside the capacitor is zero. Is there a condition such that an electric field $\vec{E}$ can exist inside the capacitor?

This might seem like a strange question, given that the boundary condition is normal $\vec{D}$ continuous, and with $\vec{D} = 0$ on the outside, then $\vec{D} = 0$ inside the capacitor as well. But consider the constitutive equation $\vec{D} = \varepsilon\vec{E}$. If $\varepsilon = 0$, then $\vec{E}$ can be finite without violating any conditions. Equation (8.50) shows that this occurs at $\omega = \omega_p$. This excitation, which consists of oscillating screening charges developing on the plates at a frequency $\omega = \omega_p$, is termed a *plasmon*. Thus the nomenclature ω_p. The physics is straightforward: with everything oscillating at $\omega = \omega_p$, the energy in the system transfers cyclically between energy stored in the field and the kinetic energy of the charges, which have a finite mass. A spherically symmetric version also exists, and can be excited by implanting a charge *q* in a metal. If *q* is positive, electrons surrounding *q* respond by rushing in, overshoot, then bounce back, the process repeating until the energy of the plasmon is dissipated.



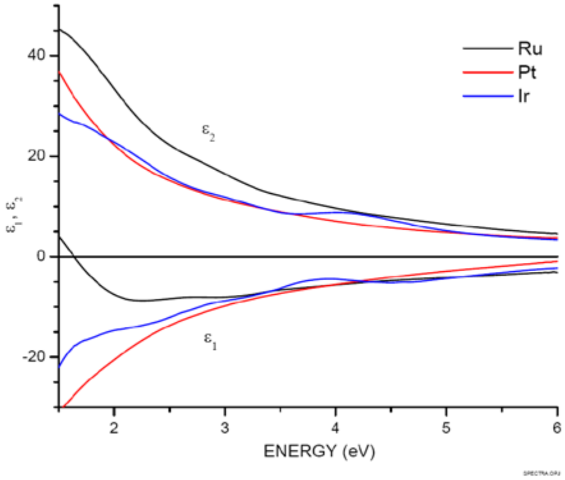
The Drude model is a poor approximation of the dielectric responses of real metals, because it does not include the effects of energy loss due to electron scattering. For example, the figure at the right shows $\varepsilon(\omega)$ data for Al from 1.5 to 6.0 eV. For this material $E_P = \hbar\omega_P \approx 15$ eV, and so is far off scale to the right. ε_1 essentially follows the Drude behavior of the previous figure except for the small structure near 1.7 eV, which is a consequence of an interband transition in the material. The main difference is the significant value of ε_2.



However, the Drude model is easily extended to describe loss. We retain the viscous-friction term, so the improved expression is

$$\varepsilon(\omega) = 1 - \frac{\omega_p^2}{\omega^2 + i2\omega\Gamma} \qquad (8.51)$$

With Γ finite, $\varepsilon_2(\omega)$ extends into the 1.5 – 6.0 eV range, providing a better description of Al, as is seen above. Transition metals such as Ni and Fe have partially filled d bands, so the density of states into which electrons can scatter is very large. As a result, losses and therefore Γ are large as well. This is indicated by the data at the right, which show $\varepsilon(\omega)$ for several Pt-group metals. The large values of ε_2 reveal that losses in these metals are large at optical frequencies. When considering reflectance, we will find that metals with large losses in the optical-frequency range are not good reflectors. Consequently, transition metals have a characteristic gray appearance compared to Ag and Al, where

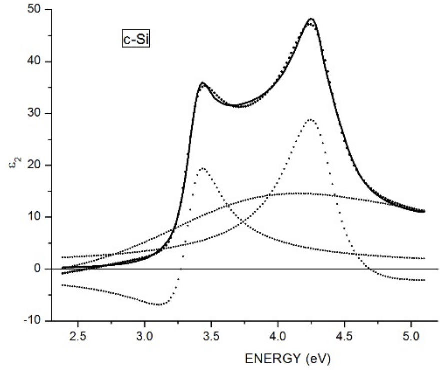


the losses are relatively small. This will be discussed in detail in Ch. 13.

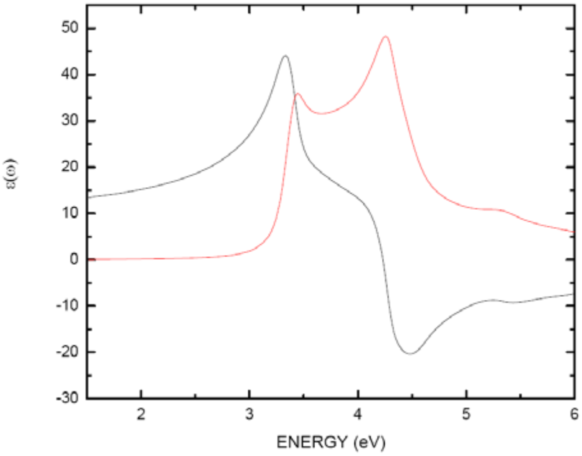
E2. Restoring forces: the Lorentz model.

Even with loss included, the free-electron model fails for semiconductors and insulators, where electrons are localized in bonds. These configurations are described by retaining all terms in the denominator of Eq. (8.39). Usually $\Gamma << \omega_o$, so Eq. (8.46d) applies. Thus ε_2 spectra of semiconductors and insulators tend to be superpositions of Lorentzian structures.

The dielectric function of Si shown in the figure illustrates this point and several others. Resonances involving lattice modes appear in the infrared, those involving bonding electrons in the visible-near uv, and those involving core levels in the soft- or hard-X-ray range. Materials typically have several resonances within each class. The crystalline Si example exhibits prominent resonances near 3.4 and 4.2 eV, a small one near 5.4 eV, and a very broad background resonance centered near 3.6 eV.



Thus a passable analytic representation of the dielectric function of Si in the visible-near uv spectral range below 5 eV would consist of three resonances. To drive this point home, the least-squares fit of three oscillators to ε_2 over this range is shown at the right. The good fit shows that this idea is basically correct. These three oscillators provide an excellent representation of $\varepsilon_1(\omega)$ and $\varepsilon_2(\omega)$ in this spectral range. The constituent lineshapes are a mix of their real and imaginary parts owing to electron interactions that are not considered here.



A broader spectral range would include similar resonances in infrared due to the ions of the crystal lattice, and in the X-ray range due to transitions from deep-lying core levels. These responses are typically much weaker. Lattice contributions involve ions, whose masses are orders of magnitude larger than that of electrons. Displacements are correspondingly reduced. The tight binding of core electrons also results in small displacements. If present, free electrons or holes in the conduction and valence bands of doped semiconductors contribute a Drude term in the infrared. The relatively wide separations in energy mean that the different classes of charges are essentially independent, so their contributions to $\varepsilon(\omega)$ are additive, as noted before.

Continuum resonances can be superposed to produce mathematical functions other than Lorenzians to better represent spectral dependences in certain regions. An example is the square-root dependence of $\varepsilon_2(\omega)$ near the fundamental absorption edge of crystalline semiconductors. Here, a continuum of Lorentzian oscillators is weighted by a density of states calculated by assuming momentum conservation in the optical transition. The square-root dependence is the result. A more elaborate example is the Tauc-Lorentz expression that describes the behavior of $\varepsilon_2(\omega)$ in the same range for amorphous semiconductors. In amorphous materials momentum no longer a good quantum number, a different mathematical form emerges. If the objective is to provide an efficient mathematical representation of the dielectric function valid over a limited range of energy, we tend to be generous and use whatever works, even if the underlying physics is difficult to justify.

An unattractive aspect of Lorentz lineshapes is their slow convergence away from the resonance. As a result, Lorentzian versions of ε_2 never truly reach zero until $\omega = 0$. This implies that materials such as glass are never quite transparent. While the Lorentizian lineshape applies to most absorptive processes near resonances, data that exhibit relatively abrupt cutoffs beyond several Γ are more typically Gaussian. The Voight lineshape, which is a convolution of a Lorentzian and Gaussian, is often used as a compromise.

E3. Low and high frequency limits; windows of transparency.

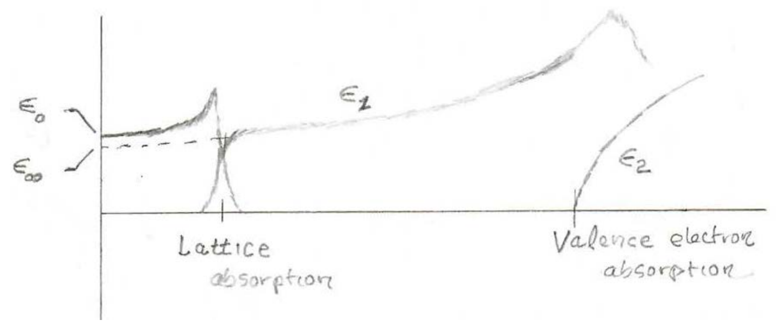
In the absence of free charge, Eq. (8.39) for a single oscillator reduces to

$$\varepsilon(\omega) = \varepsilon_1(\omega) = 1 + \frac{\omega_p^2}{\omega_o^2 - \omega^2}, \tag{8.52a}$$

$$\cong 1 + \frac{\omega_p^2}{\omega_o^2} + \frac{\omega_p^2 \omega^2}{\omega_o^4} \ldots . \tag{8.52b}$$

Thus at low frequencies, for insulators and intrinsic semiconductors $\varepsilon(\omega)$ is real, $\varepsilon(0)$ is always greater than 1, and $\varepsilon(\omega)$ always increases with increasing ω. Since the refractive index $n = \sqrt{\varepsilon}$, the low-frequency refractive indices of these materials has the same behavior.

