21 COATING OF OPTICAL SURFACES

21.1 INTRODUCTION

21.1.1 Uses. Thin films of dielectric, metallic or even semi-conducting materials are most often applied to optical components such as lenses, plates and reflectors for the purpose of altering their energy reflectances or transmittances. A great variety of distributions of spectral reflectances and transmittances can be achieved over the ultraviolet, visible and infra-red regions. However, the number of materials having suitable optical, mechanical and chemical properties for use in the ultraviolet region is severely limited. Occasionally, thin films are used for modifying, especially at oblique incidence, phase changes as well as amplitude changes upon reflection. Thin films can also serve as protective coatings for surfaces of soft materials such as aluminum or silver. In another type of application, films are deposited with non-uniform thickness in order to achieve a slight degree of aspherization of the coated surface or in order to produce wedges that transmit non-uniformly in a specified manner. In still another broad class of films, a combination of optical properties such as transmittance, and of electrical properties such as conductance is provided. A diversity of specialized films consisting of combinations of two or more materials in a multilayer is required to meet an increasing list of modern applications.

21.1.2 Properties of thin films. We shall be concerned herein with the physical principles governing the optical properties of thin films. Fortunately, thin films have been found to behave to a good first approximation as homogeneous, plane-parallel layers that can be regarded as infinite in lateral dimensions. This idealized model of single films or multilayers can be analyzed without further approximation as a boundary problem involving Maxwell's equations. Several related forms of this useful theory will be treated. Actually, thin films are not homogeneous either laterally or along the thickness direction. Small departures from the predictions of the idealized model are therefore likely to occur. Theories dealing with inhomogeneity along the thickness direction are under active investigation; but it must be expected that these theories will be of greatest value in designing films whose inhomogeneities are increased deliberately.

21.2 DEFINITIONS AND PRINCIPLES

21.2.1 The optical constants. The optical constants, n and K, of an homogeneous, isotropic film are defined in the following manner. We take the solutions for the electric vector, E, and the magnetic vector, H, in the form

$$E = U e^{-i\omega t} \; ; \; H = V e^{-i\omega t} \; ; \tag{1}$$

in which U and V are vectors; $U = (U_x, U_y, U_z)$ and $V = (V_x, V_y, V_z)$. Maxwell's curl relations become

$$\text{Curl } V + i \frac{\omega}{c} m^2 U = 0; \tag{2}$$

$$\text{Curl } U - i \frac{\omega}{c} \mu \; V = 0; \tag{3}$$

and the wave equations for U and V become

$$\nabla^2 U + \frac{\omega^2}{c^2} \mu m^2 U = 0; \tag{4}$$

$$\nabla^2 V + \frac{\omega^2}{c^2} \mu m^2 V = 0; \tag{5}$$

wherein

$$m^2 = \epsilon + i \, 4\pi \sigma/\omega \qquad \text{(defining m);} \tag{6}$$

$$m = n(1 + i K) \qquad \text{(defining n and K).} \tag{7}$$

The magnetic permeability, μ, and the dielectric constant, ϵ, are defined so that they are unity for vacuum. σ is the electric conductivity; c is velocity in vacuum and $\omega = 2\pi/T$, where T is the period of vibration of the wave. $i = \sqrt{-1}$.

$$\omega/c = 2\pi/\lambda_o = k \tag{8}$$

where λ_o denotes wavelength in vacuum. As so defined, n and K are the optical constants that are usually listed in handbooks. In most optical problems one will take $\mu = 1$. We shall regard n and K as the fundamental properties that are measured from experimental observation on the wave motion. Under this view, ϵ and σ are derived from knowledge of n and K.

21.2.2 Physical significance. So much confusion exists about the physical significance of the optical constants, n and K, that further discussion is warranted. It is often believed that n is the refractive index such that the phase velocity v is always given by c/n. Suppose that a plane wave is propagated along Z with its electric vector polarized to vibrate along X. Then $U = (U_x, 0, 0)$. Since the wave is plane, U_x is, by assumption, independent of x and y. Hence the wave equation reduces to

$$\frac{\partial^2 U_x}{\partial z^2} + \frac{\omega^2}{c^2} \mu m^2 U_x = 0. \tag{9}$$

It is easily verified by substitution into (9) that a solution is

$$U_x = E_o e^{i \frac{\omega}{c} \mu^{1/2} mz} = E_o e^{-k\mu^{1/2} nKz} e^{i \frac{\omega}{c} \mu^{1/2} nz} \tag{10}$$

One concludes from equations (1) and (10) that

$$E = (1, 0, 0) E_o e^{-k\mu^{1/2} nKz} e^{-i\omega(t - \frac{z}{v})} \tag{11}$$

wherein

$$v = c/\mu^{1/2}n \tag{12}$$

is the phase velocity of the wave in the medium, and $k\mu^{1/2} n K$ is an absorption or extinction coefficient that determines the rate of attenuation of the amplitude with increasing z. We observe from equation (11) that the parallel planes, z = constant, are planes of equiphase (wavefronts) and of equiamplitude. When the planes of equiphase and equiamplitude are parallel, the wave is said to be homogeneous. If, then, $\mu = 1$ and the wave is homogeneous, the optical constant, n, is in fact the refractive index. When a homogeneous wave is incident normally upon any system of plane parallel layers, the wave remains homogeneous. But when a homogeneous wave is incident obliquely upon a system of plane parallel layers, the wave becomes inhomogeneous in the absorbing layers, i.e. planes of equiphase and equiamplitude do not remain strictly parallel when $K \neq 0$. With inhomogeneous waves, the phase velocity is not given by equation (12) and rate of attenuation is not governed by the product $k\mu^{1/2}n$ K. A generalized form of Snell's law of refraction applies, but the actual refractive index is not the optical constant n even when $\mu = 1$. It will not be an objective of this text to dwell upon the effects exhibited by inhomogeneous waves in absorbing media; but it is worth noting that the following systems of equations will include the effects produced by inhomogeneous waves in the absorbing layers. When the system is free of absorption, the waves remain homogeneous even at oblique incidence.

21.2.3 Fresnel's coefficients for normal incidence.

21.2.3.1 Fresnel's coefficients of reflection and transmission with respect to a plane interface between two homogeneous media can be derived with a high degree of rigor. Derivations based upon Maxwell's equations and realistic boundary conditions will be found in almost all textbooks dealing with physical optics or electromagnetic theory. The following Fresnel coefficients apply to normal incidence upon the interface as illustrated in Figure 21.1. The Z-direction is chosen normal to the interface, and the point $z = 0$ shall fall at the interface. To date, the permeabilities, μ, have invariably been unity in the optical applications of thin films. In view of the unlikelihood of cases $\mu \neq 1$, the following considerations will be restricted to cases $\mu = 1$.

21.2.3.2 Let τ_o be a complex number that specifies the amplitude and phase of the incident E-vector at the left hand boundary, $z = 0$. Let r_o be a complex number that specifies the amplitude and phase of the reflected E-vector at the left hand boundary, $z = 0$. Similarly, let τ_1 specify the amplitude and phase of the transmitted E-vector at the right hand boundary, $z = 0$. It can be shown that

$$\frac{r_o}{\tau_o} = \frac{m_o - m_1}{m_o + m_1} \equiv W_1 , \tag{13}$$

and that

$$\frac{\tau_1}{\tau_o} = \frac{2m_o}{m_o + m_1} , \tag{14}$$

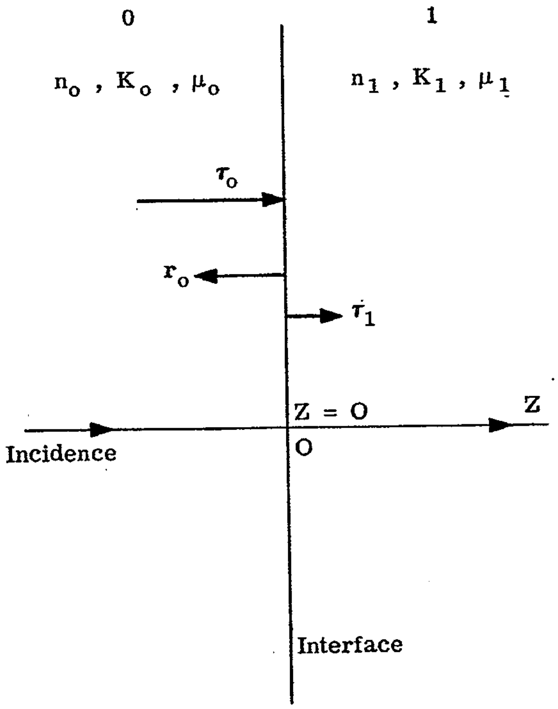


Figure 21. 1- Notation with respect to Fresnel's coefficients for normal incidence.

in which

$$m_j = n_j (1 + i K_j) ; \quad j = 0 \text{ and } 1. \tag{15}$$

The ratios r_o/τ_o and τ_1/τ_o are, respectively, Fresnel's coefficient of reflection and transmission at normal incidence. One may take τ_o as unity without any essential loss of generality.

21. 2. 3. 3 If neither medium is absorbing, $K_o = K_1 = 0$. Hence for interfaces between non-absorbing media the Fresnel coefficients reduce to the better known results

$$r_o/\tau_o = \frac{n_o - n_1}{n_o + n_1} , \tag{13a}$$

and

$$\tau_1/\tau_o = \frac{2n_o}{n_o + n_1} , \tag{14a}$$

in which n_o and n_1 are physically the refractive indices of the two media. When $n_1 > n_o$, one may write

$$r_o/\tau_o = \frac{|n_o - n_1|}{n_o + n_1} \quad e^{\pm i\pi} . \tag{13b}$$

Thus one concludes that,with respect to the electric vector, the phase change on reflection is ± π radians when $n_1 > n_o$. The phase change on transmission across the interface is always zero when the two media are non-absorbing, for then the ratio τ_1/τ_o is necessarily real and positive.

21. 2. 3. 4 As a second example, consider incidence from a non-absorbing medium upon an absorbing medium. Since $m_o = n_o$ and $m_1 = n_1 + i n_1 K_1$, one obtains from equation (13)

$$\frac{r_o}{\tau_o} = \frac{n_o - n_1 - i n_1 K_1}{n_o + n_1 + i n_1 K_1} = \frac{n_o^2 - n_1^2 (1 + K_1^2) - i\, 2 n_o n_1 K_1}{(n_o + n_1)^2 + n_1^2 K_1^2} . \tag{16}$$

For reflection from air to metals, $n_1^2 (1 + K_1^2)$ invariably exceeds n_o^2 . Hence the real and imaginary parts of the Fresnel reflection coefficient r_o/τ_o will usually be negative and the phase angle on reflection will fall in the third quadrant. Thus with $r_o/\tau_o = |r_o/\tau_o|\ e^{i\theta}$, $180^o < \theta < 270^o$. From equation (14),

$$\frac{\tau_1}{\tau_o} = \frac{2n_o}{n_o + n_1 + in_1 K_1} = 2\ \frac{n_o(n_o + n_1) - in_o n_1 K_1}{(n_o + n_1)^2 + n_1^2 K_1^2} \ . \tag{17}$$

The phase angle introduced by transmission through the interface will fall in the fourth quadrant.

21.2.4 Fresnel's coefficients for oblique incidence.

21.2.4.1 The direction, Z, has been taken along the normal to the interface. It is convenient to choose the X-direction in the plane of incidence as illustrated in Figure 21.2. X, Z is then the plane of incidence. Introduce for brevity,

$$p_o = \sin i_o ; \quad q_o = \cos i_o ; \tag{18}$$

where i_o is the angle of incidence. Let

$$M_\nu = (m_\nu^2 - m_o^2 p_o^2)^{1/2} \tag{19}$$

wherein the suffix $\nu = 0, 1$ and refers to the media of Figure 21.2, and wherein m_ν is defined by equation (15). M_ν is complex imaginary whenever m_o or m_1 is complex imaginary. It can happen, as in total internal reflection, that $(m_\nu^2 - m_o^2 p_o^2)$ is real and less than zero. In such cases

$$M_\nu = i\ |m_\nu^2 - m_o^2 p_o^2|^{1/2} ; \quad i = \sqrt{-1} \ . \tag{19a}$$

21.2.4.2 We need to introduce two more quantities W_ν and F_ν. These are

$$W_\nu = \frac{M_{\nu-1} - M_\nu}{M_{\nu-1} + M_\nu} \ , \tag{20}$$

and

$$F_\nu = \frac{m_\nu^2\ M_{\nu-1} - m_{\nu-1}^2\ M_\nu}{m_\nu^2\ M_{\nu-1} + m_{\nu-1}^2\ M_\nu} \ . \tag{21}$$

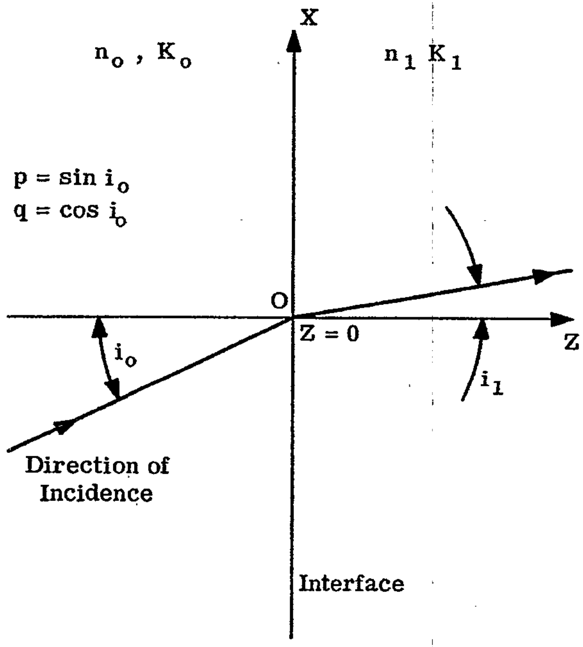


Figure 21.2 - Notation with respect to Fresnel's coefficients for oblique incidence. Plane XZ is chosen as the plane of incidence. i_o is the angle of incidence.

The laws of reflection depend upon the state of polarization of the incident wave. It suffices to consider the Fresnel coefficients of reflection and transmission for two states of polarization. In one of these states, the electric vector is perpendicular to the plane of incidence so that $E = (0, E_y, 0)$. In the second state of polarization, the magnetic vector is perpendicular to the plane of incidence so that $H = (0, H_y, 0)$. A minimum number of four Fresnel coefficients becomes necessary.

21. 2. 4. 3 Consider first the state of polarization in which the electric vector is perpendicular to the plane of incidence. Let τ_o and r_o be complex numbers that specify the amplitude and phase of the incident and reflected electric vector at the origin, O, at left hand boundary of the interface, $z = 0$, Figure 21. 2. Let, similarly, τ_1 specify the amplitude and phase of the transmitted electric vector at the point $x = z = 0$ at the right hand boundary of the interface. The ratios r_o/τ_o and τ_1/τ_o are now Fresnel's coefficients of reflection and transmission, respectively, for the electric vector. These ratios are given by

$$\frac{r_o}{\tau_o} = W_1 = \frac{M_o - M_1}{M_o + M_1} , \tag{22}$$

and

$$\frac{\tau_1}{\tau_o} = \frac{2M_o}{M_o + M_1} . \tag{23}$$

The similarity of equations (13) and (22), and of equations (14) and (23) should be noted. These results become alike when $p_o = 0$ (normal incidence).