The existence of different classes of charge, for example the ions that comprise the crystal lattice, the valence electrons that form compounds and crystals, and core electrons that compensate the positive charge of the nuclei, all exhibit this behavior at energies much smaller than their resonant energies. If these resonances are well separated in energy, as is usually the case, then as ω



is decreased the resonances successively add their contributions to $\varepsilon_1(\omega)$, as indicated schematically in the figure at the bottom of the previous page. In this case the 1 in Eqs. (8.52) is replaced by the static dielectric constant ε_o, or

$$\varepsilon_1(\omega) = \varepsilon_o + \frac{\omega_p^2}{\omega_o^2 - \omega^2} \tag{8.53}$$

In this case ω_p^2 and ω_o^2 do not correspond to any particular transitions, but are treated as phenomenological parameters. If the electronic contribution is extrapolated to $\omega = 0$, i.e., the lattice constant is ignored, then ε_o becomes the infrared dielectric constant ε_∞ and ω_p^2 and ω_o^2 assume different meanings. Although the term "dielectric constant" is in common usage, the only time it is appropriate to use the term "constant" in connection with $\varepsilon(\omega)$ is in these two limits.

In modeling applications, Eq. (8.53) is termed the Sellmeyer approximation. It can also be expressed in terms of wavelength, where the wavelength $\lambda = hc/E$, where hc = 1.23955 eV μm for propagation in air. The two expansions in common use are

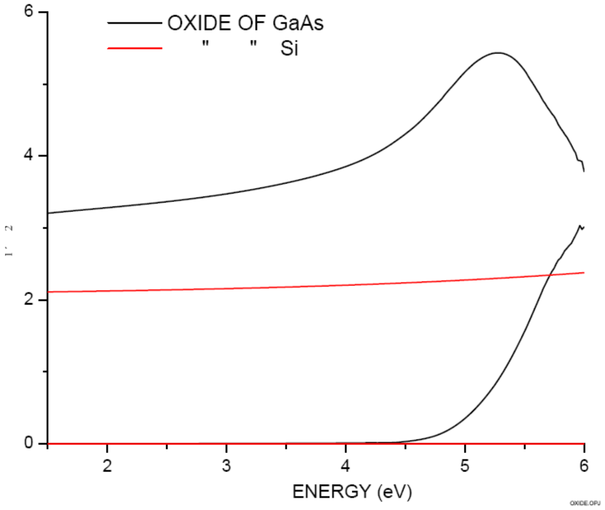
$$\varepsilon(\lambda) \approx \varepsilon_\infty + \frac{A\lambda^2}{\lambda^2 - \lambda_o^2} \tag{8.54a}$$

$$\approx \varepsilon_\infty + \frac{A'}{\lambda^2} + \frac{B'}{\lambda^4} + \ldots \tag{8.54b}$$

Equation (8.54a) is written in terms of the empirical parameters A and λ_o^2. The second, using A' and B', is termed the Cauchy representation.

The regions between absorption classes are called windows of transparency. The same general characteristics term is usually used in connection with the visible range, which extends from about 1.7 to 3.1 eV (wavelengths of 700 to 400 nm.) An example is provided in the figure, which shows $\varepsilon_1(\omega)$ for the oxides of Si and GaAs from 1.5 to 6.0 eV. The absorption edge for the oxide on GaAs occurs within this range, while that of SiO2 occurs near 9 eV and hence is out of the range of the figure.

The increase of $\varepsilon(\omega)$ with frequency is termed dispersion, and has at least two practical consequences. First, it is impossible for a lens made of a single material to focus light at the same point for long and short wavelengths. Achromatic lenses can be fabricated that approximate this behavior, but do so by combining concave and convex lenses



of different materials in ways that capitalize on differences in dispersions.

The second consequence concerns communications technology. Here, information is transmitted as optical pulses in optical fibers. Pulses are engineered to have as short a duration as possible. However, Fourier analysis reminds us that a short pulse is the result of the phase-coherent superposition of wavelengths in a range centered about the carrier wavelength, and that the shorter the pulse, the broader the range of wavelengths that it contains. With dispersion, the components of the pulses travel at different speeds and gradually get out of phase, with the longer-wavelength components traveling faster. As a result, the pulse broadens. The limiting distance between repeaters or substations is that where separate pulses can no longer be distinguished. A considerable amount of engineering effort has gone into addressing this problem. These aspects will be considered in detail next semester.

F. Nonlinear and crystal optics.

The simple models discussed above give only a brief introduction to the wide range of dielectric-response phenomena that can be addressed through atomic-scale formulations. Two generalizations worth considering explicitly are anharmonicity and anisotropy. These are the entry points to nonlinear optics and crystal optics, respectively. We consider anharmonicity first.

Nonlinear-optical phenomena such as second-harmonic or sum-frequency generation can be described by adding an anharmonic restoring force $\vec{\kappa}_2 \cdot \Delta\vec{r}\Delta\vec{r}$ to Eq. (8.33):

$$m\frac{d^2\vec{r}}{dt^2} = q\vec{E}e^{-i\omega t} - b\frac{d\vec{r}}{dt} - \underset{\sim}{\kappa}(\vec{r}-\vec{r}_o) - \underset{\sim}{\kappa}_2(\vec{r}-\vec{r}_o)(\vec{r}-\vec{r}_o)\,. \qquad (8.57)$$

Here, $\underset{\sim}{\kappa}$ and $\underset{\sim}{\kappa}_2$ are second- and third-rank tensors, respectively. To avoid complications that are unnecessary in the present context, we assume that the material is isotropic so $(\vec{r}-\vec{r}_o) = \Delta\vec{r}$ is in the direction of $\vec{E}$, so all terms reduce to scalars. Equation (8.57) can therefore be written

$$\left(m\frac{d^2\Delta r}{dt^2} + 2\Gamma\frac{d\Delta r}{dt} + \kappa\Delta r + \kappa_2\Delta r\Delta r \right) = qEe^{-i\omega t}\,. \qquad (8.58)$$

In the absence of the anharmonic term, the solution $\Delta r = \Delta r(t)$ has the same time dependence as the applied field. Hence the related anharmonic term has a time dependence $e^{-i2\omega t}$. Therefore, $\Delta\vec{r}(t)$ must be generalized to include a second-harmonic term:

$$\Delta\vec{r}(t) = \Delta\vec{r}_1 e^{-i\omega t} + \Delta\vec{r}_2 e^{-i2\omega t}\,. \qquad (8.59)$$

In practice such nonlinearities are very small, so no higher harmonics need be considered and the equations can be solved iteratively, using the anharmonic term as the source term for the time dependence $e^{-i2\omega t}$. The zero-order solution is the same as before:

$$\Delta r_1 = \frac{qE}{\kappa - m\omega^2 - 2im\omega\Gamma} = \frac{qE/m}{\omega_o^2 - \omega^2 - 2i\omega\Gamma}\,. \tag{8.60}$$

Working through the math, the second-harmonic term is therefore

$$\Delta r_2 = \frac{\kappa_2 \Delta r_1^2}{\kappa - 4m\omega^2 - 4im\omega\Gamma} = \frac{\kappa_2 (qE/m)^2}{(\kappa - 4m\omega^2 - 4im\omega\Gamma)(\omega_o^2 - \omega^2 - 2i\omega\Gamma)^2}\,. \tag{8.61}$$

These simple harmonic-oscillator models contain a significant amount of physics, and lead to relatively efficient descriptions of nonlinear-optical data. For example, if two or more fields with different frequency dependences are present, sum and difference mixing terms result. In the early days of lasers, it was noted that the second-order susceptibility is proportional to product of the first-order susceptibilities at the three frequencies that are involved in the general (two-driving-field) version of Eq. (8.61). This is known as Miller's Rule, and it follows directly from the structure of Eq. (8.61). Strictly speaking, Miller's Rule is a good approximation only in the static limit.

We consider anisotropy next. Suppose that a material is described in the *xy* plane by an anisotropic restoring force such that the Hooke's-Law tensor in the laboratory *x*- and *y*-axis frame can be written as

$$\underline{\kappa} = \begin{pmatrix} \kappa & \Delta\kappa \\ \Delta\kappa & \kappa \end{pmatrix}. \tag{8.62}$$

Suppose also that a field $\vec{E}(t)e^{-i\omega t} = (\hat{x}E_x + \hat{y}E_y)e^{-i\omega t}$ is applied to the material. Then if there is no dissipation, the force law requires

$$m\frac{d^2}{dt^2}\begin{pmatrix} \Delta x \\ \Delta y \end{pmatrix} e^{-i\omega t} = -m\omega^2 \begin{pmatrix} \Delta x \\ \Delta y \end{pmatrix} e^{-i\omega t} = q\begin{pmatrix} E_x \\ E_y \end{pmatrix} e^{-i\omega t} - \begin{pmatrix} \kappa & \Delta\kappa \\ \Delta\kappa & \kappa \end{pmatrix}\begin{pmatrix} \Delta x \\ \Delta y \end{pmatrix} e^{-i\omega t}\,, \tag{8.63}$$

which can be rewritten

$$\begin{pmatrix} \kappa - m\omega^2 & \Delta\kappa \\ \Delta\kappa & \kappa - m\omega^2 \end{pmatrix}\begin{pmatrix} \Delta x \\ \Delta y \end{pmatrix} = q\begin{pmatrix} E_x \\ E_y \end{pmatrix}. \tag{8.64}$$

Thus $\varepsilon = \overleftrightarrow{\varepsilon}$ will be a tensor, because each Cartesian component of

$$\vec{D} = \overleftrightarrow{\varepsilon} \cdot \vec{E} = \vec{E} + 4\pi\, n\, q\, \Delta\vec{r} \tag{8.65}$$

will receive contributions from both E_x and E_y.

The efficient way to solve this problem is to determine the normal modes of the system. In the force-equation context, normal modes are linear combinations of the displacement components in the laboratory coordinate system that result in a restoring-force tensor $\overleftrightarrow{\kappa}$ that is diagonal. This reduces the original coupled two- or three-dimensional problem into two or three independent one-dimensional calculations. If no magnetic field is present, this process defines new coordinate axes where the displacements are independent. By taking components of the applied field along these new directions, these one-dimensional force equations can be solved exactly, as we have already done for isotropic systems, and the results superposed.

What is a normal mode? By definition, a normal mode in a displacement context is a linear combination of displacements such that when the restoring-force matrix acts on it, it returns the same linear combination multiplied by a scaling factor. As an example, let

$$\begin{pmatrix} \kappa & \Delta\kappa \\ \Delta\kappa & \kappa \end{pmatrix}\begin{pmatrix} \Delta x \\ \Delta y \end{pmatrix} = \gamma \begin{pmatrix} \Delta x \\ \Delta y \end{pmatrix} = \begin{pmatrix} \gamma & 0 \\ 0 & \gamma \end{pmatrix}\begin{pmatrix} \Delta x \\ \Delta y \end{pmatrix}, \tag{8.66}$$

where γ is the scaling factor to be determined, and κ and $\Delta\kappa$ are the Hooke's-Law spring constants parallel and perpendicular, respectively, to the bond. In usual diagonalization discussions, one jumps immediately to

$$\begin{pmatrix} \kappa-\gamma & \Delta\kappa \\ \Delta\kappa & \kappa-\gamma \end{pmatrix}\begin{pmatrix} \Delta x \\ \Delta y \end{pmatrix} = 0, \tag{8.67}$$

which is obtained by moving the γ matrix in Eq. (8.66) to the left side of the equation. Nontrivial solutions occur only if the determinant of the matrix is zero. The resulting equation is

$$(\kappa-\gamma)^2 - (\Delta\kappa)^2 = 0. \tag{8.68}$$

The solutions are

$$\gamma_{\pm} = \kappa \pm \Delta\kappa. \tag{8.69}$$

The eigenvectors are obtained by substituting these solutions in the defining matrix. The result is

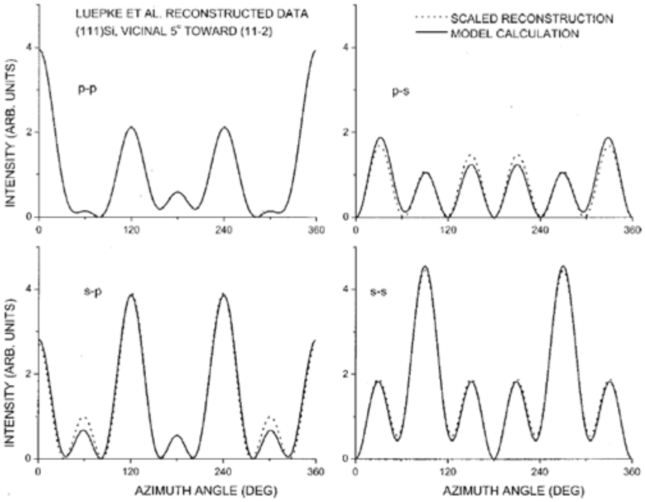
$$\begin{pmatrix} \mp\Delta\kappa & \Delta\kappa \\ \Delta\kappa & \mp\Delta\kappa \end{pmatrix}\begin{pmatrix} \Delta x \\ \Delta y \end{pmatrix} = 0. \tag{8.70}$$

Hence the two normal modes are $\Delta\hat{u} = \Delta u(\hat{x}+\hat{y})/\sqrt{2}$ with an effective spring constant $\kappa' = \kappa + \Delta\kappa$, and $\Delta\hat{v} = \Delta v(\hat{x}-\hat{y})/\sqrt{2}$ with an effective spring constant $\kappa' = \kappa - \Delta\kappa$. The problem is most easily treated by projecting the incident field along these two directions, solving the two resulting one-dimensional equations, and superposing the results.

If a magnetic field is present, the resulting equations are one-dimensional but the normal modes themselves are linear combinations of the displacements along the laboratory axes. We will take this up in more detail in Ch. 12.

Anharmonicity and anisotropy were recently combined to describe the second-harmonic generation that occurs at the surface of centrosymmetric crystals, specifically silicon. Silicon is centrosymmetric, so second-harmonic generation is forbidden in the bulk. However, it is allowed at surfaces and interfaces, where the symmetry is lower. Thus Si was an attractive material on which to do surface-sensitive second-harmonic generation. Typical single wavelength measurements were done using polarized light at non-normal incidence, then rotating the sample and measuring the resulting variation in scattered intensity. Four combinations of measurements are possible: illumination by transverse-electric (TE) or transverse-magnetic (TM) radiation, followed by TE or TM detection.

Representative data are shown in the figure at the right for the four combinations of polarization. The theoretical calculations shown for comparison were done with the mechanical-oscillator model using only 3 parameters. It is clear that the model fits the data excellently well. nonlinear-optical phenomena. These three parameters, replace the 14 tensor components and 11 Fourier components required to fit these data.



As can be appreciated, this approach shows that nonlinear-optical phenomena can be viewed in general as the result of deviations from linearity caused by overdriving a system, analogous to the type of distortion encountered for when an audio system is overdriven. In the nonlinear-optical context, the anharmonic motion gives rise to radiation at the relevant harmonics, which is seen as second- or third-harmonic generation, sum-frequency generation, etc. Raman scattering provides another example. Here, the nonlinearity is a result of the spring "constant" being modulated by a phonon of frequency ω_v, which leads to sidebands at $\omega \pm \omega_v$.

G. Poles, Green functions, and the response in the time domain.

We already drew some conclusions about general behavior of $\varepsilon(\omega)$, specifically the reality condition summarized as Eqs. (8.42) and (8.44). In this section we consider general properties due to causality. The path starts with the Green function $G(t,t')$ of the force equation, where t and t' are the times associated with the observer and source, respectively. Causality requires that $G(t,t')$ must be zero for $t < t'$. After completing the development, we find that for passive materials causality requires the poles of $\varepsilon(\omega)$ to lie in the lower half of the complex ω plane.

The derivation starts by defining $G(t,t')$. While $G(t,t')$ is typically written to give the response $\vec{P}(t)$ resulting from the "materials" part $\vec{P}(\omega)$ of $\varepsilon(\omega)$, we define it here to give the displacement $\Delta\vec{r}(t)$ so we can use the force operator directly. The $\vec{P}(t)$ version follows by multiplying the $G(t,t')$ that we obtain by nq. Accordingly, the equation that the Green function must satisfy is

$$\left(m\frac{d^2}{dt^2} + 2m\Gamma\frac{d}{dt} + \kappa \right)\Delta\vec{r}(t) = q\vec{E}(t) , \tag{8.71}$$

where the quantity in large parentheses is the force operator of the harmonic oscillator.

Recalling our experience with Poisson's Equation, we therefore look for a function $G(t,t')$ such that

$$\left(m\frac{d^2}{dt^2} + 2m\Gamma\frac{d}{dt} + \kappa \right) G(t,t') = \delta(t-t') . \tag{8.72}$$

When we solved Poisson's Equation, convenience suggested that the prefactor of the delta function should be (-4π). Here, the "convenience" suggests that the prefactor should be 1. Accordingly, $\Delta\vec{r}(t)$ is given by

$$\Delta\vec{r}(t) = \int_{-\infty}^{\infty} dt' G(t,t') q\vec{E}(t') , \tag{8.73}$$

where q is included with $\vec{E}(t')$ to be consistent with Eq. (8.71). This result is completely general, valid for any $\vec{E}(t)$. Therefore, it can describe transient responses as well as the steady-state solutions that we have been studying up to now.

To determine $G(t,t')$, we note that the delta function is mostly zero, so we construct it from solutions of the homogeneous equation. As noted in our treatment of electrostatics, $G(t,t')$ must be continuous at the singularity and have a discontinuous derivative there. To find these solutions consider the trial function $C_o e^{-i\omega t}$ and determine C_o and ω. We find that the homogeneous version of Eq. (8.72) is satisfied if

$$\omega = \omega_\pm = -i\Gamma \pm \sqrt{\omega_o^2 - 4\Gamma^2} = -i\Gamma \pm \omega_o ' . \tag{8.74}$$

These solutions can be written

$$\Delta r_\pm(t) = C_\pm e^{-\Gamma t} e^{\pm i\omega_o' t} . \tag{8.75}$$

Because they also set the denominator of Eq. (8.39) equal to zero,we identify these as they are the elementary excitations of the system.