21. 2. 4. 4 Consider next the state of polarization in which the magnetic vector is perpendicular to the plane of incidence. Let T_o and R_o be complex numbers that specify the amplitude and phase of the incident and reflected H-vectors, respectively, at point $x = 0$ at the left hand boundary of the interface at $z = 0$. Let T_1 specify the amplitude and phase of the transmitted H-vector at point $x = 0$ at the right hand boundary of the interface. The ratios R_o/T_o and T_1/T_o define Fresnel's coefficient of reflection and transmission, respectively, for the perpendicular component of the magnetic vector. These ratios are given by

$$\frac{R_o}{T_o} = F_1 = \frac{m_1^2 M_o - m_o^2 M_1}{m_1^2 M_o + m_o^2 M_1} , \tag{24}$$

and

$$\frac{T_1}{T_o} = \frac{2m_1^2 M_o}{m_1^2 M_o + m_o^2 M_1} . \tag{25}$$

21. 2. 4. 5 As an application or test of equations (22) and (24), consider total internal reflection. This phenomenon occurs when neither medium absorbs, so that $m_o = n_o$ and $m_1 = n_1$, and when $n_o > n_1$. Total internal reflection occurs when the angle of incidence $i_o \geq \sin^{-1} \frac{n_1}{n_o}$, i. e. when $n_o^2 p_o^2 \geq n_1^2$. Therefore, from equation (19a) it follows that M_1 is a pure imaginary number for angles i_o beyond the critical angle for total internal reflection. Since m_o, m_1 and M_o are real numbers, the numerators of equations (22) and (24) are complex conjugates with respect to the denominators. Hence it follows at once that both $|r_o/\tau_o|^2$ and $|R_o/T_o|^2$ are unity. The energy reflectance is therefore total, as required by experiment. One can verify that the numerator of equation (24) is zero at Brewster's angle. This means that the H-vector will not be reflected when it is perpendicular to the plane of incidence, or, equivalently, that the E-vector will not be reflected at Brewster's angle when it vibrates in the plane of incidence.

21. 2. 4. 6 Whereas $|r_o/\tau_o|^2$ and $|R_o/T_o|^2$ can always be interpreted as energy reflectances for the states of polarization to which they apply, the square of the absolute values of Fresnel's coefficients of transmission τ_1/τ_o and T_1/T_o do not necessarily equal the actual energy transmittances. This matter will be treated in some detail since it has been the source of much confusion. With respect to thin films or interfaces, one invariably wishes to compute energy transmittance for cases in which the initial and last media are non-absorbing. Accordingly, emphasis will be placed upon non-absorbing initial and final media.

21. 2. 5 <u>The electromagnetic field when the electric vector is perpendicular to the plane of incidence.</u>

21. 2. 5. 1 Let us suppose that an homogeneous incident wave has been generated in the initial medium such that the incident electric vector is the known vector,

$$E_{\text{incident}} = (0, 1, 0)\, \tau_o\, e^{-i\omega t}\, e^{ikm_o(p_o x + q_o z)} \tag{26}$$

From equations (1) and (26),

$$U_{\text{incident}} = (0,\ 1,\ 0)\ \tau_o\ e^{ikm_o (p_o x + q_o z)} \tag{26a}$$

Curl relation, equation (3), serves to determine V from U. For plane waves incident in the X, Z plane, vectors U and V are independent of y. Consequently equation (3) yields directly the result,

$$V_x = \frac{i}{k\mu}\ \frac{\partial U_y}{\partial z}, \quad V_y = 0, \quad V_z = \frac{-i}{k\mu}\ \frac{\partial U_y}{\partial x}; \tag{27}$$

a result that holds in any medium. Since U_y is given from equation (26a), one can compute vector, V, from equation (27) and then write the incident H-vector from equation (1). One obtains in a straightforward manner for the case $\mu = 1$,

$$H_{\text{incident}} = (-\,q_o,\ 0,\ p_o)\ m_o\, \tau_o\ e^{-i\omega t}\ e^{ikm_o(p_o x + q_o z)}, \tag{26b}$$

The incident electromagnetic field becomes known when τ_o is assigned. The time averaged Poynting vector, S, is given except for an unimportant factor by the vector product

$$S = E \times \bar{H} + \bar{E} \times H . \tag{28}$$

From equations (26), (26b) and (28),

$$S_{\text{incident}} = (p_o,\ 0,\ q_o)\ 2n_o \left|\tau_o\right|^2\ e^{-2kn_o K_o (p_o x + q_o z)}, \tag{26c}$$

since $m_o + \bar{m}_o = 2n_o$. This energy flux is along the direction of propagation of the incident wave.

21.2.5.2 The incident E-vector is reflected with the Fresnel reflection coefficient r_o/τ_o. Consequently,

$$E_{\text{reflected}} = (0,\ 1,\ 0)\ \frac{r_o}{\tau_o}\ \ \tau_o\ e^{-i\omega t}\ e^{ikm_o(p_o x - q_o z)} . \tag{29}$$

Just as equation (27) served for determining H_{incident} from E_{incident}, it serves again for determining $H_{\text{reflected}}$ from $E_{\text{reflected}}$. One obtains straightforwardly,

$$H_{\text{reflected}} = (q_o,\ 0,\ p_o)\ m_o\ \frac{r_o}{\tau_o}\ \ \tau_o\ e^{-i\omega t}\ e^{ikm_o(p_o x - q_o z)} . \tag{29a}$$

The Fresnel coefficient r_o/τ_o of equation (22) determines the reflected electromagnetic field. Upon evaluating the Poynting vector, S, from equations (28), (29) and (29a), one obtains

$$S_{\text{reflected}} = (p_o,\ 0,\ -\,q_o)\ 2n_o \left|\frac{r_o}{\tau_o}\right|^2 \left|\tau_o\right|^2\ e^{-2kn_o K_o (p_o x - q_o z)}, \tag{29b}$$

an energy flux along the direction of the reflected wave.

21.2.5.3 The electric vector transmitted across the interface $z = 0$, Figure 21.2, has the form*

$$E_{\text{transmitted}} = (0,\ 1,\ 0) \left(\frac{\tau_1}{\tau_o}\right) \tau_o\ e^{-i\omega t}\ e^{ik(m_o p_o x + M_1 z)} . \tag{30}$$

Correspondingly, equation (27) yields

$$H_{\text{transmitted}} = (-\,M_1,\ 0,\ m_o p_o) \left(\frac{\tau_1}{\tau_o}\right) \tau_o\ e^{-i\omega t}\ e^{ik(m_o p_o x + M_1 z)} . \tag{30a}$$

The wave described by equations (30) and (30a) is in general inhomogeneous,** It is more convenient for many

*It can be verified easily by substitution, that vector U defined by equations (1) and (30) satisfies wave equation (4) in medium number one, provided that M_1 obeys equation (19).

**Suppose, for example, that $m_o = n_o$ but that m_1 is complex. The first medium is then non-absorbing and the second medium is absorbing. M_1 is now complex imaginary. Write $M_1 = R_e(M_1) + i\, I_m(M_1)$. Since

$$e^{ik(m_o p_o x + M_1 z)} = e^{-kI_m(M_1)z}\ \ e^{ik[n_o p_o x + R_e(M_1) z]}$$

the planes of equiamplitude are parallel to the interface $z = 0$ whereas the planes of equiphase are the planes $n_o p_o x + R_e(M_1)\, z$ = constant.

purposes to write the second exponential in equations (30) and (30a) in the expanded form,

$$e^{ik(m_o p_o x + M_1 z)} = e^{-k[n_o K_o p_o x + I_m (M_1) z]} \; e^{ik[n_o p_o x + R_e (M_1) z]} , \quad (30b)$$

wherein $R_e (M_1)$ and $I_m (M_1)$ denote, respectively, the real and imaginary parts of M_1 . The Poynting vector, S, can now be found in a straightforward manner by applying equation (28) to equations (30), (30a) and (30b). The most general result is

$$S_{\text{transmitted}} = \left[n_o p_o,\ 0,\ R_e (M_1)\right] \ 2 \left|\frac{\tau_1}{\tau_o}\right|^2 \left|\tau_o\right|^2 \ e^{-2k[n_o K_o p_o x + I_m (M_1) z]} \quad (30c)$$

By comparing the three components $n_o p_o$, 0 and $R_e (M_1)$ with the arguments $n_o p_o x + R_e(M_1) z$ of the second exponential of equation (30b), one finds that $S_{\text{transmitted}}$ is an energy flow along the direction of propagation of the equiphase surfaces (wavefronts) in the second medium when the electric vector is perpendicular to the plane of incidence. Equation (30c) becomes most significant and useful when the initial medium has negligible absorption so that $K_o = 0$; for then the Poynting vector is constant in planes z = constant. Explicitly,

$$S_{\text{transmitted}} = \left[n_o p_o,\ 0,\ R_e(M_1)\right] \ 2 \left|\frac{\tau_1}{\tau_o}\right|^2 \left|\tau_o\right|^2 \ e^{-2k I_m (M_1) z} . \quad (30d)$$

If, also, the second medium is non-absorbing, $m_1 = n_1$ and M_1 is real. Attenuation with z does not occur because $I_m (M_1)$ is zero.

21.2.5.4 Since the x - and z - components of $S_{\text{transmitted}}$ are proportional to $n_o p_o$ and $R_e(M_1)$, respectively, the angle, i_1, between $S_{\text{transmitted}}$ and the Z-axis is given by

$$\tan i_1 = n_o p_o / R_e (M_1) , \quad (30e)$$

or

$$\sin i_1 = n_o p_o / \left[n_o^2 \, p_o^2 + R_e^2 (M_1)\right]^{1/2} . \quad (30f)$$

As stated, this direction of the Poynting vector, S, is parallel to the direction of propagation of the wavefronts. Let

$$(n_a)_1 = \left[n_o^2 \, p_o^2 + R_e^2 (M_1)\right]^{1/2} . \quad (31)$$

Equation (30f) now states that the law of refraction is

$$(n_a)_1 \ \sin i_1 = n_o p_o = n_o \sin i_o \quad (32)$$

in which $(n_a)_1$ is in fact the actual refractive index of medium number one, Figure 21.2. When $m_o = n_o$ and $m_1 = n_1$, $(n_a)_1 = n_1$ so that the more general law of equation (32) degenerates into the usual form known as Snell's law. Equation (32) serves to determine the direction of the Poynting vector and the "rays" in medium number one.

21.2.5.5 The manner in which the actual refractive index, $(n_a)_1$, depends upon the angle of incidence, i_o , and upon $n_1 K_1$, is described by the table in Table 21.1 for the case in which $n_o = 1$, $K_o = 0$, and $n_1 = 1.75$. $(n_a)_1$ increases with the angle of incidence and with the product $n_1 K_1$. Because $n K$ is less than 0.02 in the usual lenses, plates, etc., the more general law of equation (32) is not of great importance to geometrical optics.

21.2.5.6 The absolute value of the vector $n_o p_o$, 0, $R_e(M_1)$ in equation (30d) is $n_o^2 \, p_o^2 + R_e^2 (M_1)^{1/2}$. Hence equation (30d) can be written in the more useful and significant form

$$\left|S_{\text{transmitted}}\right| = 2 \, (n_a)_1 \left|\frac{\tau_1}{\tau_o}\right|^2 \left|\tau_o\right|^2 \ e^{-2k I_m (M_1) z} . \quad (33)$$

The relations between the Fresnel coefficients, r_o/τ_o and τ_1/τ_o, and the energy reflectance and energy transmittance, respectively, can now be clarified unambiguously. First, we note from equations (26c) and (29b) that at $z = 0$,

$$\frac{\left|S_{\text{reflected}}\right|}{\left|S_{\text{incident}}\right|} = \left|\frac{r_o}{\tau_o}\right|^2 \quad \text{(energy reflectance)} , \quad (34)$$

a result that holds whether or not m_o is complex. Secondly, we note from equations (33) and (26c), that

i_o / $n_1 K_1$	0^o	10^o	20^o	40^o	60^o	80^o
0.01	1.750000	1.750000	1.750001	1.7500045	1.750009	1.750013
0.02	1.750000	1.750001	1.7500045	1.750018	1.750037	1.750053
0.04	1.750000	1.7500045	1.750015	1.750071	1.750148	1.750212
0.06	1.750000	1.750010	1.750041	1.750160	1.750333	1.750493
0.08	1.750000	1.7500205	1.750072	1.750284	1.750591	1.750843
0.10	1.750000	1.750028	1.750113	1.750444	1.750921	1.751314
0.50	1.750000	1.750656	1.752605	1.760028	1.770224	1.778142
1.00	1.750000	1.752131	1.758389	1.781126	1.809765	1.830155
2.00	1.750000	1.754882	1.768962	1.817250	1.872569	1.908867
4.00	1.750000	1.757218	1.7778585	1.846750	1.922468	1.970524

Table 21.1- Table of values of the actual refractive index $(n_a)_1$ of the second medium as a function of the angle of incidence, i_o, and of $n_1 K_1$ for the case in which the first medium does not absorb and has the optical constant $n_o = 1$. The optical constant $n_1 = 1.750000$.

at $z = 0$,

$$\frac{|S_{\text{transmitted}}|}{|S_{\text{incident}}|} = \frac{(n_a)_1}{n_o}\left|\frac{\tau_1}{\tau_o}\right|^2 \qquad (35)$$

a result that holds when $m_o = n_o$. Consider next the conservation of energy flow across any element of area, ΔA, of the interface $z = 0$, Figure 21.3. Energy is conserved provided that

$$|S_{\text{reflected}}| \quad \Delta A_r + |S_{\text{transmitted}}| \, \Delta A_t = |S_{\text{incident}}| \quad \Delta A_i \,, \qquad (36)$$

in which the elements of area are interrelated as indicated in Figure 21.3. Division of equation (36) by the right hand member produces the following important result,

$$\frac{|S_{\text{reflected}}|}{|S_{\text{incident}}|}\frac{\Delta A_r}{\Delta A_i} + \frac{|S_{\text{transmitted}}|}{|S_{\text{incident}}|}\frac{\Delta A_t}{\Delta A_i} = 1. \qquad (37)$$

The first left hand member is energy reflectance from the element of area ΔA. Because $\Delta A_r = \Delta A_i$, comparison of the first left hand member of equations (37) and (34), shows that the ratio $|r_o/\tau_o{}^2$ is in fact energy reflectance. The second left hand member of equation (37) is energy transmittance of the element of area ΔA. Since $\Delta A_t/\Delta A_i = \cos i_1/\cos i_o$, equations (37) and (35) show that the energy transmittance of any element ΔA of the interface, $z = 0$, is given by

$$\text{Energy transmittance} = \frac{(n_a)_1}{n_o}\left|\frac{\tau_1}{\tau_o}\right|^2 \frac{\cos i_1}{\cos i_o}\,, \qquad (38)$$

in which $(n_a)_1$ is the actual refractive index of the last medium, and τ_1/τ_o is Fresnel's transmission coefficient for the case in which the E-vector is perpendicular to the plane of incidence.

21.2.5.7 We shall verify that equations (22) and (23) for the Fresnel coefficients are consistent with the law of conservation of energy. Upon introducing equations (34) and (35) into equation (37), one obtains the condition for conservation of energy at the interface $z = 0$ in the form

$$\left|\frac{r_o}{\tau_o}\right|^2 + \frac{(n_a)_1 \cos i_1}{n_o \cos i_o}\left|\frac{\tau_1}{\tau_o}\right|^2 = 1. \qquad (39)$$

Hence one should obtain

$$\frac{|M_o - M_1|^2}{|M_o + M_1|^2} + \frac{(n_a)_1 \cos i_1}{n_o \cos i_o}\,\frac{4|M_o|^2}{|M_o + M_1|^2} = 1. \qquad (40)$$

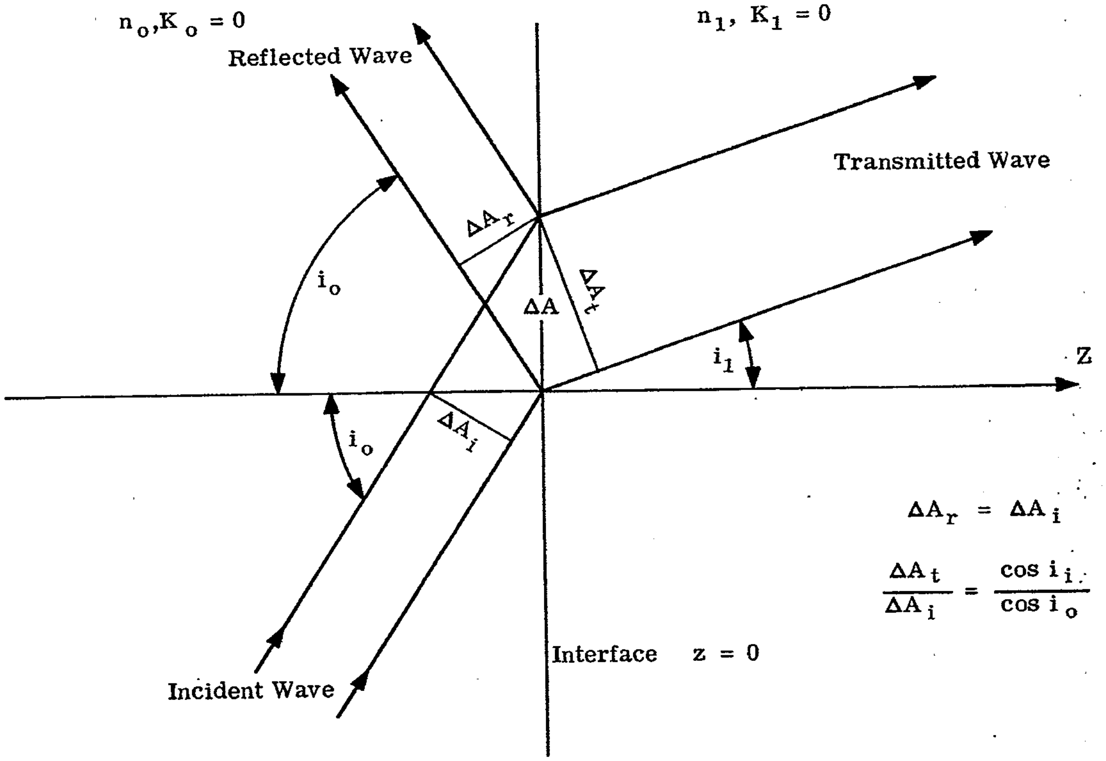


Figure 21. 3- Notation with respect to the flow of energy flux in the Poynting vectors for cases in which absorption is negligible in the initial and final media.

To avoid difficulties, let us test the case $m_o = n_o$ and $m_1 = n_1$. In this case $n_o \cos i_o = M_o$ and $(n_a)_1 \cos i_1 = M_1$. Hence equation (40) becomes the required identity.

21. 2. 6 <u>The electromagnetic field when the magnetic vector is perpendicular to the plane of incidence.</u>

21. 2. 6. 1 When the y-component of the H-vector is given, the curl relation, equation (2), determines the vector U. In the present case

$$U_x = \frac{-i}{km^2} \frac{\partial V_y}{\partial z} \; ; \quad U_y = 0 \; ; \quad U_z = \frac{i}{km^2} \frac{\partial V_y}{\partial x} \; . \qquad (41)$$

Consistent with equation (26), we suppose that the incident magnetic vector is given as the homogeneous wave

$$H_{\text{incident}} = (0, 1, 0) \; T_o \; e^{ikm_o (p_o x + q_o z)} \; . \qquad (42)$$

Then from equations (41) and (42),

$$E_{\text{incident}} = (q_o , 0, - p_o) \frac{T_o}{m_o} \; e^{ikm_o (p_o x + q_o z)} . \qquad (42a)$$

Therefore,

$$S_{\text{incident}} = (p_o , 0, q_o) \frac{2 n_o}{|m_o|^2} \left| T_o \right|^2 e^{-2kn_o K_o (p_o x + q_o z)} . \qquad (42b)$$

Since the Fresnel reflection coefficient is now R_o/T_o , equation (24),

$$H_{\text{reflected}} = (0, 1, 0) \left(\frac{R_o}{T_o}\right) T_o \; e^{ikm_o (p_o x - q_o z)} , \qquad (43)$$

$$E_{\text{reflected}} = (- q_o, 0, - p_o) \; \frac{T_o}{m_o} \left(\frac{R_o}{T_o}\right) e^{\,ikm_o (p_o x - q_o z)} , \qquad (43a)$$

and

$$S_{\text{reflected}} = (p_o , 0, - q_o) \frac{2 n_o}{|m_o|^2} (T_o)^2 \left|\frac{R_o}{T_o}\right|^2 e^{-2kn_o K_o (p_o x + q_o z)} . \qquad (43b)$$

The incident and reflected Poynting vectors point along the direction of propagation of the corresponding waves.

21. 2. 6. 2 As in equation (30),

$$H_{\text{transmitted}} = (0,\ 1,\ 0) \left(\frac{T_1}{T_o}\right) T_o\ e^{ik(m_o p_o x + M_1 z)}, \qquad (44)$$

therefore ,

$$E_{\text{transmitted}} = (M_1,\ 0,\ -m_o p_o)\ \frac{1}{m_1^2} \left(\frac{T_1}{T_o}\right) T_o\ e^{ik(m_o p_o x + M_1 z)}. \qquad (44a)$$

When $m_o = n_o$,

$$S_{\text{transmitted}} = \left(n_o p_o \left[\frac{1}{m_1^2} + \frac{1}{\overline{m_1^2}}\right],\ 0,\ \frac{M_1}{m_1^2} + \frac{\overline{M}_1}{\overline{m_1^2}}\right) \left|\frac{T_1}{T_o}\right|^2 \left|T_o\right|^2 e^{-2k I_m (M_1) z}. \qquad (44b)$$

Since

$$\frac{1}{m_1^2} + \frac{1}{\overline{m_1^2}} = \frac{m_1^2 + \overline{m_1^2}}{|m_1|^4} = \frac{2 R_e (m_1^2)}{|m_1|^4}, \qquad (44c)$$

the x-component of S would change sign with $R_e (m_1^2)$. The possibility $R_e (m_1^2) < 0$ would violate also equation (6) because a negative dielectric constant, ϵ_1, has no physically acceptable meaning. Thus, measured values of n and K can be valid [1a] only when

$$R_e (m^2) = n^2 (1 - K^2) = \epsilon > 0. \qquad (44d)$$

With respect to the listed values of n and K for metals, one generally finds that $K^2 > 1$ so that $n^2 (1 - K^2) < 0$. Application of the theory to cases in which the listed K-values exceed unity is to be regarded, therefore, with skepticism.[1] When $R_e (m_1^2) > 0$, one finds quite directly from equation (44b) that

$$S_{\text{transmitted}} = \left(n_o p_o,\ 0,\ R_e (M_1)\right) 2\ \frac{R_e (m_1^2)}{|m_1|^4} \left|\frac{T_1}{T_o}\right|^2 \left|T_o\right|^2 e^{-2k I_m (M_1) z}$$

$$+ (0,\ 0,\ 1)\ 2\ I_m (M_1)\ \frac{I_m (m_1^2)}{|m_1|^4} \left|\frac{T_1}{T_o}\right| \left|T_o\right|^2 e^{-2k I_m (M_1) z}. \qquad (44e)$$

The first right hand vector is parallel to the normals to the equiphase surfaces (wavefronts) in the second medium. The second right hand vector is normal to the interface and vanishes as the imaginary part of m approaches zero. As $n_1 K_1$ is increased from zero, the transmitted Poynting vector, S, departs from the direction of the wave normals and tends toward perpendicularity with the interface when the incident H-vector is perpendicular to the plane of incidence.

21. 2. 6. 3 We shall restrict our considerations of equation (44e) to cases in which K_1 is so small that the vector with components (0, 0, 1) is negligible. The transmitted Poynting vector, S, and the wave normals are then practically parallel. As in 21. 2. 5, we obtain

$$\left|S_{\text{transmitted}}\right| = \frac{2 (n_a)_1}{|m_1|^4} R_e (m_1^2) \left|\frac{T_1}{T_o}\right|^2 \left|T_o\right|^2 e^{-2k I_m (M_1) z}, \qquad (44f)$$

in which $(n_a)_1$ is given by equation (31) and in which S points along the direction determined by equation (32). From equations (42b) and (43b), one finds that at $z = 0$

$$\frac{\left|S_{\text{reflected}}\right|}{\left|S_{\text{incident}}\right|} = \left|\frac{R_o}{T_o}\right|^2 \quad \text{(energy reflectance)}, \qquad (45)$$

a result that holds whether or not m_o is complex. One finds from equations (44f) and (42b) that at the interface, $z = 0$, with the approximation $R_e (m_1^2) = n_1^2$

$$\frac{\left|S_{\text{transmitted}}\right|}{\left|S_{\text{incident}}\right|} = \frac{(n_a)_1\ n_1^2}{|m_1|^4}\ n_o \left|\frac{T_1}{T_o}\right|^2, \qquad (45a)$$

a result that has been specialized to the case $m_o = n_o$. Under the restriction that the transmitted Poynting vector is practically parallel to the wave normals in the second medium, equation (37) holds again. Hence

[1] See P. Drude pg. 368 Theory of Optics, Longmans Green & Co. 1902.

[1a] For a more modern viewpoint relative to cases in which $n^2(1-K^2) = \epsilon < 0$, consult the physical interpretation of negative dielectric constants by Max Born and Emil Wolf, Principles of Optics, Pergamon Press, (1959), pp 618 and 623.

$R_o/T_o\ ^2$ is in fact energy reflectance. The steps leading to equation (38) now yield, instead of equation (38), the result,

$$\text{Energy transmittance} = \frac{(n_a)_1\ n_1^2\ n_o}{|m_1|^4}\left|\frac{T_1}{T_o}\right|^2 \frac{\cos i_1}{\cos i_o}, \qquad (45b)$$

wherein T_1/T_o is Fresnel's coefficient of transmission of the H-vector when this vector is perpendicular to the plane of incidence. If $m_1 = n_1$, $(n_a)_1 = n_1$ and

$$\text{Energy transmittance} = \frac{n_o}{n_1}\left|\frac{T_1}{T_o}\right|^2 \frac{\cos i_1}{\cos i_o}, \qquad (45c)$$

equations (45b) and (45c) should be compared with equation (38) for the state of polarization in which the E-vector is perpendicular to the plane of incidence. It will be seen that the ratios of the refractive indices are the inverse of one another. The Fresnel coefficients R_o/T_o and T_1/T_o of equations (24) and (25) are consistent with the requirement that the corresponding flow of energy shall be conserved at the interface $z = 0$.

21.2.7 Summary with respect to the Fresnel coefficients.

21.2.7.1 The Fresnel coefficients r_o/τ_o and τ_1/t_o given by equations (22) and (23), respectively, determine the reflected and transmitted E-vector when this vector is perpendicular to the plane of incidence. The corresponding incident, reflected and transmitted electromagnetic fields are given by equations (26), (29) and (30), respectively. Whereas $|r_o/\tau_o|^2$ is energy reflectance, the quantity $|\tau_1/\tau_o|^2$ is not, in general, energy transmittance. A study of the incident, reflected and transmitted Poynting vectors shows that energy transmittance across the interface between the two media is given by equation (38).

21.2.7.2 The Fresnel coefficients R_o/T_o and T_1/T_o refer to reflection and transmission of the H-vector when it is perpendicular to the plane of incidence. With this state of polarization the incident, reflected and transmitted electromagnetic fields are determined by equations (42), (43) and (44). Examination of the time-averaged Poynting vectors shows that $|R_o/\tau_o|^2$ is energy reflectance but that $|T_1/T_o|^2$ is not necessarily energy transmittance. Equation (45b) is an approximate one for computing energy transmittance. As the absorption of the second medium vanishes, equation (45b) becomes more exact and approaches the result given by equation (45c) for the case in which neither medium is absorbing.

21.2.7.3 When the first medium is non-absorbing and the second medium is absorbing, equation (31) and (32) show how the wave normals are refracted. This law of refraction is the same for both states of polarization. Whereas the time-averaged Poynting vector points along the wave normals when the E-vector is perpendicular to the plane of incidence, this Poynting vector does not always do so when the H-vector is perpendicular to the plane of incidence.

21.2.8 Normal incidence upon multilayers.

21.2.8.1 Once the Fresnel coefficients for an interface between two media have been established, it is not necessary to resubmit the numerous interfaces of a multilayer to the Maxwell equations and the boundary conditions in order to construct the theory of thin films. Instead, the following instructive method can be utilized. We consider the useful case of normal incidence in order to simplify the presentation. We may suppose, without any essential loss of generality, that the electric vector is perpendicular to the plane X, Z, Figure 21.4. In other words, we choose Z along the normal to the interfaces and take Y as the direction of vibration of the electric vector. Because the incident waves are assumed to be plane, the magnetic vector now vibrates along the X-direction.

21.2.8.2 Waves propagated to the right and left in Figure 21.4 are regarded as transmitted and reflected waves, respectively. In dealing with more than one interface, it is convenient to take τ_ν as a complex number that specifies the amplitude and phase of the "transmitted" electric vector at the right hand boundary of the νth medium or layer for the range of ν from 0 to N. As indicated in Figure 21.4, τ_{N+1} specifies the amplitude and phase of the wave in the last medium N + 1 at the left hand boundary of this medium. Similarly, r_ν is a complex number that specifies the amplitude and phase of the reflected electric vector at the right hand boundary of the ν th medium or layer. A reflected wave does not exist in the last medium because it extends indefinitely along Z. The ratio r_ν/τ_ν is complex reflectance at the right hand boundary of the νth medium or layer. We define

$$\rho_\nu \equiv r_\nu/\tau_\nu \quad \text{(complex reflectance).} \qquad (46)$$

21.2.8.3 The Fresnel coefficient of reflectance from medium number 0 to medium number 1 is given by $W_1 = (M_o - M_1)/(M_o + M_1)$ as in equation (22). More generally, the Fresnel coefficient of reflectance at the interface between the $\nu - 1$th and νth layer is given by W_ν, equation (20), when the light is incident

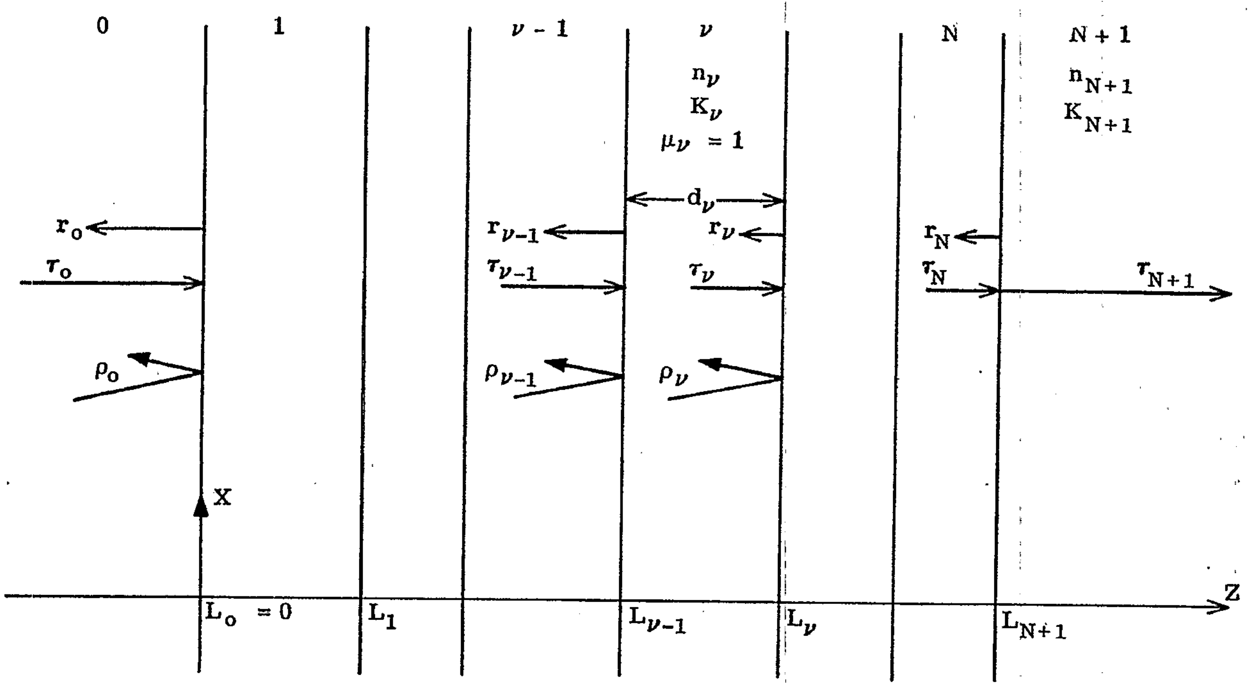


Figure 21. 4- Notation with respect to a system of N layers at normal incidence. The electric vector is taken perpendicular to the XZ plane. Incidence is from the O th medium upon layer #1. Both the initial and final medium #N + 1 are assumed to extend indefinitely along Z.

from the ν-1th layer upon the νth layer. When the direction of incidence is reversed, the effect is to interchange the integers ν and ν-1 in equation (20) so that Fresnel's coefficient of reflection becomes $-W_\nu$. In dealing with a system of layers it is therefore convenient to call W_ν the Fresnel coefficient of reflectance, when the electric vector is polarized to vibrate along the Y-direction. The ratios $\rho_\nu = r_\nu/\tau_\nu$ are equal to the Fresnel coefficients of reflectance only in special cases such as, for example, the single interface of Figure 21. 1.