Now, taking into account that $G(t,t) = 0$ for $t < t'$ and that it must be continuous at $t = t'$, the appropriate linear combination of the two solutions is

$$G(t,t') = C_o e^{-\Gamma t} \sin\left(\omega_o '(t-t')\right) u(t-t') . \tag{8.76}$$

where $u(t-t')$ is the unit-step function. To determine C_o, substitute Eq. (8.76) into Eq. (8.72) and integrate over a small interval δt about t'. The result is

$$\int_{t'-\delta t}^{t'+\delta t} dt \left(m\frac{d^2}{dt^2} + b\frac{d}{dt} + \kappa \right) G(t,t')$$

$$= m\frac{dG}{dt}\bigg|_{t'-\delta t}^{t'+\delta t} = m\frac{d}{dt}\left(C_o e^{-\Gamma t} \sin\omega_o '(t-t') \right)\bigg|_{t'+\delta t} . \tag{8.77a,b}$$

$$= m\omega_o ' C_o e^{-\Gamma t'} \tag{8.77c}$$

$$= \int_{t'-\delta t}^{t'+\delta t} dt\delta(t-t') = 1. \qquad (8.77d,e)$$

Therefore

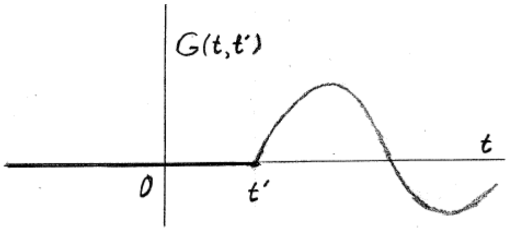
$$G(t,t') = \frac{1}{m\omega_o'} e^{-\Gamma(t-t')} \sin\left(\omega_o'(t-t')\right) u(t-t') \qquad (8.78)$$

for the damped harmonic oscillator. As a cross-check, the inverse Fourier transform of $G(t,t')$ is

$$\frac{1}{2\pi}\int_{-\infty}^{\infty} dtG(t,t')e^{i\omega t} = \frac{1}{2\pi m}\frac{1}{\omega_o^2 - \omega^2 - 2i\omega\Gamma}$$

$$= \frac{\varepsilon(\omega)-1}{2\pi m\omega_p^2}. \qquad (8.79)$$

A schematic of $G(t,t')$ is shown in the figure. This is exactly what we would expect for the response $\Delta\vec{r}(t)$ if an initially stationary block of mass *m* connected by a spring to a wall were hit by a hammer at $t = t'$. Its position would be continuous but its speed would be discontinuous. Once set in motion it oscillates at a frequency ω_o' with the amplitude dying out with a time constant $\tau = 1/\Gamma$. Accordingly, Eq. (8.73) can be viewed as the result of the ultimate superposition of infinitesimal hammers delivering forces of amplitudes $q\vec{E}(t')$ at times t'. Applications of Eq. (8.73) are left as homework assignments.



Because Eq. (8.73) is completely general, it describes transients as well as the steady-state solutions that we have considered so far. The difference can be appreciated by considering a frictionless spring/mass combination of resonance frequency ω_o driven by a field $\vec{E}e^{-i\omega t}$. The steady-state solution is

$$(\omega^2 - \omega_o^2)\Delta\vec{r}e^{-i\omega t} = \frac{q}{m}\vec{E}e^{-i\omega t}. \qquad (8.80)$$

This creates an obvious problem if $\omega = \omega_o$. Taken at face value, it states that either the amplitude of $\Delta\vec{r}$ is infinite, or it is impossible to apply an electric field to an object at its resonance frequency. The dilemma spotlights a generally unrecognized built-in difficulty with steady-state solutions, but one that is easily resolved by considering transients. The resolution of this dilemma is left as a homework assignment.

Our goal remains to determine the constraint that causality, $G(t,t') = 0$ for $t < t'$, places on $\varepsilon(\omega)$. To do this we must introduce the frequency domain. This can be done by writing $\Delta\vec{r}(t)$ as a Fourier transform

$$\Delta\vec{r}(t) = \int_{-\infty}^{\infty} d\omega \Delta\vec{r}(\omega) e^{-i\omega t} \, , \tag{8.81}$$

using the physics convention $e^{-i\omega t}$ for consistency. Now $\varepsilon(\omega)$ and $\Delta\vec{r}(\omega)$ are connected through

$$\vec{D} = \varepsilon\vec{E} = \vec{E} + 4\pi\vec{P} = \vec{E} + 4\pi nq\Delta\vec{r} \, , \tag{8.82}$$

so

$$\Delta\vec{r}(\omega) = \frac{\varepsilon(\omega)-1}{4\pi nq}\vec{E}(\omega) \, . \tag{8.83}$$

Therefore

$$\Delta r(t) = \int_{-\infty}^{\infty} d\omega \frac{\varepsilon(\omega)-1}{4\pi nq}\vec{E}(\omega) e^{-i\omega t} \, . \tag{8.84}$$

Now $\vec{E}(\omega)$ can also be written as a Fourier transform. The result is

$$\Delta r(t) = \int_{-\infty}^{\infty} d\omega \frac{\varepsilon(\omega)-1}{4\pi nq^2} e^{-i\omega t} \left(\frac{1}{2\pi} \int_{-\infty}^{\infty} dt' q\vec{E}(t') e^{i\omega t'} \right) \tag{8.85a}$$

$$= \int_{-\infty}^{\infty} dt' q\vec{E}(t') \left(\int_{-\infty}^{\infty} d\omega \frac{\varepsilon(\omega)-1}{8\pi^2 nq^2} e^{-i\omega(t-t')} \right), \tag{8.85b}$$

where in Eq. (8.85b) we have gathered terms and the order of integration has been reversed. Comparing Eq. (8.85b) to Eq. (8.73) we see that the Green function for $\Delta\vec{r}(t)$ is

$$G(t,t') = \frac{1}{8\pi^2 nq^2} \int_{-\infty}^{\infty} d\omega \left(\varepsilon(\omega)-1\right) e^{-i(t-t')\omega} . \tag{8.86}$$

While the Green function that we have been using explicitly up to now is that of the harmonic oscillator, the derivation of Eq. (8.86) is independent of any particular model. It assumes only linearity. Hence it is completely general, establishing the connection between $G(t,t')$ and $\varepsilon(\omega)$ independent of its origin or its spectral dependence.

The validity of Eq. (8.86) for the harmonic oscillator is easily verified by substituting it in Eq. (8.72), interchanging the order of differentiation and integration, and using the identity

$$\delta(t-t') = \int_{-\infty}^{\infty} e^{-i\omega(t-t')} d\omega$$

To obtain the constraint on $\varepsilon(\omega)$ imposed by Eq. (8.86) it is necessary to do more mathematics. This is accomplished in the next section.

H. The Cauchy Integral Formula: reality and causality.

To proceed further we must introduce the Cauchy integral theorem. An understanding of this theorem leads to a better understanding of linearity, reality, causality, and the physical meaning of poles in the complex plane. It also leads to Kramers-Kronig relations and sum rules. The conclusions reached here are completely general, applying well beyond the context of E&M. Accordingly, this section is unapologetically mathematical, because the language needs to be established before the physics is extracted.

H1. Complex variables, the Cauchy Integral Formula, and residues.

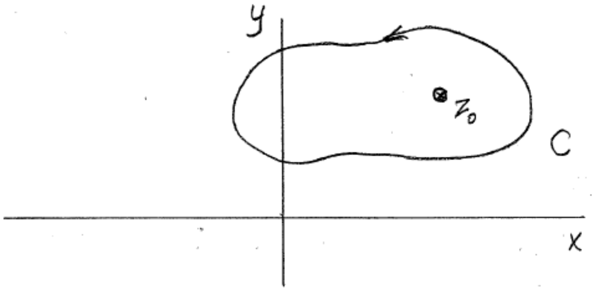
The theory of complex variables could occupy an entire course, but we need only to become familiar with the Cauchy Integral Formula and its generalization, Cauchy's Residue Theorem. The development treats ω as a complex variable, but this should cause no difficulty since we have done this before. Contour (line) integrals are also an intrinsic part of the theory, but these have been used as well.