21. 2. 8. 4 We have seen, as in equation (23), that Fresnel's coefficient of transmission across the interface from medium number 0 into medium number 1 is equal to $2M_o/(M_o + M_1)$. More generally, Fresnel's coefficient of transmission across the interface from the $(\nu - 1)$th into the νth medium is equal to $2M_{\nu-1}/(M_{\nu-1} + M_\nu)$. Fresnel's coefficient of transmission through this interface in the opposite direction is equal to $2M_\nu/(M_{\nu-1} + M_\nu)$. For normal incidence $M_\nu = m_\nu = n_\nu(1 + i K_\nu)$ since $p_o = 0$ in equation (19). Let

$$\beta_\nu \equiv \frac{4\pi}{\lambda} m_\nu d_\nu = \frac{4\pi}{\lambda} n_\nu d_\nu + i \frac{4\pi}{\lambda} n_\nu K_\nu d_\nu , \qquad (47)$$

where d_ν is the thickness of the νth layer. When $K_\nu = 0$, β_ν is twice the optical path (in radians) of the νth layer.

21. 2. 8. 5 The following equilibrium theory for the flow of the electric vector through the multilayer can now be derived in a simple manner. We fix our attention upon the equilibrium flow at the $(\nu-1)$th and νth layers as illustrated in Figure 21. 5. Consider the flow described by $r_{\nu-1}$. First, the flow $\tau_{\nu-1}$ is reflected at the interface at the right hand side of the $(\nu$-1)th layer in the direction of $r_{\nu-1}$ as the flow $\tau_{\nu-1} W_\nu$. Secondly, the flow r_ν arrives at the left side of the νth layer as the flow $r_\nu e^{i\frac{\beta\nu}{2}}$ and then passes through the interface subject to the Fresnel coefficient of transmission $2M_\nu/(M_{\nu-1} + M_\nu)$. Hence

$$r_{\nu-1} = \tau_{\nu-1} W_\nu + r_\nu e^{i\frac{\beta\nu}{2}} \frac{2M_\nu}{M_{\nu-1} + M_\nu} . \qquad (48)$$

Similarly,

$$\tau_\nu = \tau_{\nu-1} \frac{2M_{\nu-1}}{M_{\nu-1} + M_\nu} e^{i\frac{\beta\nu}{2}} - r_\nu W_\nu e^{i\beta\nu} . \qquad (48a)$$

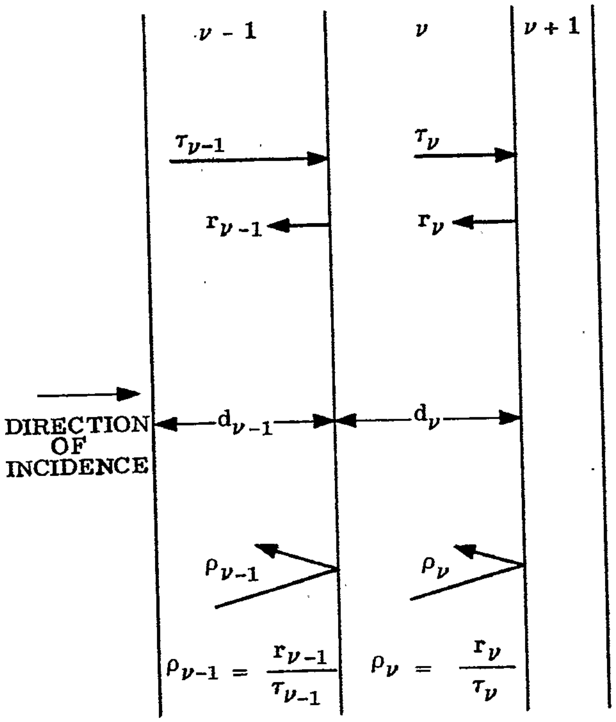


Figure 21. 5- Flow conditions at the (ν-1) th and ν th layers.

The second right hand member has the factor $e^{i\beta\nu}$ because the flow r_ν must cross the νth layer twice. Furthermore, the flow r_ν is reflected at the left hand interface of the νth layer so that the corresponding Fresnel coefficient of reflection is $-W_\nu$. From equation (48),

$$\frac{r_{\nu-1}}{\tau_{\nu-1}} = \rho_{\nu-1} = W_\nu + \frac{r_\nu}{\tau_{\nu-1}} e^{i\frac{\beta\nu}{2}} \frac{2M_\nu}{M_{\nu-1} + M_\nu} . \qquad (48b)$$

From equation (48a),

$$\tau_{\nu-1} = (\tau_\nu + r_\nu W_\nu e^{i\beta\nu}) \frac{M_{\nu-1} + M_\nu}{2M_{\nu-1}} e^{-i\frac{\beta\nu}{2}} . \qquad (48c)$$

By eliminating $\tau_{\nu-1}$ from equation (48b) with the aid of equation (48c), one obtains

$$\rho_{\nu-1} = W_\nu + \frac{r_\nu e^{i\beta\nu}}{\tau_\nu + r_\nu W_\nu e^{i\beta\nu}} \frac{4M_{\nu-1}M_\nu}{(M_{\nu-1} + M_\nu)^2} . \qquad (48d)$$

Dividing numerator and denominator in the right hand member by τ_ν yields the result

$$\rho_{\nu-1} = W_\nu + \frac{\rho_\nu e^{i\beta\nu}}{1 + W_\nu \rho_\nu e^{i\beta\nu}} \frac{4M_{\nu-1}M_\nu}{(M_{\nu-1} + M_\nu)^2} . \qquad (48e)$$

From equation (48e),

$$\rho_{\nu-1} = \frac{\rho_\nu e^{i\beta\nu}\left[W_\nu^2 + 4M_{\nu-1}M_\nu/(M_{\nu-1} + M_\nu)^2\right] + W_\nu}{1 + W_\nu \rho_\nu e^{i\beta\nu}} . \qquad (48f)$$

It follows directly from the definition of W_ν, equation (20), that

$$W_\nu^2 + 4M_{\nu-1}M_\nu/(M_{\nu-1} + M_\nu)^2 = 1 . \qquad (49)$$

Hence

$$\rho_{\nu-1} = \frac{\rho_\nu \, e^{i\beta\nu} + W_\nu}{1 + W_\nu \, \rho_\nu \, e^{i\beta\nu}} \,, \qquad (50)$$

a recursion formula that enables one to compute $\rho_{\nu-1}$ from ρ_ν and the given properties β_ν and W_ν of the νth layer. Equation (50) is the well known result that follows rigorously from the Maxwell equations and the boundary conditions appropriate to a multilayer.

21.2.8.6 The method for computing the complex reflectance ρ_o of the entire multilayer is now clear. Since $\rho_{N+1} = 0$ because no reflected wave exists in the final medium, number N + 1, it follows from equation (50) that

$$\rho_N = W_{N+1} = \frac{M_N - M_{N+1}}{M_N + M_{N+1}} \,. \qquad (51)$$

This result is to be expected because ρ_N ought to be the Fresnel coefficient of reflection at the last interface of the multilayer. Since ρ_N becomes known, one computes ρ_{N-1} from equation (50). This equation is then applied consecutively to determine $\rho_{N-2}, \rho_{N-3} \cdots\cdots\cdots \rho_o$. If ρ_o is expressed in the form

$$\rho_o = |\rho_o| \; e^{i\theta_o} \,, \qquad (52)$$

then $|\rho_o|$ is amplitude reflectance and θ_o is phase change on reflection of the entire multilayer at the first interface z = 0, Figure 21.4. The phase change on reflection is phase retardation (equivalent to an increase in optical path when $\theta_o > 0$).

21.2.8.7 The complex transmittance τ_{N+1}/τ_o of the system of layers is easily derived as follows. By dividing both sides of equation (48c) by τ_ν one obtains

$$\frac{\tau_\nu}{\tau_{\nu-1}} = \frac{2\, M_{\nu-1}}{M_{\nu-1} + M_\nu} \; \frac{e^{i\frac{\beta\nu}{2}}}{1 + W_\nu \, \rho_\nu \, e^{i\beta\nu}} \,. \qquad (53)$$

Now,

$$\frac{\tau_N}{\tau_o} = \frac{\tau_1}{\tau_o} \; \frac{\tau_2}{\tau_1} \cdots\cdots \frac{\tau_N}{\tau_{N-1}} = \prod_{\nu=1}^{N} \frac{2\, M_{\nu-1}}{M_{\nu-1} + M_\nu} \; \frac{e^{i\frac{\beta\nu}{2}}}{1 + W_\nu \, \rho_\nu \, e^{i\frac{\beta\nu}{2}}} \qquad (53a)$$

wherein $\prod$ denotes a product. Furthermore,

$$\frac{\tau_{N+1}}{\tau_N} = \frac{2\, M_N}{M_N + M_{N+1}} \,, \qquad (53b)$$

the Fresnel coefficient of transmission of the last interface. Therefore

$$\frac{\tau_{N+1}}{\tau_o} = \prod_{\nu=1}^{N+1} \frac{2\, M_{\nu-1}}{M_{\nu-1} + M_\nu} \; \prod_{\nu=1}^{N} \frac{e^{\frac{i\beta\nu}{2}}}{1 + W_\nu \, \rho_\nu \, e^{i\beta\nu}} \,, \qquad (54)$$

where τ_{N+1}/τ_o is the complex transmittance from the left hand side of the first interface to the right hand side of the last interface. We observe that the complex transmittance is not merely the product of the Fresnel coefficients of transmission of the N + 1 interfaces and of a factor that includes the optical path through the layers. Consider, for example, the case in which all $K_\nu = 0$ so that $M_\nu = n_\nu$ and $\beta_\nu = \frac{4\pi}{\lambda} n_\nu d_\nu$. From equation (54)

$$\frac{\tau_{N+1}}{\tau_o} = e^{i \frac{2\pi}{\lambda} \sum_{\nu=1}^{N} n_\nu d_\nu} \; \prod_{\nu=1}^{N+1} \frac{2\, n_{\nu-1}}{n_{\nu-1} + n_\nu} \; \prod_{\nu=1}^{N} \frac{1}{1 + W_\nu \, \rho_\nu \, e^{i\beta\nu}} \qquad (54a)$$

The quantity $\sum_{\nu=1}^{N} n_\nu d_\nu$ is the optical path through the layer system. The product from $\nu = 1$ to $\nu = N + 1$ is the product of the Fresnel coefficients of transmission of the interfaces. The product from $\nu = 1$ to $\nu = N$ is due to interreflections within the multilayer system. Since this product may not be real, the phase change introduced into the wave as it traverses the multilayer is not in general equal to the optical path through the layer. In designing a lens system of high optical quality, it finally becomes necessary to consider the phase changes introduced by transmission through the various layers or multilayers as the number of coated surfaces is increased. Consequently, equation (54) is of interest to both the optical designer and the designer of thin films. We shall see that similar equations hold for oblique incidence.

21.2.8.8 For reasons discussed at the end of Section 21.2.5, $|\rho_o|^2 = |r_o/\tau_o|^2$ is always energy reflectance of the multilayer. Because the incidence is normal, $i_o = i_{N+1} = 0$. Therefore, as in equation (38),

$$\text{Energy transmittance} = \frac{(n_a)_{N+1}}{n_o} \left|\frac{\tau_{N+1}}{\tau_o}\right|^2 \qquad (55)$$

in which τ_{N+1}/τ_o is given by equation (54). Because $p_o = 0$ at normal incidence, a result similar to equation (31) gives $(n_a)_{N+1} = R_e(M_{N+1}) = n_{N+1}$. Consequently,

$$\text{Energy transmittance} = \frac{n_{N+1}}{n_o} \left|\frac{\tau_{N+1}}{\tau_o}\right|^2 \qquad (55a)$$

at normal incidence when the initial medium is non-absorbing. When the initial and final media are identical and non-absorbing, the energy transmittance of the multilayer is simply $|\tau_{N+1}/\tau_o|^2$.

21.2.9 Oblique incidence upon multilayers; the electric vector perpendicular to the plane of incidence.

21.2.9.1 The theory of Section 21.2.8 applies again with minor changes. Let β_ν be written in the more general form

$$\beta_\nu = \frac{4\pi}{\lambda} M_\nu d_\nu , \qquad (56)$$

a result that reduces to equation (47) when $p_o = 0$ (normal incidence). Then,

$$\rho_{\nu-1} = \frac{\rho_\nu e^{i\beta\nu} + W_\nu}{1 + W_\nu \rho_\nu e^{i\beta\nu}} \qquad (57)$$

in which W_ν and M_ν are defined by equations (20) and (19). Because ρ_N is given by equation (51), one can compute $\rho_{N-1}, \rho_{N-2} \cdots\cdots\cdots \rho_o$ consecutively from equation (57) to obtain the complex reflectance ρ_o of the multilayer at the left hand boundary of the first interface, $z = 0$, when the electric vector is perpendicular to the plane of incidence. Furthermore, the complex transmittance, τ_{N+1}/τ_o, from the left hand side of the first interface to the right hand side of the last interface, Figure 21.6, is given again by equation (54). The quantity, $|\rho_o|^2$, is energy reflectance of the entire multilayer. The energy transmittance is given by a result similar to that of equation (38), specifically:

$$\text{Energy transmittance} = \frac{(n_a)_{N+1} \cos i_{N+1}}{n_o \cos i_o} \left|\frac{\tau_{N+1}}{\tau_o}\right|^2 , \qquad (58)$$

in which

$$(n_a)_{N+1} = \left[n_o^2\ p_o^2 + R_e^2(M_{N+1})\right]^{1/2} \qquad (59)$$

and

$$(n_a)_{N+1} \sin i_{N+1} = n_o p_o , \qquad (60)$$

wherein the wave normals (rays) make the angles i_o and i_{N+1} with Z, in the first and last medium, respectively. See Figure 21.6. It has been assumed that the medium of incidence is non-absorbing in writing equation (58). If the first and last media are identical and non-absorbing, the energy transmittance of the multilayer is given by $|\tau_{N+1}/\tau_o|^2$.

21.2.9.2 For some purposes, it is desirable to find the directions of the wave normals in the various layers. When $U = (0, U_y, 0)$ and the time factor is $e^{-i\omega t}$, the wave equation (4) is satisfied in the νth medium or layer by a transmitted wave of the form

$$U_y = \tau_\nu e^{ikm_o p_o x} e^{-ikM_\nu(z - L_\nu)} \qquad (61)$$

provided that M_ν obeys equation (19) where τ_ν specifies the amplitude and phase of the transmitted E-vector at point $x = 0$ in the interface $z = L_\nu$. Equation (61) may be written in the form

$$U_y = \tau_\nu e^{-ik M_\nu L_\nu} e^{-k(n_o K_o p_o x + I_m(M_\nu) z)} e^{ik(n_o p_o x + R_e(M_\nu) z)} \qquad (61a)$$

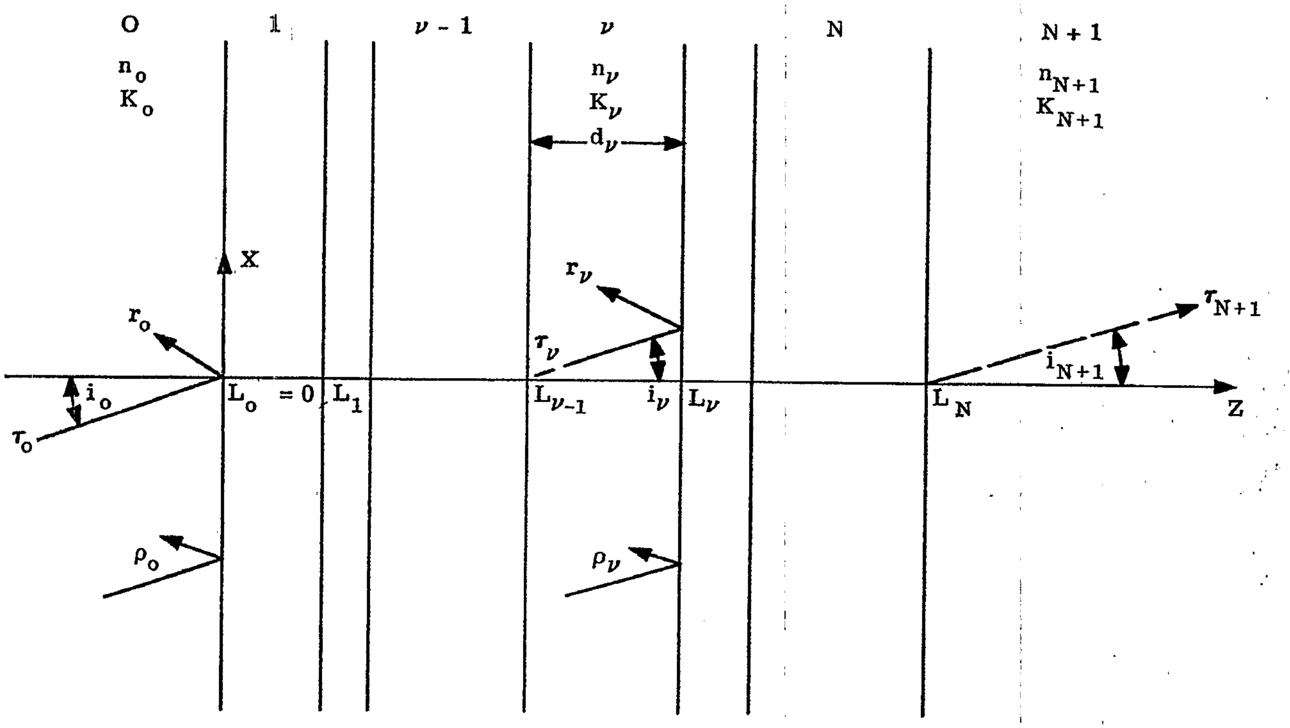


Figure 21. 6- Notation with respect to a multilayer of N films when the incidence is oblique and the E-vector is perpendicular to the plane of incidence. The angles i_ν refer to the directions of the wave normals.

Every equiphase surface in any layer ν is therefore a surface

$$n_o\, p_o\, x + R_e\,(M_\nu)\, z = \text{constant.} \tag{62}$$

If, then, the normals to the equiphase surface make the angle i_ν with Z, it follows directly* from equation (62) that

$$\sin i_\nu = n_o\, p_o / \left[n_o^2\, p_o^2 + R_e^2\,(M_\nu)\right]^{1/2},$$

or that

$$(n_a)_\nu \sin i_\nu = n_o \sin i_o\,, \tag{63}$$

wherein

$$(n_a)_\nu \equiv \left[n_o^2\, p_o^2 + R_e^2\,(M_\nu)\right]^{1/2} \tag{64}$$

is the actual refractive index of the νth layer. Hence the generalized Snell's Law described by equations (63) and (64) holds with respect to all of the media of the system. In particular, equations (59) and (60) are special cases of equations (64) and (63).

21. 2. 9. 3 When required, the electromagnetic field in any layer or medium can be computed as follows. In each medium

$$(E_\nu)_{\text{transmitted}} = (0,\ 1,\ 0)\cdot\tau_\nu\ e^{-i\omega t}\ e^{ikm_o p_o x}\ e^{ikM_\nu(z - L_\nu)}, \tag{65}$$

wherein $E_{\text{transmitted}}$ is the wave propagated to the right, Figure 21. 6. The corresponding H-vector is now computed from equations (27) and stated with the aid of equation (1). With respect to the wave propagated to the

* See discussions of the normal form of the equation of a straight line in textbooks on analytic geometry.

left in each medium

$$(E_\nu)_{\text{reflected}} = (0,\ 1,\ 0)\ r_\nu\ e^{-i\omega t}\ e^{ikm_o p_o x}\ e^{-ikM_\nu(z - L_\nu)}. \qquad (66)$$

The corresponding "reflected" H-vector is computed again from equation (27) and stated with the aid of equation (1). When all ρ_ν's have been determined, each τ_ν can be computed from $\tau_{\nu-1}$ by means of equation (53). Of course, τ_o must be given or assigned. Next, all r_ν's can be calculated from the known ratios $\rho_\nu = r_\nu/\tau_\nu$ and the known values of τ_ν. The theory is therefore a complete theory for the case in which the electric vector is perpendicular to the plane X, Z of incidence, Figure 21. 6.

21. 2. 10 Oblique incidence upon multilayers; the magnetic vector perpendicular to the plane X, Z of incidence.

21. 2. 10. 1 As can be expected, the theory of oblique incidence upon multilayers with the magnetic vector perpendicular to the plane X, Z of incidence differs from the case in which the electric vector is perpendicular to the plane of incidence only in the Fresnel coefficients that become involved. Let

$$\gamma_\nu = R_\nu/T_\nu\ , \qquad (67)$$

where T_ν is a complex number that specifies the amplitude and phase of the H-vector propagated to the right, Figure 21. 7, at x = 0 in the right hand boundary of the νth medium, and where R_ν specifies the amplitude and phase of the H-vector propagated to the left at point x = 0 in the right hand boundary of the νth medium, for the range of ν from $\nu = 0$ to $\nu = N$. $R_{N+1} = 0$ because no reflected wave exists in the last medium. T_{N+1} specifies the amplitude and phase of the transmitted H-vector at the left hand boundary of the (N + 1) th medium. Then

$$\gamma_{\nu-1} = \frac{\gamma_\nu\ e^{i\beta\nu} + F_\nu}{1 + F_\nu\ \gamma_\nu\ e^{i\beta\nu}}\ , \qquad (68)$$

in which β_ν is given by equation (56) and the Fresnel coefficients of reflection F_ν are given by equation (21).

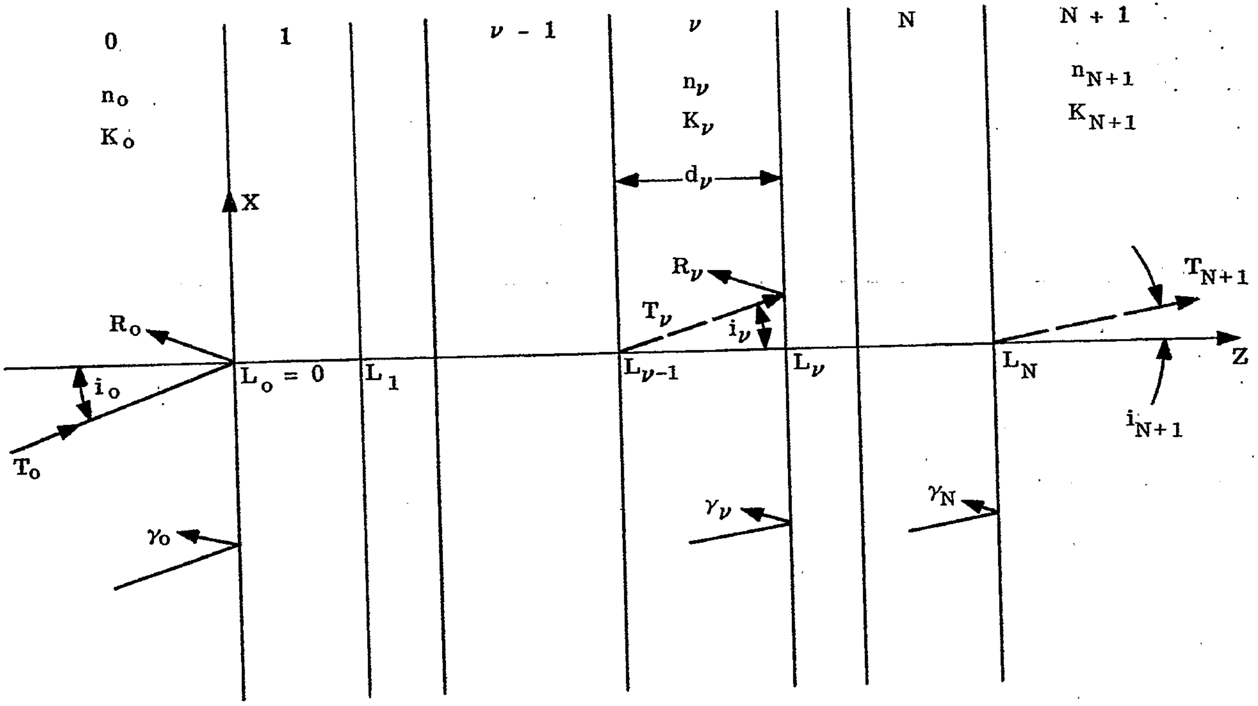


Figure 21. 7- Notation for the case in which the magnetic vector is polarized to vibrate perpendicular to the plane X, Z of incidence. The wave normals have the same angles i_ν with Z as in Figure 21. 5.

Since

$$\gamma_N = F_{N+1}, \qquad (69)$$

the recursion formula (68) enables one to compute consecutively all complex reflectances, γ_ν from γ_N down to γ_o. These complex reflectances, γ_ν, are defined at the right hand boundary of the νth medium, as in the case of the complex reflectances ρ_ν.

21.2.10.2 The complex transmittance, T_{N+1}/T_o, from the left hand boundary of the first interface, z = 0, to the right hand boundary of the last interface, $z = L_N$, is given by

$$\frac{T_{N+1}}{T_o} = \prod_{\nu=1}^{N+1} \frac{2\,M_{\nu-1}\;m_\nu^2}{M_{\nu-1}\;m_\nu^2 + M_\nu\;m_{\nu-1}^2} \prod_{\nu=1}^{N} \frac{e^{\frac{i\beta\nu}{2}}}{1 + F_\nu\,\gamma_\nu\;e^{i\beta\nu}}, \qquad (70)$$

in which $2M_{\nu-1}\,m_\nu^2 / \left[M_{\nu-1}\;m_\nu^2 + M_\nu\;m_{\nu-1}^2\right]$ is the Fresnel coefficient of transmission in the forward direction through the interface between the $(\nu - 1)$th and νth medium. Equation (70) should be compared with equation (54).

21.2.10.3 The energy reflectance is always given in this state of polarization by $|\gamma_o|^2$. Calculation of the energy transmittance involves difficulties, as discussed at the end of Section 21.2.6, unless the time-averaged Poynting vector is practically parallel to the wave normals. If K_{N+1} is so small that the Poynting vector is sensibly parallel to the wave normals in the last medium, then, as in equation (45b),

$$\text{Energy transmittance} = \frac{n_{N+1}^2 n_o (n_a)_{N+1} \cos i_{N+1}}{|m_{N+1}|^4 \cos i_o} \left|\frac{T_{N+1}}{T_o}\right|^2, \qquad (71)$$

in which the medium of incidence is assumed to be non-absorbing. If, also, the last medium is non-absorbing,

$$\text{Energy transmittance} = \frac{n_o \cos i_{N+1}}{n_1 \cos i_o} \left|\frac{T_{N+1}}{T_o}\right|^2. \qquad (71a)$$

21.2.10.4 The electromagnetic field in the νth medium is determined as follows. In each medium

$$(H_\nu)_{\text{transmitted}} = (0,\ 1,\ 0)\ T_\nu\ e^{-i\omega t}\ e^{ikm_o p_o x}\ e^{ikM_\nu(z - L_\nu)}, \qquad (72)$$

for the wave propagated to the right, Figure 21.7. Compare with equation (65). The corresponding E-vector is determined by equation (41) and stated with the aid of equation (1). With respect to the wave propagated to the left in the νth medium,

$$(H_\nu)_{\text{reflected}} = (0,\ 1,\ 0)\ R_\nu\ e^{-i\omega t}\ e^{ikm_o p_o x}\ e^{-ikM_\nu(z - L_\nu)}. \qquad (73)$$

The corresponding reflected E-vector can be computed from equation (41) and stated with the aid of equation (1). When the method of 21.2.8 is applied to the case in which the H-vector is perpendicular to the X, Z plane, one finds instead of equation (53), that

$$\frac{T_\nu}{T_{\nu-1}} = \frac{2\,M_{\nu-1}\;m_\nu^2}{M_{\nu-1}\;m_\nu^2 + M_\nu\;m_{\nu-1}^2} \cdot \frac{e^{i\frac{\beta\nu}{2}}}{1 + F_\nu\,\gamma_\nu\;e^{i\beta\nu}}. \qquad (74)$$

If all values of γ_ν have been computed from equation (68), every T_ν from T_1 to T_N can be computed from equation (74). Since $\gamma_\nu = R_\nu/T_\nu$, every R_ν can be computed so that the coefficients R_ν and T_ν of the electromagnetic fields become known.

21.2.10.5 Comparison of equations (65) and (66) with equations (72) and (73) shows that the corresponding exponential factors are alike. Consequently, the wave normals are refracted from one layer into the next so as to make the same angle, i_ν, with Z whether the E-vector or the H-vector is perpendicular to the plane of incidence.

21.2.11 An approximate method of computation based upon the complex reflectances.

21.2.11.1 Comparison of equations (50), (57) and (68) shows that the recursion formula (50) for normal incidence can be used as the prototype in dealing with the complex reflectances. It is necessary only to enter the appropriate set of Fresnel coefficients, W_ν or F_ν, and to insert the value of $p_o = \sin i_o$ into β_ν as defined by equations (56) and (19). The following approximation simplifies and expedites the determination of ρ_o. The approximation becomes excellent for films that are intended to exhibit low reflectance and have relatively small Fresnel coefficients, W_ν or F_ν. Inspection of equation (50) reveals at once that when both W_ν and

ρ_ν are small, $\rho_{\nu-1}$ should be given with good approximation by

$$\rho_{\nu-1} = \rho_\nu \; e^{i\beta\nu} + W_\nu . \qquad (75)$$

Thus

$$\rho_N = W_{N+1} ,$$

$$\rho_{N-1} = W_{N+1} \; e^{i\beta_N} + W_N ,$$

$$\rho_{N-2} = W_{N+1} \; e^{i(\beta_N + \beta_{N-1})} + W_N \; e^{i\beta_{N-1}} + W_{N-1} ,$$

and therefore,

$$\rho_o = \sum_{\nu=1}^{N+1} W_\nu \exp i \sum_{\mu=1}^{\nu-1} \beta\mu . \qquad (76)$$

When the sum (76) is computed as the complex number

$$\rho_o = R_e (\rho_o) + i \, I_m (\rho_o) , \qquad (76a)$$

$$|\rho_o|^2 = R_e^2 (\rho_o) + I_m^2 (\rho_o) . \qquad (76b)$$

21.2.11.2 This approximate method for computing ρ_o is used mainly during rapid exploration for likely multilayer systems that do not contain absorbing layers and that are intended to produce low reflectance. When absorption is absent, one may compute $|\rho_o|^2$ directly from the sum

$$|\rho_o|^2 = \sum_{\nu=1}^{N+1} W_\nu + 2 \sum_{\mu=1}^{N} \sum_{\nu=2}^{N+1} W\mu \; W_\nu \; \cos(\beta_\mu + \beta_{\mu+1} + \cdots\cdots \beta_{\nu-1}), \qquad (76c)$$

in which $\mu < \nu$. Comparison of $|\rho_o|_a^2$ computed from the approximate equation (76c) with the values $|\rho_o|^2$ computed from equation (50) is made in Figure 21.8 for the case of a low reflecting trilayer. The Fresnel coefficients W_2 and W_3 are quite large numerically.