Like Gauss' Theorem, Cauchy's Integral Formula is one of the miracles of mathematics. It can be summarized as follows. Let *C* be a closed curve in the complex $z = x + iy$ plane, and let $f(z)$ be a holomorphic (differentiable, to us physicists) function of *z* with no poles in the region enclosed by *C*. Then

$$\frac{1}{2\pi i}\oint_C \frac{f(z)}{z - z_o} dz = f(z_o) \text{ if } z_o \text{ is within C;} \tag{8.87a}$$

$$= 0 \text{ otherwise.} \tag{8.87b}$$

A schematic is shown in the figure at the right. The path *C* is traversed in the *counterclockwise* direction (in cylindrical coordinates, think increasing φ), using the right-hand rule (think $\hat{z}$ pointing upward toward you). Because this is a *line* integral over the path *C*, not an integral over the area enclosed by *C*, if the path is followed in the clockwise direction, then the sign of the integral is reversed. Nothing is said about *C* other than that it is a closed loop, and that the integration proceeds in the counterclockwise direction when viewed from above. If $f(z)$ itself contains poles $z = z_\nu$, then $(z - z_o)^{-1}$ is considered part of $f(z)$ and Eq. (8.87a) written as

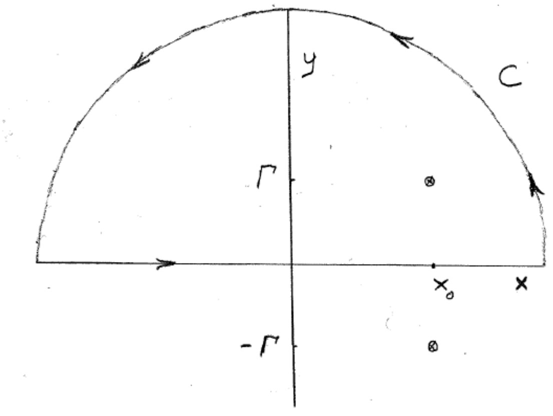


$$\oint_C dz\, f(z) = 2\pi i \sum_{\nu=1}^{N} f(z)(z - z_\nu)\Big|_{z_\nu} . \tag{8.88}$$

The terms in the sum on the right are termed *residues*.

In the theory of complex variables “holomorphic” is equivalent to “analytic”, where “analytic” is a broader term that is used to describe any function that can by represented as a convergent power series (Taylor series) in a neighborhood of any point in its domain. While it appears reasonable that “holomorphic” and “analytic” are equivalent, this result is difficult to prove. Fortunately, we do not have to prove it to use it. “Domain” refers to any connected open subset of a space.

This is as far as we need to go in developing the theory of complex variables. By now you are probably wondering why even take complex variables this far? The answer is that there are many instances in physics where Fourier integrals of the type $f(t) = \int_{-\infty}^{\infty} d\omega\, f(\omega) e^{-i\omega t}$ and its inverse $f(\omega) = \dfrac{1}{2\pi}\int_{-\infty}^{\infty} dt'\, f(t') e^{i\omega t'}$ are encountered. Equation (8.88) provides an easy way to evaluate them. Choose a contour *C* that consists of the entire real axis, which is the integral that we want, and a semicircular loop of infinite radius for either positive or negative $\mathrm{Im}(\omega)$, which we hope contributes nothing but must be checked in any case. The residues of $f(\omega)$ are then calculated for its poles within *C*. Finally, the residues are summed and the result multiplied by $2\pi i$. This completes the evaluation of the integral.



As an example, consider

$$I = \int_{-\infty}^{\infty} dx \frac{1}{(x - x_o)^2 + \Gamma^2}\,. \qquad (8.89)$$

The integrand has poles at $z = x_o \pm i\Gamma$, as indicated in the figure above. Let *C* consist of the real axis and the semicircle $Re^{i\theta}$, where $R \to \infty$ and $0 \le \theta \le \pi$. The line integral over the semicircle vanishes, since

$$I_\theta = \lim_{R\to\infty} \int_0^{\pi} id\theta\, Re^{i\theta} \frac{1}{(Re^{i\theta} - x_o)^2 + \Gamma^2} = 0\,. \qquad (8.90)$$

Thus evaluating Eq. (8.89) at the pole $z = x_o + i\Gamma$ gives the value of the integral. Explicitly:

$$I = \int_{-\infty}^{\infty} dx \frac{1}{(x - x_o)^2 + \Gamma^2} = 2\pi i \frac{1}{(x_o + i\Gamma) - x_o + i\Gamma}\,.$$

$$= \frac{\pi}{\Gamma}, \qquad (8.91)$$

which agrees with the value obtained from a table of integrals.

The integral can also be evaluated by closing the contour in the lower half plane. The sign of the residue changes, but this change is cancelled either because the loop is traversed in the “wrong” direction, or if the “right” direction is used, the integral in Eq. (8.98) is done backwards, i.e., from $+\infty$ to $-\infty$. The net result is the same.

Returning to causality, if the result of the integration in Eq. (8.86) is truly a response function, then it must be rigorously zero for $t < t'$, that is, before the drive is applied. We now use Cauchy’s Theorem to prove that this condition is satisfied. Let *C* consist of the real axis and a semicircle of radius *R* receding to infinity. To determine which semicircle to use, let $t'-t > 0$, with $\omega = Re^{i\theta} = R(\cos\theta + i\sin\theta)$. With $t'-t > 0$ we must obviously use the upper semicircle, where $0 \le \theta \le \pi$. With $R \to \infty$ its contribution to the contour integral vanishes. Regarding the residue term, we showed earlier that causality requires all poles to lie in the lower half plane. Since there are no poles in the upper half plane, the residue contribution is zero. Hence for $t < t'$, $G(t,t') = 0$ for $t < t'$. This is consistent with causality. Again, the result is completely general.

We now check the specific case of Eq. (8.86) for $t - t' > 0$. In this case the contour must be closed in the lower half plane. The infinite semicircle again makes no contribution, so the result is the Fourier transform that we want. A small collection of minus signs are involved, so intermediate steps are provided. In addition, this is good practice in evaluating integrals by residues. The residues of Eq. (8.86) are

$$\text{at } \omega = \omega_o - i\Gamma: \quad \frac{-\omega_p^2}{(4\pi)(2\omega_o)} e^{-\Gamma(t-t')} e^{-i\omega_o(t-t')}; \qquad (8.92a)$$

$$\text{at } \omega = -\omega_o - i\Gamma: \quad \frac{-\omega_p^2}{(4\pi)(-2\omega_o)} e^{-\Gamma(t-t')} e^{i\omega_o(t-t')}. \qquad (8.92b)$$

Adding the two contributions, multiplying by $2\pi i$, and accounting for the $(1/2\pi)$ prefactor on the integral gives

$$\frac{2\pi i}{2\pi} e^{-\Gamma(t-t')} \frac{\omega_p^2}{4\pi} \frac{2i}{2\omega_o} \left(\frac{-e^{-i\omega_o(t-t')} + e^{-i\omega_o(t-t')}}{2i} \right)$$

$$= -\frac{\omega_p^2}{4\pi\omega_o} e^{-\Gamma(t-t')} \sin\omega_o(t-t'). \qquad (8.93)$$

To obtain $G(t,t')$, the sign must be changed because the contour was followed in the “wrong” direction. This result is limited to $t > t'$. Combining the two solutions

$$G(t,t') = \frac{\omega_p^2}{4\pi\omega_o} e^{-\Gamma(t-t')} \sin\omega_o(t-t')\, u(t-t'). \qquad (8.94a)$$

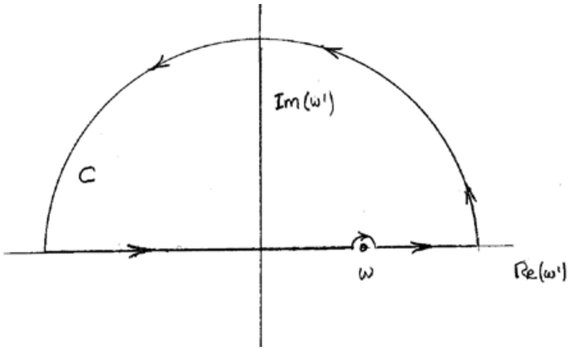
$$= \frac{nq^2}{m\omega_o} e^{-\Gamma(t-t')} \sin\omega_o(t-t')\, u(t-t') \qquad (8.94b)$$

The prefactor $1/4\pi$ follows because $G(t-t')$ is the Green function for $\vec{P}(t)=nq\Delta\vec{r}$, not $\vec{D}(t)$. By replacing ω_p^2 with its constituent parameters, we convert $G(t,t')$ to a form where it can be compared to the original. The results agree; the loop has been closed.

I. Kramers-Kronig relations and sum rules.

Kramers-Kronig (KK) transforms are important because they relate the real part ε_1 of ε to the imaginary part ε_2 and vice versa through integrals that are essentially weighted averages. Thus ε_1 and ε_2 are not independent, with implications discussed at the end of the section. These integrals connecting the real and imaginary parts are known as the KK transforms.

KK transforms are special cases of a more general result, the Sokhotskii-Plemelj Theorem. A related special case is the Hilbert transform. Interestingly, the entire history is one of discovery and rediscovery. The original form was obtained by Sokhotskii in 1868. In 1905 Hilbert developed a related expression, the Hilbert transform, as a byproduct of a solution to a problem originally posed by Riemann. Plemelj presented his own version of the proof of Riemann's conjecture in 1908, extending Sokhotskii's earlier result. The Kramers-Kronig transform, yet another special case, was discovered by Kronig and 1926 and rediscovered by Kramers in 1927.



The KK transform is

$$\varepsilon(\omega)-1=-\frac{i}{\pi}P\int_{-\infty}^{\infty}\frac{d\omega'}{\omega'-\omega}\left(\varepsilon(\omega')-1\right). \tag{8.95}$$

Here, *P* represents the principal-part operation, which is defined below. The *i* in the prefactor shows that ε_1 and ε_2 are not independent but are connected by integral relations. This is a direct consequence of causality, as also described below. Additional consequences are sum rules, which are described in the following section.

The relation is derived by evaluating

$$I(\omega)=\oint_C\frac{d\omega'}{\omega'-\omega}\left(\varepsilon(\omega')-1\right) \tag{8.96}$$

over the contour shown above. *C* covers the entire real axis except for the point ω, where *C* is deformed by a small semicircle of radius $\delta\omega$ to exclude it. The associated contour integral therefore has four segments, specifically

$$\oint_C \frac{d\omega'}{\omega'-\omega}\left(\varepsilon(\omega')-1\right)=$$

$$\int_{-\infty}^{\omega-\delta\omega} \frac{d\omega'}{\omega'-\omega}\left(\varepsilon(\omega')-1\right)+\int_{\pi}^{0} \frac{id\theta\ \ \delta\omega\, e^{i\theta}}{(\omega+\delta\omega\, e^{i\theta})-\omega}\left(\varepsilon(\omega+\delta\omega\, e^{i\theta})-1\right)$$

$$+\int_{\omega+\delta\omega}^{\infty} \frac{d\omega'}{\omega'-\omega}\left(\varepsilon(\omega')-1\right)+\lim\nolimits_{R\to\infty}\int_{0}^{\pi} \frac{id\theta\ \ Re^{i\theta}}{Re^{i\theta}-\omega}\left(\varepsilon(\omega')-1\right)=0\,. \qquad (8.97)$$

$I(\omega)$ vanishes because of causality: no poles exist in the upper half of the complex ω' plane. For $\delta\omega \to 0$ the second integral on the right reduces to

$$\lim\nolimits_{\delta\omega\to 0}\int_{\pi}^{0} \frac{id\theta\ \ \delta\omega e^{i\theta}}{(\omega+\delta\omega e^{i\theta})-\omega}\left(\varepsilon(\omega+\delta\omega e^{i\theta})-1\right)=-i\pi\left(\varepsilon(\omega)-1\right). \qquad (8.98)$$

Because $\lim_{\omega\to\infty}\left(\varepsilon(\omega)\right)=1-\omega_p^2/\omega^2$, the fourth integral vanishes in the $R\to\infty$ limit as well. The remaining two integrals can be combined as

$$\int_{-\infty}^{\omega_o-\delta} \frac{d\omega'}{\omega'-\omega}[\varepsilon(\omega')-1]+\int_{\omega_o+\delta}^{\infty} \frac{d\omega'}{\omega'-\omega}[\varepsilon(\omega')-1]=P\int_{-\infty}^{\infty} \frac{d\omega'}{\omega'-\omega}[\varepsilon(\omega')-1]\,. \qquad (8.99)$$

The result is Eq. (8.96).

While most of *C* is routine, the principal-part operation requires attention. To evaluate this contribution, on the real axis, terminate the left and right line integrals at points ω_o and ω_1, then consider the segments from ω_o to $\omega-\delta\omega$ and $\omega+\delta\omega$ to ω_1, respectively, where $(\omega-\omega_0)=(\omega_1-\omega)$ so everything is symmetric. Then for these two segments, the principal-part operation is defined as

$$P\int_{\omega_o}^{\omega_1} \frac{d\omega'}{\omega'-\omega}f(\omega')=\lim\nolimits_{\delta\omega\to 0}\left(\int_{\omega_0}^{\omega-\delta\omega} \frac{d\omega'}{\omega'-\omega}f(\omega')+\int_{\omega+\delta\omega}^{\omega_1} \frac{d\omega'}{\omega'-\omega}f(\omega')\right). \qquad (8.100)$$

Next, expand

$$f(\omega')=f(\omega)+(\omega'-\omega)\left.\frac{df}{d\omega'}\right|_{\omega'=\omega}, \qquad (8.101)$$

which is sufficiently accurate if the total extent $(\omega_1-\omega_0)$ is small. Substituting, we find

$$P\int_{\omega_0}^{\omega_1} \frac{d\omega'}{\omega'-\omega}f(\omega')=\lim\nolimits_{\delta\omega\to 0}\left(f(\omega)\ln\left(\frac{-\delta\omega}{\omega_0-\omega}\frac{\omega_1-\omega}{\delta\omega}\right)\right)+(\omega_1-\omega_0)\left.\frac{df(\omega')}{d\omega'}\right|_{\omega'=\omega}$$

$$=(\omega_1-\omega_0)\left.\frac{df(\omega')}{d\omega'}\right|_{\omega'=\omega}. \qquad (8.102)$$

When the remainder of the path *C* is added to Eq. (8.102), the integral is evaluated.

By separating Eq. (8.95) into its real and imaginary parts, $\varepsilon_1(\omega)$ is seen to be a weighted average of $\varepsilon_2(\omega)$ and *vice versa*. However, we can proceed further. Noting that $\varepsilon(\omega) = \varepsilon^*(-\omega)$, Eq. (8.95) can be rewritten in terms of positive frequencies only as

$$\varepsilon_1(\omega) = 1 + \frac{2}{\pi} P \int_0^\infty \frac{\omega' d\omega'}{\omega'^2 - \omega^2} \varepsilon_2(\omega') \, ; \tag{8.103a}$$

$$\varepsilon_2(\omega) = -\frac{2\omega}{\pi} P \int_0^\infty \frac{d\omega'}{\omega'^2 - \omega^2} [\varepsilon_1(\omega') - 1] \, . \tag{8.103b}$$

These forms have the additional advantage that the image poles that are always neglected in local expansions are incorporated automatically, so the results are necessarily more accurate than the results of KK integrals applied to single-pole expansions. This is particularly important in the limit of large ω, where ignoring image poles causes $(\varepsilon(\omega) - 1)$ to decrease as ω^{-1} instead of ω^{-2}.

With the heavy math out of the way, we investigate some very practical consequences. For example we might want a large refractive index $n = \sqrt{\varepsilon}$ in a material that is transparent in the visible part of the optical spectrum. This is the condition required for white paint, which consists of transparent inclusions in a transparent organic binder. The paint is white for the same reason that clouds are white: the particles doing the scattering are significantly larger than the wavelength of light so any radiation entering the paint gets scattered, with all of it eventually finding its way out. The particles as well as the binder must be transparent, otherwise only a filtered (colored) version of the light will emerge. Since the binder in dried paint is a transparent organic material with a refractive index $n \approx 1.5$ in the visible spectral range, the scattering inclusions must have as large a refractive index as possible to maximize the refractive-index difference to promote scattering. Although it is transparent and cheap, SiO_2 would fail miserably because it has a refractive index of approximately 1.46 in the visible range (see a previous figure), and hence would be an optical match to the binder. The paint would be cloudy of not transparent.

Examining Eq. (8.103a), the largest $\varepsilon_1 = n^2$ consistent with transparency in the visible range occurs for a material that strongly absorbs light just outside the high end of the visible range, that is, at a wavelength as close as possible to, but energetically slightly above, 400 nm. Several materials have this property, among them “white lead” ($2PbCO_3{\cdot}Pb(OH)_2$) and titania (titanium dioxide, TiO_2). For many years white lead was the material of choice because it was cheap. However, lead is toxic, so “white-lead” paints were banned in the US in 1977. White paints are now based on titania, which has an absorption edge 3.24 eV = 383 nm, just outside our visible spectral range.

Paper is another example of the use of a dielectric mismatch to promote scattering and generate a white appearance. Here the “inclusions” are cellulose fibers with $n \approx 1.5$, and the “binder” is air, with $n = 1$. The large difference in refractive indices leads to efficient scattering, so paper works. Its scattering effectiveness can be reduced quite easily by wetting it with water ($n = 1.33$) or particularly oil ($n \sim 1.5$).

Until about 30 years ago measurements of reflectance spectra *R* and application of the K-K relations were the only way of obtaining $\varepsilon(\omega)$ in the visible range. Spectroscopists would measure *R* over as wide a range as possible, calculate $\sqrt{R(\omega)} = r(\omega)$, take the K-K transform of $r(\omega)$ to generate the phase angle $\theta(\omega)$, then calculate $\varepsilon(\omega)$ using the complex reflectance $\tilde{r}(\omega) = r(\omega)e^{i\theta(\omega)}$ in the two-phase model that will be discussed in a later chapter. This was not satisfactory for several reasons: data could not be obtained over sufficiently wide spectral ranges for transforms to be reliable, and reflectance is very difficult to measure accurately. In addition, it was basically impossible to assess the quality of the surface being measured and thereby to eliminate artifacts from inadvertent oxide or adsorbate layers. Spectroscopic ellipsometry (SE), which can measure complex reflectance ratios directly, put an end to determining dielectric functions from reflectance spectra.

Nevertheless, reflectance is still the approach of choice for monitoring the deposition of optical thin films, mainly because of its simplicity. Optical films are usually thick enough to show interference phenomena. These applications reduce to measuring differences in film thicknesses, so high precision is required, but not high accuracy. While better measurement methods exist, sometimes a simpler approach is good enough.