21.2.11.3 The approximate method of equation (76) is the algebraic equivalent of the graphical polygon method used in early calculations on mono and bilayers by C. H. Cartwright [2] and others.

21.2.12 Method of admittances; E-vector perpendicular to the plane of incidence.

21.2.12.1 Because the theories of multilayers and transmission lines are similar, many investigators prefer to treat a multilayer as a transmission line. One of the earlier publications dealing with multilayers in terms of the admittances of transmission lines is due to B. Salzberg [3]. It will be one aim of the following presentation to unify and to show the relationships that exist between the optical method of the reflectances and the electrical method of the admittances. These two methods are equivalent and complementary. Each method possesses some advantages over the other in dealing with thin films.

21.2.12.2 From equations (65) and (66)

$$(E_{y,\nu})_{\text{transmitted}} = \tau_\nu \; e^{-i\omega t} \; e^{ikm_o p_o x} \; e^{ikM_\nu (z - L_\nu)} , \qquad (77)$$

and

$$(E_{y,\nu})_{\text{reflected}} = r_\nu \; e^{-i\omega t} \; e^{ikm_o p_o x} \; e^{-ikM_\nu (z - L_\nu)} , \qquad (77a)$$

wherein the subscript ν refers to the wave in the νth medium or layer. Correspondingly from equations (27) and (1),

$$(H_{x,\nu})_{\text{transmitted}} = - \tau_\nu \; M_\nu \; e^{-i\omega t} \; e^{ikm_o p_o x} \; e^{ikM_\nu (z - L_\nu)} , \qquad (77b)$$

and

$$(H_{x,\nu})_{\text{reflected}} = r_\nu \; M_\nu \; e^{-i\omega t} \; e^{ikm_o p_o x} \; e^{-ikM_\nu (z - L_\nu)} . \qquad (77c)$$

[2] C. H. Cartwright et al, U. S. Patent 2, 281, 474.

[3] Bernard Salzberg, J. Opt. Soc. Amer., 40, 465-470 (1950).

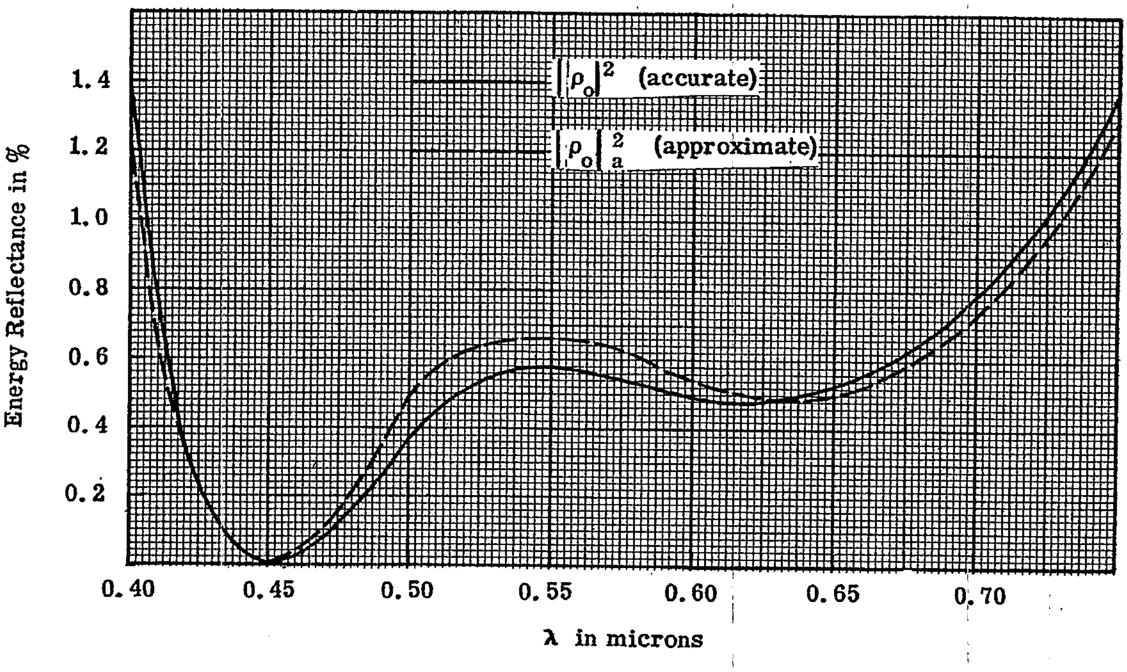


Figure 21. 8- Comparison of the computed energy reflectances $|\rho_o|^2$ and $|\rho_o|_a^2$ for a tri-layer in which $W_1 = -0.15966$; $W_2 = -0.28866$; $W_3 = 0.18765$ and $W_4 = 0.05882$. At the wavelength $\lambda_o = 0.55$ microns, $\beta_1 = 235.8^o$; $\beta_2 = 32^o$ and $\beta_3 = 360^o$.

The admittance Y_ν for the case in which the electric vector is perpendicular to the plane of incidence is defined so that

$$Y_\nu = \frac{(H_{x,\nu})_{\text{transmitted}} + (H_{x,\nu})_{\text{reflected}}}{(E_{y,\nu})_{\text{transmitted}} + (E_{y,\nu})_{\text{reflected}}}, \tag{78}$$

evaluated at the left hand boundary, $z = L_{\nu-1}$, of the νth medium or layer. Since $L_{\nu-1} - L_\nu = - d_\nu$, equations (77) through (78) give almost directly the result

$$Y_\nu = M_\nu \frac{-\tau_\nu \; e^{-i\frac{\beta\nu}{2}} + r_\nu \; e^{i\frac{\beta\nu}{2}}}{\tau_\nu \; e^{-i\frac{\beta\nu}{2}} + r_\nu \; e^{i\frac{\beta\nu}{2}}}, \tag{78a}$$

or

$$Y_\nu = M_\nu \frac{-1 + \rho_\nu \; e^{i\beta\nu}}{1 + \rho_\nu \; e^{i\beta\nu}}, \tag{78b}$$

because $\rho_\nu = r_\nu/\tau_\nu$. Upon solving equation (78b) for $\rho_\nu \; e^{i\beta\nu}$, one obtains

$$\rho_\nu \; e^{i\beta\nu} = \frac{M_\nu + Y_\nu}{M_\nu - Y_\nu}. \tag{78c}$$

Hence the admittances Y_ν can be computed from the reflectances ρ_ν and vice versa.

From equation (57),

$$\rho_o = \frac{\rho_1 \; e^{i\beta_1} + W_1}{1 + W_1 \; \rho_1 \; e^{i\beta_1}}. \tag{79}$$

Upon eliminating $\rho_1 \; e^{i\beta_1}$ with the aid of equation (78c) and making use of the identity

$$M_\nu \; \frac{1 + W_\nu}{1 - W_\nu} = M_{\nu-1}, \tag{80}$$

one finds that the complex reflectance, ρ_o , of the multilayer, and the admittance, Y_1 , are connected by the equation

$$\rho_o = \frac{M_o + Y_1}{M_o - Y_1} . \tag{79a}$$

From equation (78b),

$$Y_{\nu-1} = M_{\nu-1} \frac{\rho_{\nu-1} e^{i\beta_{\nu-1}} -1}{\rho_{\nu-1} e^{i\beta_{\nu-1}} +1} . \tag{81}$$

Just as equation (79a) is obtained from equation (79),

$$\rho_{\nu-1} = \frac{M_{\nu-1} + Y_\nu}{M_{\nu-1} - Y_\nu} . \tag{82}$$

By eliminating $\rho_{\nu-1}$ from equation (81) with the aid of equation (82) and utilizing the identity

$$\frac{e^{-ix} - 1}{e^{-ix} + 1} = - \tan h \frac{ix}{2} = - i \tan \frac{x}{2} ,$$

one obtains without difficulty the recursion formula for $Y_{\nu-1}$ in the form

$$Y_{\nu-1} = M_{\nu-1} \frac{Y_\nu - i M_{\nu-1} \tan (\beta_{\nu-1}/2)}{M_{\nu-1} - i Y_\nu \tan (\beta_{\nu-1}/2)} . \tag{81a}$$

Because $\rho_{N+1} = 0$ (no reflected wave in the last medium), it follows from equation (78c) as a boundary condition that

$$Y_{N+1} = - M_{N+1} . \tag{81b}$$

Hence the recursion formula (81a) enables one to compute all Y_ν's from Y_N down to Y_1 . With Y_1 thus determined, equation (79a) can be applied to find the complex reflectance, ρ_o , of the multilayer.

21.2.12.3 The complex transmittance, τ_{N+1}/τ_o , of the multilayer is now given awkwardly by equation (54). From equation (78c),

$$1 + W_\nu \rho_\nu e^{i\beta\nu} = 1 + W_\nu \left(\frac{M_\nu + Y_\nu}{M_\nu - Y_\nu}\right) = \frac{M_\nu (1 + W_\nu) - Y_\nu (1 - W_\nu)}{M_\nu - Y_\nu} ;$$

$$= (1 - W_\nu) \frac{M_{\nu-1} - Y_\nu}{M_\nu - Y_\nu} . \tag{82a}$$

Since

$$1 - W_\nu = 2 M_\nu / (M_{\nu-1} + M_\nu) , \tag{82b}$$

$$1 + W_\nu \rho_\nu e^{i\beta\nu} = \frac{2 M_\nu}{M_{\nu-1} + M_\nu} \frac{M_{\nu-1} - Y_\nu}{M_\nu - Y_\nu} . \tag{82c}$$

Let $1 + W_\nu \rho_\nu e^{i\beta\nu}$ be eliminated from equation (54) with the aid of equation (82c). Then

$$\frac{\tau_{N+1}}{\tau_o} = \frac{2 M_N}{M_N + M_{N+1}} \exp i \left(\sum_{\nu=1}^{N} \frac{\beta_\nu}{2} \right) \prod_{\nu=1}^{N} \left[\left(\frac{M_{\nu-1}}{M_\nu} \right) \frac{M_\nu - Y_\nu}{M_{\nu-1} - Y_\nu} \right] , \tag{83}$$

a result that should be compared with equation (54).

21.2.12.4 In summary, when the electric vector is perpendicular to the plane X, Z of incidence, the complex reflectance ρ_o and the complex transmittance τ_{N+1}/τ_o of the multilayer can be computed from equations (79a) and (83) in which all the admittances, Y_ν , from Y_N down to Y_1 are determined by the recursion formula (81a). The admittance, Y_ν , at the point of entry into the ν th layer is defined by equations (78) or (78a). β_ν and M_ν are defined by equations (56) and (19), respectively. The results for normal incidence are obtained by setting $p_o = 0$ in equation (19), i.e. by setting $M_\nu = m_\nu = n_\nu (1 + i K_\nu)$.

21. 2. 13 Method of admittances; H-vector perpendicular to the plane of incidence.

21. 2. 13. 1 When the H-vector is polarized to vibrate at right angles to the plane X, Z of incidence, equations (72) and (73) show that

$$(H_{y,\nu})_{\text{transmitted}} = T_\nu \, e^{-i\omega t} \, e^{ikm_o p_o x} \, e^{ikM_\nu (z - L_\nu)} , \qquad (84)$$

and

$$(H_{y,\nu})_{\text{reflected}} = R_\nu \, e^{-i\omega t} \, e^{ikm_o p_o x} \, e^{-ikM_\nu (z - L_\nu)} . \qquad (84a)$$

Correspondingly, from equations (41) and (1),

$$(E_{x,\,\nu})_{\text{transmitted}} = \frac{M_\nu}{m_\nu^2} \, T_\nu \, e^{-i\omega t} \, e^{ikm_o p_o x} \, e^{ikM_\nu (z - L_\nu)} , \qquad (84b)$$

and

$$(E_{x,\,\nu})_{\text{reflected}} = \frac{-M_\nu}{m_\nu^2} \, R_\nu \, e^{-i\omega t} \, e^{ikm_o p_o x} \, e^{-ikM_\nu (z - L_\nu)} . \qquad (84c)$$

The admittances are designated by y_ν to distinguish them from the admittances Y_ν of equation (78) and are defined by the equation

$$y_\nu = \frac{(H_{y,\,\nu})_{\text{transmitted}} + (H_{y,\,\nu})_{\text{reflected}}}{(E_{x,\,\nu})_{\text{transmitted}} + (E_{x,\,\nu})_{\text{reflected}}} , \qquad (85)$$

evaluated at the point of entry, $z = L_{\nu-1}$, of the νth layer. Substitution of equation (84) into equation (85) yields the result

$$y_\nu = \frac{m_\nu^2}{M_\nu} \; \frac{1 + \gamma_\nu \, e^{i\beta\nu}}{1 - \gamma_\nu \, e^{i\beta\nu}} , \qquad (85a)$$

in which $\gamma_\nu = R_\nu / T_\nu$ as in equation (67). It follows from equation (85a) that

$$\gamma_\nu \, e^{i\beta\nu} = \frac{y_\nu - m_\nu^2 / M_\nu}{y_\nu + m_\nu^2 / M_\nu} . \qquad (85b)$$

21. 2. 13. 2 Corresponding to the identity (80), one finds from the definition of F_ν, equation (21), that

$$\frac{m_\nu^2}{M_\nu} \; \frac{1 - F_\nu}{1 + F_\nu} = \frac{m_{\nu-1}^2}{M_{\nu-1}} . \qquad (86)$$

Let $\gamma_\nu \, e^{i\beta\nu}$ be eliminated from equation (68) with the aid of equation (85b). By utilizing the identity (86) in the result thus obtained, one finds straightforwardly that

$$\gamma_{\nu-1} = \frac{y_\nu - m_{\nu-1}^2 / M_{\nu-1}}{y_\nu + m_{\nu-1}^2 / M_{\nu-1}} . \qquad (87)$$

In particular,

$$\gamma_o = \frac{y_1 - m_o^2 / M_o}{y_1 + m_o^2 / M_o} , \qquad (87a)$$

where γ_o is the complex reflectance of the entire multilayer evaluated at the first interface, $z = 0$, Figure 21. 7.

21. 2. 13. 3 From equation (85a),

$$y_{\nu-1} = \frac{m_{\nu-1}^2}{M_{\nu-1}} \; \frac{1 + \gamma_{\nu-1} \, e^{i\beta_{\nu-1}}}{1 - \gamma_{\nu-1} \, e^{i\beta_{\nu-1}}} \qquad (88)$$

Let $\gamma_{\nu-1}$ be eliminated from equation (88) with the aid of equation (87). As in the steps leading to equation

(81a), one finds that

$$y_{\nu-1} = \frac{m^2_{\nu-1}}{M_{\nu-1}} \left[\frac{y_\nu - i \, \frac{m^2_{\nu-1}}{M_{\nu-1}} \tan(\beta_{\nu-1}/2)}{\frac{m^2_{\nu-1}}{M_{\nu-1}} - i \, y_\nu \tan(\beta_{\nu-1}/2)} \right] . \qquad (88a)$$

This recursion formula is similar to equation (81a). It is necessary only to replace $M_{\nu-1}$ by the ratio $m^2_{\nu-1}/M_{\nu-1}$. Since $\gamma_{N+1} = 0$, equation (85b) shows that a boundary condition on y_ν is

$$y_{N+1} = m^2_{N+1}/M_{N+1} . \qquad (88b)$$

Equations (88a) and (88b) permit all values of y_ν to be computed from y_N down to y_1 from the optical properties of the multilayer system. At this point, the solution for the complex reflectance, γ_o, of the multilayer can be completed from equation (87a).