While the above is written in terms $\varepsilon(\omega)$, the real and imaginary parts of the complex refractive index

$$\tilde{n}(\omega) = \sqrt{\varepsilon(\omega)} = n + i\kappa \tag{8.104}$$

also satisfy a KK relation, although between $(n-1)$ and κ, not between *n* and κ. This follows because the poles of $\tilde{n}$ are the same as those of ε, so the integral of $(\tilde{n}-1)$ over *C* is also equal to zero. As with ε, subtracting 1 from $\tilde{n}$ is necessary to ensure that the part over the infinite semicircle vanishes. As an aside, the closed-form expression for n and κ in terms of ε_1 and ε_2 is

$$n + i\kappa = \sqrt{\frac{1}{2}\left(\varepsilon_1 + \sqrt{\varepsilon_1^2 + \varepsilon_2^2}\right)} + i\frac{\varepsilon_2}{2n}. \tag{8.105}$$

This follows immediately from the quadratic equation for *n* that results by expanding $\varepsilon_1 + i\varepsilon_2 = (n + i\kappa)^2$.

We mentioned the Hilbert transform mentioned above. Here

$$h(\omega) = \frac{1}{\pi}\int_{-\infty}^{\infty} \frac{d\omega' f(\omega')}{\omega' - \omega} = \frac{1}{\pi}\int_{-\infty}^{\infty} \frac{d\omega' e^{i\omega' t}}{\omega' - \omega}. \tag{8.106}$$

In this case the principal-part contribution to $h(\omega)$ vanishes completely in the limit ω_o, $\omega_1 \to 0$. Moreover, the integration can be done in closed form. To do this let $\omega'' = \omega' - \omega$, in which case Eq. (8.109) reduces to

$$h(\omega) = \frac{1}{\pi} \int_{-\infty}^{\infty} \frac{d\omega''}{\omega''} e^{i(\omega''+\omega)t} = \frac{e^{i\omega t}}{\pi} \int_{-\infty}^{\infty} \frac{d\omega''}{\omega''} e^{i\omega'' t}$$

$$= \frac{e^{i\omega t}}{\pi} \int_{-\infty}^{\infty} \frac{d\omega''}{\omega''} \left(\cos \omega'' t + i \sin \omega'' t \right). \qquad (8.107)$$

The first term vanishes by symmetry. The second term can be written

$$h(\omega) = \frac{e^{i\omega t}}{\pi} \int_{-\infty}^{\infty} \frac{d(\omega'' t)}{\omega'' t} i \sin(\omega'' t) = \frac{i e^{i\omega t}}{\pi} \int_{-\infty}^{\infty} dx \,\mathrm{sinc}(x)$$

$$= i e^{i\omega t} \,\mathrm{sgn}(t). \qquad (8.108)$$

Thus given the real part of a harmonic function or vice versa, the Hilbert transform delivers the complementary part. As can be guessed, the Hilbert transform is particularly useful in communications.

Sum rules. More generally, the K-K relations lead to sum rules that provide additional cross checks on data and more information about the physics and chemistry of materials. The first, obtained from Eq. (8.107a), gives the static dielectric constant of insulators as

$$\varepsilon(0) = 1 + \frac{2}{\pi} \int_{0}^{\infty} \frac{d\omega'}{\omega'} \varepsilon_2(\omega'). \qquad (8.109)$$

This further emphasizes the point made above, that the largest refractive indices are obtained when absorption edges are as low in energy as possible consistent with the application.

The second sum rule is more interesting. It shows that it is impossible to increase $\varepsilon_2(\omega)$, and therefore optical absorption, without limit. Taking the limit $\omega \to \infty$ of Eq. (8.109a) and using Eq. (8.50), we find

$$\varepsilon_1(\omega) \approx 1 - \frac{2}{\pi \omega^2} \int_{0}^{\infty} \omega' d\omega' \varepsilon_2(\omega') = 1 - \frac{\omega_p^2}{\omega^2} = 1 - \frac{4\pi n e^2}{m \omega^2}. \qquad (8.110)$$

Cancelling the common factor ω^2 leads to the *plasma sum rule*:

$$\int_{0}^{\infty} \omega' d\omega' \varepsilon_2(\omega') = \frac{\pi \omega_p^2}{2} = \frac{2\pi^2 n q^2}{m}. \qquad (8.111)$$

This puts severe upper limits on possible values of $\varepsilon_2(\omega)$, especially for high energies as encountered for example in core-level transitions in synchrotron radiation spectroscopy.

Partial integration of the plasma sum rule leads to an additional useful result. If the upper limit of the integral is stopped at ω,

$$n(\omega) = \frac{m}{2\pi^2 q^2} \int_{0}^{\omega} \omega' d\omega' \varepsilon_2(\omega'). \qquad (8.112)$$

For a given material, the right-hand side cannot exceed the known number density of a given class of charge. Thus for example if one integrates the contribution to ε_2 of the bonding-to-antibonding valence-to-conduction band transitions of Si, which start about 1 eV and end at about 20 eV, *n* should be close to the known valence-charge density of 4 electrons/atom. For the Si 2p core-level transitions near 100 eV, the scaling factor ω' in the integrand restricts ε_2 to be approximately 100× lower than what it would be in the visible-near ultraviolet region. For the two Si k_α electrons at 1739 eV, ε_2 is lower by another order of magnitude.

J. Localization effects: the Clausius-Mossotti relation.

Up to now we have assumed that the field $\vec{E}_o$ used in the force calculations is the same as the macroscopic field $<\vec{E}>$ used in Eq. (8.33). In fact the field at an induced dipole $\vec{p}$ can include contributions from the fields generated by nearby induced dipoles, so this is not always the case. A second issue regarding the physics is that the volume average of the field of a dipole is not zero, but $(-4\pi/3)$ times the dipole. In relating $\vec{E}_o$ to $<\vec{E}>$ both local-field and volume-averaging effects must be taken into account. In the limit of point dipoles and in the absence of dipole-dipole interactions, the result is the classical Clausius-Mossotti Equation,

$$\varepsilon = 1 + \frac{4\pi n\alpha}{1 - \frac{4\pi n\alpha}{3}}, \tag{8.113}$$

where n is the dipole density and α is the polarizability defining the induced dipole as $\vec{p} = \alpha\vec{E}_{loc}$, where $\vec{E}_{loc}$ is the field at the dipole.

The Clausius-Mossotti Equation is historically significant because it was the first expression to connect atomic-scale parameters, α and *n*, to a macroscopic quantity, ε. It was derived by Mossotti in 1850 in an acoustics context and rediscovered independently by Clausius in 1879 for electrostatics. It appears routinely in textbooks, where it is usually the only exposure that most physicists get in relating macroscopic observables to atomic-scale parameters.

Unfortunately, all textbook derivations of which I am aware derive Eq. (8.113) using a nonphysical model, where local fields are created by fictitious charges at the boundary of a fictitious spherical cavity. In the final step of the classical derivation, the fields generated by the dipoles near a "test" dipole are shown to cancel identically. This is surprising, because one would expect that the largest contribution to $\vec{E}_{loc}$ should come from the dipoles in the vicinity of the dipole. While this is true for cubic materials in three dimensions, this cancellation does not occur for the two-dimensional materials of current interest. Hence it cannot be ignored. One of the reasons I became interested in electrodynamics as a grad student was that I realized that the classical cavity derivation could not possibly be correct, and therefore did not describe the physics involved. There had to be a derivation consistent with basic physical principles, and the challenge was to find it.

What follows is a derivation that does follow correct principles, although as is typical of electrostatics calculations, it has a weak point regarding the treatment of an infinity. Nevertheless, it is easier to deal with questions of this type than to justify fictitious charges on a fictitious cavity. The physics is that of positive feedback: we will find that the local field that appears at the charge sites is larger than the macroscopic field that we apply, so the system effectively exhibits gain. In the present context this is a consequence of averaging. In fact, models exhibiting the equivalent of gain are fairly common in physics, for example Toyozawa's contact-exciton description of the effect of the electron-hole Coulomb interaction on dielectric properties of semiconductors, Fano's description of the interaction of a quantum absorber in a weakly absorbing background, various plasmonic expressions, and the Dyson equation of field theory, among others. All exhibit positive-feedback characteristics and fall in this category, although this is not generally noted.

Start by supposing a cubic lattice of lattice constant *a*, with a point of polarizability α at each lattice site. Assume that a constant field $\vec{E}_o$ (the field of Eq. (8.33)) is applied to the crystal. When the fields resulting from $\vec{E}_o$ and the fields generated by all the dipoles are added together, the result is a dipole $\vec{p} = \alpha\,\vec{E}_{loc}$ at each site, where $\vec{E}_{loc}$ is the local field at that site. If the direction of $\vec{E}_o$ is arbitrary, we can take components along the cubic axes, solve the problem for each component, and superpose the results. However, since the lattice is cubic, the solution in each direction will be identical, and by extension, the field at each dipole site will be $\vec{E}_{loc} \parallel \vec{E}_o$. Hence it is necessary to solve the problem only for a single component, which we take to be along *z*.