21.2.13.4 It remains to develop a more suitable formula to replace equation (70) for the complex transmittance T_{N+1}/T_o of the multilayer. From equation (85b),

$$1 + F_\nu \gamma_\nu e^{i\beta_\nu} = 1 + F_\nu \left(\frac{y_\nu - m^2_\nu/M_\nu}{y_\nu + m^2_\nu/M_\nu} \right) = \frac{y_\nu (1 + F_\nu) + \frac{m^2_\nu}{M_\nu} (1 - F_\nu)}{y_\nu + m^2_\nu/M_\nu}$$

$$= (1 + F_\nu) \frac{y_\nu + m^2_{\nu-1}/M_{\nu-1}}{y_\nu + m^2_\nu/M_\nu} ; \qquad (89)$$

with

$$1 + F_\nu = \frac{2 m^2_\nu M_{\nu-1}}{m^2_\nu M_{\nu-1} + m^2_{\nu-1} M_\nu} . \qquad (89a)$$

Let $1 + F_\nu \gamma_\nu e^{i\beta\nu}$ be eliminated from equation (70) with the aid of equations (89) and (89a). Then

$$\frac{T_{N+1}}{T_o} = \frac{2 M_N m^2_{N+1}}{M_N m^2_{N+1} + M_{N+1} m^2_N} \exp \left(i \sum_{\nu=1}^{N} \frac{\beta_\nu}{2} \right) \prod_{\nu=1}^{N} \frac{y_\nu + m^2_\nu/M_\nu}{y_\nu + m^2_{\nu-1}/M_\nu} . \qquad (89b)$$

21.2.13.5 In summary: When the magnetic vector vibrates at right angles to the plane X, Z of incidence, the complex reflectance γ_o and the complex transmittance T_{N+1}/T_o of the multilayer are determined by equations (87a) and (89b) when the admittances, y_ν, from y_N to y_1 have been computed from the recursion formula (88a). The admittances, y_ν, are defined by equations (85) or (85a), and are different from the complementary admittances, Y_ν, for the state of polarization in which the electric vector is perpendicular to the plane X, Z of incidence. β_ν and M_ν are defined by equations (56) and (19). When desired, the admittances, y_ν, can be computed from the complex reflectances, γ_ν, by means of equation (85a); or the complex reflectances, γ_ν, can be computed from the admittances, y_ν, by means of equation (85b).

21.2.14 Absentee layers.

21.2.14.1 It can be shown easily from the method of admittances that non-absorbing layers behave as if they were absent at wavelengths λ for which

$$n_\nu \, d_\nu \cos i_\nu = \mu \frac{\lambda}{2} ; \quad \mu = 1, 2, 3, 4, \text{etc.} \qquad (90)$$

We shall call such layers absentee layers. For example, at normal incidence the so called half-wave layer, the case $\mu = 1$ and $i_\nu = 0$ in equation (90), is an absentee layer. Condition (90) is satisfied when $\beta_\nu = \mu \, 2\pi$, i.e. when $\tan(\beta_\nu/2) = 0$ in equations (81a) and (88a). Consequently, $Y_{\nu-1} = Y_\nu$ and $y_{\nu-1} = y_\nu$ when equation (90) is satisfied. This means that the νth layer does not affect the admittance at the point of entry of the $(\nu-1)$th layer whether the E-vector or the H-vector vibrates at right angles to the plane X, Z of incidence. The νth layer does not influence the reflectance or the transmittance of the multilayer at any wavelength for which equation (90) is satisfied. In fact, the layer behaves as if $\mu = 0$, i.e. as if the thickness d_ν of the layer were zero.

21.2.14.2 Consider the behavior, with wavelength λ_μ or with β_ν, of a single, homogeneous, non-absorbing film deposited on any substrate. Because this film is an absentee layer at wavelengths λ_μ for which equation (90) is satisfied, it follows, for example, that the energy reflectances $|\rho_o|^2$ of the coated and uncoated surface should be alike at the wavelengths λ_μ. Actually, these reflectances are not quite alike at λ_μ because the film may absorb, may not be homogeneous,or may scatter appreciably.

21. 2. 14. 3 Absentee layers are often introduced in multilayers for the purpose of altering the behavior of the multilayer at wavelengths λ that do not satisfy equation (90).

21. 2. 15 The Q-Method.

21. 2. 15. 1 Comparison of the recursion formulae (57), (68), (81a) and (88a) shows that they are all awkward for the purposes of computation. With respect to equation (57), for example, $\rho_{\nu-1}$ is not linear in ρ_ν. This lack of linearity applies to the reflectances and the admittances alike. The reflectances enjoy a small advantage in that approximate expressions such as equation (75) exhibit linearity. It is one of the purposes of the Q-method [4] to circumvent the lack of linearity in the recursion formulae for the reflectances and the admittances.

21. 2. 15. 2 The recurrence formulae connecting successive interfacial reflectances or admittances are of the form

$$A_{\nu-1} = \frac{a_\nu A_\nu + b_\nu}{g_\nu A_\nu + h_\nu} , \qquad (91)$$

in which A_ν can represent ρ_ν, γ_ν, Y_ν or y_ν. Since the denominators of (91) are not zero, let

$$f_\nu \equiv g_\nu A_\nu + h_\nu , \qquad (91a)$$

and set

$$Q_\nu \equiv \prod_{n=\nu}^{n=m+1} f_n , \qquad (91b)$$

in which the upper limit, m, of the product is an integer such that

$$A_{m+1} = 0 ; \quad A_m = 0 . \qquad (91c)$$

Let

$$\gamma_\nu = a_\nu / g_\nu . \qquad (91d)$$

Then it can be shown that Q_ν obeys the linear recursion formula

$$Q_\nu = (\alpha_{\nu+1} g_\nu + h_\nu) Q_{\nu+1} + (b_{\nu+1} - \alpha_{\nu+1} h_{\nu+1}) g_\nu Q_{\nu+2} . \qquad (91e)$$

Furthermore,

$$A_\nu = \left[\alpha_{\nu+1} Q_{\nu+1} + (b_{\nu+1} - \alpha_{\nu+1} h_{\nu+1}) Q_{\nu+2}\right] / Q_{\nu+1} . \qquad (91f)$$

21. 2. 15. 3 Consider, for example, the application of the Q-method to the determination of the complex reflectances $\rho_{\nu-1}$ of equation (57). Thus,

$$A_{\nu-1} = \rho_{\nu-1} = \frac{\rho_\nu e^{i\beta\nu} + W_\nu}{W_\nu \rho_\nu e^{i\beta\nu} + 1} . \qquad (92)$$

Comparison of equations (91) and (92) shows that

$$a_\nu = e^{i\beta\nu} ; \quad g_\nu = W_\nu e^{i\beta\nu} ;$$
$$b_\nu = W_\nu ; \quad h_\nu = 1 . \qquad (92a)$$

Therefore,

$$\alpha_\nu = 1/W_\nu ; \qquad (92b)$$

$$f_\nu = 1 + W_\nu \rho_\nu e^{i\beta\nu} . \qquad (92c)$$

[4] H. Osterberg, J. Opt. Soc. Amer., 43, 728-732 (1953).

Because $\rho_{N+1} = 0$ in a system of multilayers containing N layers, $A_{N+1} = 0$. Hence $m = N$ in equation (91b) so that

$$Q_\nu = \prod_{n=\nu}^{N+1} f_n = \prod_{n=\nu}^{N+1} (1 + W_n \rho_n e^{i\beta n}). \tag{92d}$$

In particular,

$$Q_{N+1} = 1 ; \tag{92e}$$

$$Q_N = 1 + W_N W_{N+1} e^{i\beta_N} ; \rho_N = W_{N+1} . \tag{92f}$$

From equations (91e), (92a) and (92b),

$$Q_\nu = Q_{\nu+1} + \left[Q_{\nu+1} + (W_{\nu+1}^2 - 1) Q_{\nu+2} \right] \frac{W_\nu e^{i\beta\nu}}{W_{\nu+1}} . \tag{92g}$$

Every Q_ν from $\nu = N + 1$ down to $\nu = 1$ is determined from equations (92e) through (92g). Since $A_\nu = \rho_\nu$, equations (91f), (92a) and (92b) now yield ρ_ν in the form

$$\rho_\nu = \frac{Q_{\nu+1} + (W_{\nu+1}^2 - 1) Q_{\nu+2}}{W_{\nu+1} Q_{\nu+1}} . \tag{92h}$$

The complex reflectance, ρ_o, of the entire multilayer is therefore given by

$$\rho_o = \frac{Q_1 + (W_1^2 - 1) Q_2}{W_1 Q_1} . \tag{93}$$

21.2.15.4 With respect to the computation of the complex transmittance, τ_{N+1}/τ_o, of the multilayer, from equation (54), we note that

$$\prod_{\nu=1}^{N} \frac{1}{1 + W_\nu \rho_\nu e^{i\beta\nu}} = f_{N+1} \prod_{\nu=1}^{N+1} \frac{1}{f_\nu} = \frac{1}{Q_1} , \tag{94}$$

because $f_{N+1} = 1$. Hence the complex transmittance, τ_{N+1}/τ_o, of the multilayer is given by the comparatively simple result,

$$\frac{\tau_{N+1}}{\tau_o} = \frac{2^{N+1}}{Q_1} \exp\left(i \sum_{\nu=1}^{N} \frac{\beta_\nu}{2} \right) \prod_{\nu=1}^{N+1} \frac{M_{\nu-1}}{M_{\nu-1} + M_\nu} , \tag{95}$$

when the electric vector is perpendicular to the plane of incidence. Since β_ν is defined by equation (56),

$$\frac{\beta_\nu}{2} = \frac{2\pi d_\nu}{\lambda} \left[R_e (M_\nu) + i \, I_m (M_\nu) \right] . \tag{95a}$$

21.2.15.5 Suppose that the system has no absorption. Then $M_\nu = n_\nu \cos i_\nu$, a real number, and $\frac{\beta_\nu}{2} = \frac{2\pi}{\lambda} n_\nu d_\nu \cos i_\nu$. It will be seen from Figure 21.6 that $\frac{\beta_\nu}{2}$ is the optical path along the rays only when the incidence is normal so that $\cos i_\nu = 1$. If the incidence is normal, and if the system has no absorption, $\sum_{\nu=1}^{N} \frac{\beta_\nu}{2}$ is physically the optical path through the multilayer. But the optical path will not be equal to the phase change suffered by the wave on passing through the multilayer system unless Q_1 is real, a condition that holds only at certain wavelengths for a given multilayer.

21.2.15.6 The phase change (retardation) suffered by the wave in passing through the multilayer is equal to $\arg (\tau_{N+1}/\tau_o)$. The portion determined by $\arg (\bar{Q}_1) = \arg (1/Q_1)$ can be oscillatory with wavelength and is therefore quite dispersive. We learn that the quantity Q_1 has definite physical significance and that the Q-method is not merely a more convenient method for computing reflectance and transmittance of multilayers.

21.2.15.7 Examination of equation (92g) shows that when every β_ν is an integral multiple of the smallest value of β, each Q_ν will be a terminating exponential series of the Fourier type. Hence the powerful methods of Fourier series may be brought to bear. In the simplest case, each layer can be a quarter-wave layer at a specified wavelength. Furthermore, equation (92g) is a difference equation, consequently the methods of difference equations can often be utilized to simplify the determination of Q_1.

21. 2. 16 The zero condition.

21. 2. 16. 1 One of the most important applications of single or multilayers is to reduce the energy reflectance of the coated surface. Whereas it is not always practical to attain zero reflectance, the designer of low reflecting films attempts to achieve zero reflectance at one or more wavelengths. Unfortunately, it is not possible to obtain zero reflectance over an extended range of wavelengths. Each method of attack on the design of thin films will include an analytical statement of the zero condition, i. e. the condition for obtaining zero reflectance.

21. 2. 16. 2 With respect to the method involving the interfacial reflectances ρ_ν or γ_ν, the zero condition requires that ρ_o or $\gamma_o = 0$ as indicated in Figure 21. 9. From equation (57),

$$\rho_o = \frac{\rho_1 e^{i\beta_1} + W_1}{1 + W_1 \rho_1 e^{i\beta_1}} , \tag{96}$$

and from equation (68),

$$\gamma_o = \frac{\gamma_1 e^{i\beta_1} + F_1}{1 + F_1 \gamma_1 e^{i\beta_1}} . \tag{97}$$

Accordingly, the zero condition assumes the form

$$\rho_1 e^{i\beta_1} = -W_1 , \tag{98}$$

or

$$\gamma_1 e^{i\beta_1} = -F_1 , \tag{98a}$$

depending upon whether the electric or the magnetic vector, respectively, is perpendicular to the plane of incidence. In considering normal incidence, one ordinarily chooses equation (98) and sets $i_o = 0$ in determining M_ν.

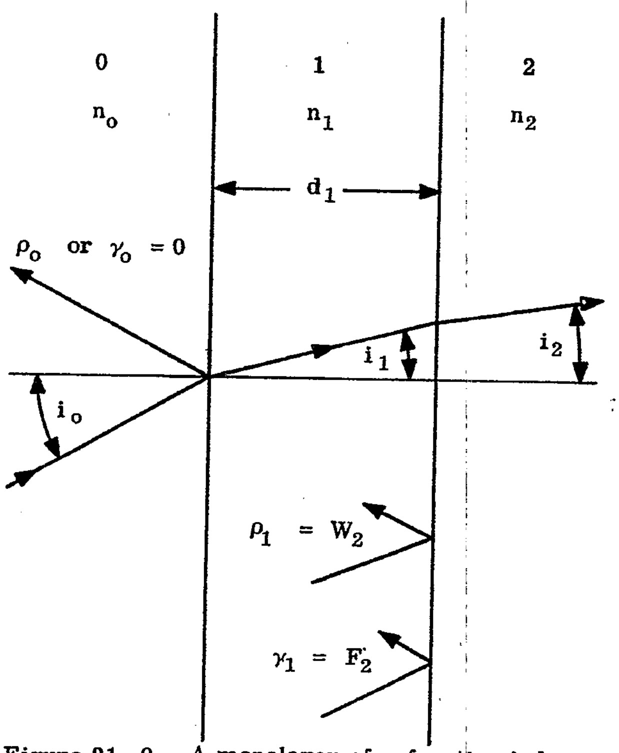


Figure 21. 9- A monolayer of refractive index n_1 and thickness d_1 between two media having refractive indices n_o and n_2.

21.2.16.3 In dealing with the admittances Y_ν and y_ν, equations (79a) and (87a) show that the zero condition is

$$Y_1 = - M_o , \quad (99)$$

or

$$y_1 = m_o^2/M_o , \quad (99a)$$

according as the electric vector or the magnetic vector is perpendicular to the plane of incidence.

21.2.16.4 With respect to the Q-method, equation (93) shows that the zero condition is $Q_1 + (W_1^2 -1) Q_2 = 0$, or

$$Q_1 = (1 - W_1^2) Q_2 , \quad (100)$$

when the electric vector is perpendicular to the plane of incidence.