Next, calculate the average dipole density $<\vec{P}>$. With a dipole at each lattice site, the microscopic dipole density of the material is

$$\vec{p}(\vec{r}) = \sum_{\vec{R}_n} \alpha\vec{E}_{loc}\delta(\vec{r}-\vec{R}_n), \tag{8.114}$$

where $\vec{R}_n$ is a lattice vector. By Eq. (7.27c),

$$<\vec{P}> = \int_V d^3r'W(\vec{r}-\vec{r}\,')\sum_{\vec{R}_n}\alpha\vec{E}_{loc}\delta(\vec{r}\,'-\vec{R}_n) \;= \sum_{\vec{R}_n}\alpha\vec{E}_{loc}W(\vec{r}-\vec{R}_n). \tag{8.115}$$

Converting the sum into an integral yields

$$<\vec{P}> = \sum_{\vec{R}_n}\alpha\vec{E}_{loc}W(\vec{r}-\vec{R}_n) \;= \int_V d^3r'\,n\,\alpha\vec{E}_{loc}W(\vec{r}-\vec{r}\,') = n\,\alpha\vec{E}_{loc}. \tag{8.116}$$

The next two problems are related. These are (1) to determine the local field at each dipole, and (2) to calculate the average field $<\vec{E}>$. The calculation begins by determining the complete microscopic field. This consists of the applied field $\vec{E}_o$ plus that of all the dipoles. The field due to a dipole at the origin is

$$\vec{E}_{dip} = -\nabla\left(\frac{\vec{p}\cdot\vec{r}}{r^3}\right) = \frac{3(\vec{r}\cdot\vec{p})\vec{r} - r^2\vec{p}}{r^5}, \qquad (8.117a,b)$$

where $\vec{r}$ is the location of the observer. Since the dipoles are all located at lattice sites $\vec{R}_n$, and all dipoles are identical, it follows that the field at any point inside the material is given by

$$\vec{E}(\vec{r}) = \vec{E}_o + \sum_{\vec{R}_n}{}' \frac{3((\vec{r}-\vec{R}_n)\cdot\vec{p})(\vec{r}-\vec{R}_n) - (\vec{r}-\vec{R}_n)^2\vec{p}}{|\vec{r}-\vec{R}_n|^5}, \qquad (8.118)$$

where the prime on the summation means that we exclude the self-field at $\vec{R}_n = 0$.

Now if this expression is general, it must be valid at any lattice site. Consequently, we can evaluate it at the origin to find the field there, which is by definition $\vec{E}_{loc}$. The result is

$$\vec{E}_{loc} = \vec{E}(0) = \vec{E}_o + \sum_{\vec{R}_n}{}'\alpha\frac{3(\vec{R}_n\cdot\vec{E}_{loc})\vec{R}_n - (\vec{R}_n)^2\vec{E}_{loc}}{|\vec{R}_n|^5}. \qquad (8.119)$$

This is a standard self-consistency expression.

The evaluation of Eq. (8.119) for a simple cubic lattice is straightforward. Starting with the 6 nearest neighbors and with $\vec{E}_o \parallel \hat{z}$, the contribution of the two dipoles along the *z* axis is exactly that of the 4 dipoles in the plane perpendicular to this axis, but of opposite sign. Hence the first shell contributes nothing. The same is found for all succeeding shells. The surprising result is that $\vec{E}_{loc} = \vec{E}_o$. As noted above, if a local-field effect exists, we should expect to be largest for the nearest dipoles, but this is not realized in practice. However, this cancellation does not occur for two-dimensional materials, or for dipoles near a surface or interface. Here the dipole contribution must be taken into account.

It is still necessary to relate $\vec{E}_o$ to $<\vec{E}>$. The calculation is

$$<\vec{E}> = \int_V d^3r' W(\vec{r}-\vec{r}')[\vec{E}_o + \sum_{\vec{R}_n}\frac{3((\vec{r}'-\vec{R}_n)\cdot\vec{p})(\vec{r}'-\vec{R}_n) - (\vec{r}'-\vec{R}_n)^2\vec{p}}{|\vec{r}'-\vec{R}_n|^5}]. \qquad (8.120)$$

This apparently formidable calculation can be brought to a successful conclusion by noting first that, by the properties of *W*, the average of $\vec{E}_o$ is $\vec{E}_o$, as expected. To assess the value of the second term, change variables to $\vec{r}'' = \vec{r}' - \vec{R}_n$, then interchange the order of summation and integration. The result is

$$<\vec{E}> = \vec{E}_o + \sum_{\vec{R}_n}\int_V d^3r'' W(\vec{r}-\vec{r}''+\vec{R}_n)\frac{3(\vec{r}''\cdot\vec{p})\vec{r}'' - (\vec{r}'')^2\vec{p}}{(\vec{r}'')^5}. \qquad (8.121)$$

In changing the sum to an integral, the standard dipole density factor *n* is introduced. The resulting integral over *W* can be done explicitly, since this particular dummy variable does not appear in the dipole-field term. The result is

$$< \vec{E} > = \vec{E}_o - n\int_V d^3r''\nabla_{\vec{r}''}\left(\frac{\vec{r}''\cdot\vec{p}}{(r'')^3}\right), \tag{8.122}$$

where the dipole field is written as the negative gradient of the dipole potential.

To proceed further, use the gradient theorem from the inside front cover of Jackson to turn the volume integral into a surface integral. At the same time, simplify the notation by replacing the dummy integration variable $\vec{r}''$ with $\vec{r}$. The result is

$$< \vec{E} > = \vec{E}_o - n\int_S d^2r\,\hat{n}\left(\frac{\vec{r}\cdot\vec{p}}{r^3}\right). \tag{8.123}$$

Now $\hat{n} = \hat{x}\sin\theta\cos\phi + \hat{y}\sin\theta\sin\phi + \hat{z}\cos\theta$, $\hat{r} = r(\hat{x}\sin\theta\cos\phi + \hat{y}\sin\theta\sin\phi + \hat{z}\cos\theta)$, and $\vec{p} = p\hat{z}$. Making these substitutions gives

$$< \vec{E} > = \vec{E}_o - n\int_S r^2 d(\cos\theta)d\phi(\hat{x}\sin\theta\cos\phi + \hat{y}\sin\theta\sin\phi + \hat{z}\cos\theta)\left(\frac{p\cos\theta}{r^2}\right)$$

$$= \vec{E}_o - \frac{4\pi n}{3}\alpha\vec{E}_o \,. \tag{8.124}$$

This result is surprising in two ways. First, the internal field $\vec{E}_o$ at any dipole site is greater than the macroscopic field $< \vec{E} >$ that is applied. The reason is clear: the volume integral of the electric field due to a dipole $\vec{p}$ is not zero, but rather $(-4\pi/3)\vec{p}$. Thus the macroscopic external field is reduced relative to the internal field that created the dipoles. Second, as long as the sphere is big enough to contain the dipole (and any finite radius *a* will do), the result is independent of the radius of the sphere. This means that the contribution of the dipole to the average internal field is fully developed for arbitrarily small radii, and that increasing radii beyond this leads to no additional contribution. We will use this result in Ch. 8, where we derive the Maxwell Garnett effective-medium expression for the dielectric response of a composite material.

Given the connection between $\vec{E}_o$ and $< \vec{E} >$, we can now find *ε*. Clearly, the correct version of Eqs. (8.29) is

$$< \vec{D} > = \varepsilon < \vec{E} > = < \vec{E} > + 4\pi < \vec{P} >$$

$$= < \vec{E} > + \frac{4\pi n\alpha}{1 - \frac{4\pi n\alpha}{3}} < \vec{E} > \cdot \tag{8.125}$$

With field-averaging included, this gives

$$\varepsilon = 1 + \frac{4\pi n\alpha}{1 - \frac{4\pi n\alpha}{3}} \cdot \tag{8.126}$$

Equation (8.127) is usually written

$$\frac{4\pi n\alpha}{3} = \frac{\varepsilon - 1}{\varepsilon + 2}, \tag{8.127}$$

giving the atomic-scale parameters *n* and α in terms of the macroscopic quantity ε. Note that Eq. (8.127) used without local-field corrections yields the "bare" dielectric response

$$\varepsilon = 1 + 4\pi\, n\alpha\,, \tag{8.128}$$

which assumes that $\vec{E}_{loc} = <\vec{E}>$, as noted at the beginning of this section. Thus localization of the polarizable species has a significant effect.

The above uses several steps beyond the standard textbook derivation, but has the advantages that all steps follow basic physics and/or mathematics principles, and that the physics behind the C-M relation is clearly identified. The functional form Eq. (8.128) appears in many contexts, for example in the Toyozawa contact-exciton approximation of the Coulomb interaction between the electron and hole in an optical absorption process, and in the currently highly cited 1961 paper of Fano, where the interaction of a local oscillator with a background continuum yields an expression of the same form. Although Toyozawa attributed the resultant distortion of the "bare" $\varepsilon(\omega)$ spectrum to the attractive interaction between the electron and hole, the fact that excitons are localized entities means that localization may also contribute to the picture. This aspect of the dielectric response of causal systems has not been recognized.

A final comment regarding local fields: the exact cancellation of the sum over dipoles works only for an infinite material. If the lattice is terminated, as for example at a surface, cancellation is not exact and the dipoles there give an additional correction to the local field. This surface-local-field effect is needed to describe some otherwise puzzling aspects of the optical properties of semiconductors. It also shows that terminating an infinite bulk lattice must inevitably modify the dielectric properties of a material in the outer several atomic layers, not only for quantum-mechanical reasons.