21.3 ZERO REFLECTANCE FROM NON-ABSORBING MONOLAYERS AND SUBSTRATES

21.3.1 Introduction. The problem is to design a film that produces zero reflectance. The variables of the film are its refractive index, n_1, and its thickness, d_1. The usual restriction of the discussion to normal incidence will not be made because this restriction avoids too many pertinent and practical facts associated with oblique incidence. The following discussion will hinge upon the method of the complex reflectances. The complex reflectances ρ_o and γ_o are given by equations (96) and (97), respectively, with

$$\rho_1 = W_2 , \text{ and } \gamma_1 = F_2 . \quad (101)$$

Equation (101) applies to monolayers, i.e. to cases $N = 1$. From equation (96), the energy reflectance, $|\rho_o|^2$, is given by

$$|\rho_o|^2 = \frac{W_1\bar{W}_1 + W_2 \bar{W}_2 \exp(i\beta_1) \overline{\exp(i\beta_1)} + W_1 \bar{W}_2 \overline{\exp(i\beta_1)} + \bar{W}_1 W_2 \exp(i\beta_1)}{1 + W_1 \bar{W}_1 W_2 \bar{W}_2 \exp(i\beta_1) \overline{\exp(i\beta_1)} + W_1 W_2 \exp(i\beta_1) + \bar{W}_1 \bar{W}_2 \overline{\exp(i\beta_1)}} . \quad (102)$$

Similarly, from equations (101) and (97)

$$\gamma_o{}^2 = \frac{F_1 \bar{F}_1 + F_2 \bar{F}_2 \exp(i\beta_1) \overline{\exp(i\beta_1)} + F_1 \bar{F}_2 \overline{\exp(i\beta_1)} + \bar{F}_1 F_2 \exp(i\beta_1)}{1 + F_1 \bar{F}_1 F_2 \bar{F}_2 \exp(i\beta_1) \overline{\exp(i\beta_1)} + F_1 F_2 \exp(i\beta_1) + \bar{F}_1 \bar{F}_2 \overline{\exp(i\beta_1)}} . \quad (103)$$

21.3.2 Total internal reflection with M_2 pure imaginary. Let us consider first the class of cases in which $n_o p_o > n_2$ but in which n_1 is chosen so that $n_o p_o < n_1$. Then according to equation (19), M_1 is real but M_2 is pure imaginary. Consequently, from equations (20) and (21), W_1 and F_1 are real and W_2 and F_2 are complex such that

$$W_2 \bar{W}_2 = 1 ; \quad F_2 \bar{F}_2 = 1 . \quad (104)$$

Furthermore, $\beta_1 = 4\pi M_1 d_1/\lambda$ will be real. Equation (102) assumes now the simplified form

$$\rho_o{}^2 = \frac{1 + W_1 \bar{W}_1 + 2|W_1||W_2| \cos[\beta_1 + \arg(W_2) - \arg(W_1)]}{1 + W_1 \bar{W}_1 + 2|W_1||W_2| \cos[\beta_1 + \arg(W_2) + \arg(W_1)]} . \quad (105)$$

Because W_1 is real, $\arg(W_1) = 0$ and $|\rho_o|^2 = 1$. Similarly, $|\gamma_o|^2 = 1$ irrespective of the value of β_1. If n_1 is chosen so that $n_1^2 > n_o^2 p_o^2$, the film cannot alter the energy reflectance of the coated interface and the total reflection remains complete irrespective of the state of polarization of the incident beam. On the other hand, the phase change on reflection can be modified.

21.3.3 Total internal reflection with both M_1 and M_2 pure imaginary. Consider next the class of cases in which $n_o p_o$ exceeds both n_1 and n_2. Both M_1 and M_2 are then pure imaginary. Equations (20) and (21) now show that W_2 and F_2 are real but that W_1 and F_2 are complex such that

$$W_1 \bar{W}_1 = 1 ; \quad F_1 \bar{F}_1 = 1 . \quad (106)$$

Furthermore, since $\beta_1 = 4\pi M_1 d_1/\lambda$,

$$\exp(i\beta_1) = \overline{\exp(i\beta_1)} = \exp(-4\pi\,|n_o^2\,p_o^2 - n_1^2|^{1/2}\;d_1/\lambda), \qquad (107)$$

an attenuation factor that we shall designate temporarily by A. From equations (102), (106) and (107)

$$\rho_o{}^2 = \frac{1 + W_2^2 A^2 + 2A\,|W_1|\,W_2 \cos\left[\arg(W_1) - \arg(W_2)\right]}{1 + W_2^2 A^2 + 2A\,|W_1|\,W_2 \cos\left[\arg(W_1) + \arg(W_2)\right]} . \qquad (108)$$

Since W_2 is real, $\arg(W_2) = 0, \pi, 2\pi$, etc. Hence $|\rho_o|^2 = 1$ irrespective of the thickness of the film. A similar conclusion holds when the magnetic vector is perpendicular to the plane of incidence. If i_o is chosen so that $n_o \sin i_o$ exceeds n_1 and n_2, the film cannot modify the energy reflectance of the coated interface.

21.3.4 Total internal reflectance when M_1 is pure imaginary. Total internal reflection may or may not occur when $n_o \sin i_o$ exceeds n_1 but not n_2. In this class of cases, M_1 is pure imaginary and M_2 is real. It can be seen from equations (20) and (21) that the Fresnel coefficients of reflection are complex such that

$$W_1\,\overline{W}_1 = 1;\; W_2\,\overline{W}_2 = 1;\; F_1\,\overline{F}_1 = 1;\; F_2\,\overline{F}_2 = 1. \qquad (109)$$

In addition, β_1 obeys equation (107). Equation (102) assumes the form

$$|\rho_o|^2 = \frac{1 + A^2 + A\,(W_1\,\overline{W}_2 + \overline{W}_1\,W_2)}{1 + A^2 + A\,(W_1\,W_2 + \overline{W}_1\,\overline{W}_2)} , \qquad (110)$$

in which $A = \exp(i\beta_1)$, a real attenuation factor. Because $A \to 0$ as $d_1 \to \infty$, $|\rho_o|^2 \to 1$ as the thickness d_1 of the film is increased. In fact, A falls very rapidly with increasing d_1. A relatively thin film can therefore produce almost total * internal reflection. One can show that,subject to equation (109),

$$W_1\,\overline{W}_2 + \overline{W}_1\,W_2 \leqq W_1\,W_2 + \overline{W}_1\,\overline{W}_2 . \qquad (111)$$

Hence $|\rho_o|^2 \leqq 1$.

21.3.5 Zero reflectance; the E-vector perpendicular to the plane of incidence. Zero reflectance is possible with monolayers when $n_o \sin i_o$ is less than n_1 or n_2, i.e. when both M_1 and M_2 are real. Since $\rho_1 = W_2$ for monolayers (case $N = 1$), the zero condition of equation (98) becomes

$$W_2\,e^{i\beta_1} = -W_1 . \qquad (112)$$

Because W_1 and W_2 are real, equation (112) requires that $\exp(i\beta_1)$ be real. Two choices are possible

$$\beta_1 = \begin{cases} \mu\,\pi;\ \mu \text{ an odd integer}; & (113)\\ \nu\,2\pi;\ \nu \text{ any integer}; & (113a)\end{cases}$$

$$\beta_1 = \frac{4\pi}{\lambda}\,n_1\,d_1\,\cos i_1 . \qquad (113b)$$

We shall restrict our attention to the more interesting and important**choice (113). For all choices of μ, $\exp(i\beta_1) < 0$. Hence from equation (112) one must have

$$W_1 = W_2 . \qquad (114)$$

Therefore from equation (20) and (114),

$$(M_o - M_1)\,(M_1 + M_2) = (M_o + M_1)\,(M_1 - M_2) ,$$

so that

$$M_1 = \sqrt{M_o\,M_2} . \qquad (114a)$$

*Films belonging to the class discussed in this section are often said to be films that frustrate total internal reflection. For the simple case treated here, total reflection is frustrated more thoroughly as the thickness of the film is decreased.

** The choice (113a) leads to the theory of the "soap film." Thus at normal incidence the optical path $n_1\ d_1$ is an integral number of half-wavelengths.

At normal incidence equations (113), (113b) and (114a) reduce to the pair of well known and independent conditions

$$n_1 d_1 = \mu \frac{\lambda}{4} , \tag{115}$$

where μ = odd, and $n_1 = \sqrt{n_o n_2}$.

At oblique incidence, it is often advantageous to assemble the independent conditions (113) and (114a) in the more explicit form

$$n_1 d_1 \cos i_1 = \mu \frac{\lambda}{4} \ ; \ \mu \text{ odd} ;$$

and (116)

$$\frac{n_1}{\sqrt{n_o n_2}} = \frac{\sqrt{\cos i_o \ \cos i_2}}{\cos i_1} ,$$

in which, from Snell's law, $n_o \sin i_o = n_1 \sin i_1 = n_2 \sin i_2$. The choice of n_1 and d_2 for zero reflectance depends therefore upon the angle, i_o , of incidence. The first condition of equation (115) is often called the quarter wave condition. The effective "interference path," $n_1 d_1 \cos i_1$, is equal to the optical path, $n_1 d_1$, of the film at normal incidence. The interference path, $n_1 d_1 \cos i_1$, always obeys the quarter wave condition.

21.3.6 Zero reflectance; the H-vector perpendicular to the plane of incidence. As in Section 21.3.5, zero reflectance is possible when both M_1 and M_2 are real; but, as we shall see, the zero conaition corresponding to (114a) is significantly different. Since $\gamma_1 = F_2$ for monolayers, the zero condition (98a) assumes the form

$$F_2 \ e^{i\beta_1} = - F_1 . \tag{117}$$

With non-absorbing media, it will become clear from equation (21) that the Fresnel coefficients, F_1 and F_2, of reflection must be real when M_1 and M_2 are real. Conditions (113) and (113a) apply again. Corresponding to the choice (113), equation (117) requires that

$$F_1 = F_2 . \tag{118}$$

From equations (118) and (21) one obtains straightforwardly,

$$\frac{M_1}{\sqrt{M_o \ M_2}} = \frac{n_1^2}{n_o \cdot n_2} , \tag{118a}$$

a condition that should be compared with equation (114a). Upon introducing $M_1 = n_1 \cos i_1$, $M_2 = n_2 \cos i_2$ and $M_o = n_o \sin i_o$, one finds instead of equation (116) that

$$\frac{n_1}{\sqrt{n_o n_2}} = \frac{\cos i_1}{\sqrt{\cos i_o \ \cos i_2}} . \tag{119}$$

21.3.7 Summary.

21.3.7.1 We learn that a film that will produce zero reflectance at oblique incidence when the E-vector is perpendicular to the plane of incidence cannot be expected to produce zero reflectance when the H-vector is perpendicular to the plane of incidence. This conclusion could have been expected; for when the magnetic vector is perpendicular to the plane of incidence, the reflectance is automatically zero at Brewster's angle of incidence without the use of a film whereas,the reflectance is not zero when the electric vector is perpendicular to the plane of incidence. We conclude also that a monolayer cannot in principle produce strictly zero energy reflectance with unpolarized, incident light at other than normal incidence.

21.3.7.2 The reflectance method can be applied in a systematic manner to design bilayers, trilayers, etc. that produce zero reflectance at one or more wavelengths. Except with certain simplified and restricted combinations, the details of the analysis become exceedingly tedious as the number, N, of layers is increased beyond N = 3.

21.4 MATRIX METHODS

21.4.1 Introduction. Matrix methods possess distinct advantages for computation with desk or automatic calculators. For example, the nonlinear recursion formulae that relate successive interfacial reflectances are avoided. We shall construct matrix methods for computing the complex amplitudes, τ_ν and r_ν, for for cases in which the electric vector is perpendicular to the plane of incidence, and for computing the complex amplitudes, T_ν and R_ν, for cases in which the magnetic vector is perpendicular to the plane of incidence. One may treat other states of polarization by splitting the field into two parts, in one of which the E-vector is perpendicular to the plane of incidence, and in the other of which the magnetic vector is perpendicular to the plane of incidence. The complex amplitudes, τ_ν and r_ν, retain the same physical significance as described in Sections 21.2.8 and 21.2.9. The complex amplitudes, T_ν and R_ν, retain the same significance as in Section 21.2.10.

21.4.2 Matrix methods; the E-vector perpendicular to the plane of incidence.

21.4.2.1 Equations (65) and (66) describe the electric vector propagated to the right and left, respectively, in the ν^{th} layer or medium, Figure 21.6. The electric vector has only the y-component. Thus,

$$\begin{aligned}(E_{y,\nu})_{\text{transmitted}} &= \tau_\nu \; e^{-i\omega t} \; e^{ikm_o p_o x} \; e^{ikM_\nu (z - L_\nu)} ,\\ (E_{y,\nu})_{\text{reflected}} &= r_\nu \; e^{-i\omega t} \; e^{ikm_o p_o x} \; e^{-ikM_\nu (z - L_\nu)} .\end{aligned} \tag{120}$$

From equation (27) the corresponding tangential component $H_{x,\nu}$ of the magnetic vector is determined by

$$H_{x,\nu} = \frac{i}{k} \; \frac{\partial E_{y,\nu}}{\partial z} . \tag{121}$$

Hence,

$$\begin{aligned}(H_{x,\nu})_{\text{transmitted}} &= - M_\nu \; (E_{y,\nu})_{\text{transmitted}} ,\\ (H_{x,\nu})_{\text{reflected}} &= M_\nu \; (E_{y,\nu})_{\text{reflected}} .\end{aligned} \tag{122}$$

Let the total tangential components in the νth layer or medium be denoted by $H_{T,\nu}$ and $E_{T,\nu}$. Then, by definition,

$$\begin{aligned}E_{T,\nu} &= (E_{y,\nu})_{\text{transmitted}} + (E_{y,\nu})_{\text{reflected}} ,\\ H_{T,\nu} &= (H_{x,\nu})_{\text{transmitted}} + (H_{x,\nu})_{\text{reflected}} .\end{aligned} \tag{123}$$

The total tangential components $E_{T,\nu}$ and $H_{T,\nu}$ are continuous across every interface of the multilayer. From equations (120) and (122), $E_{T,\nu}$ and $H_{T,\nu}$ are continuous across the interface $z = L_{\nu-1}$ provided that

$$r_{\nu-1} + \tau_{\nu-1} = r_\nu \; e^{i\frac{\beta\nu}{2}} + \tau_\nu \; e^{-i\frac{\beta\nu}{2}} , \tag{124}$$

$$r_{\nu-1} - \tau_{\nu-1} = \frac{M_\nu}{M_{\nu-1}} \left[r_\nu \; e^{i\frac{\beta\nu}{2}} - \tau_\nu \; e^{-i\frac{\beta\nu}{2}} \right] , \tag{124a}$$

since $k M_\nu (L_\nu - L_{\nu-1}) = k M_\nu d_\nu = \beta_\nu/2$. Therefore,

$$2 r_{\nu-1} = \frac{r_\nu}{M_{\nu-1}} \; e^{i\frac{\beta\nu}{2}} (M_{\nu-1} + M_\nu) + \frac{\tau_\nu}{M_{\nu-1}} \; e^{-i\frac{\beta\nu}{2}} (M_{\nu-1} - M_\nu) , \tag{125}$$

$$2 \tau_{\nu-1} = \frac{r_\nu}{M_{\nu-1}} \; e^{i\frac{\beta\nu}{2}} (M_{\nu-1} - M_\nu) + \frac{\tau_\nu}{M_{\nu-1}} \; e^{-i\frac{\beta\nu}{2}} (M_{\nu-1} + M_\nu) . \tag{125a}$$

It may be noted that division of equations (125) leads to the result of equation (57).

21.4.2.2 The following matrix algebra is all that is needed in executing the matrix method. A matrix describes a linear transformation from one pair of variables x_1, y_1 to a second pair x_2, y_2. Thus the linear transformation

$$\begin{aligned}x_2 &= a_{11} x_1 + a_{12} y_1 ,\\ y_2 &= a_{21} x_1 + a_{22} y_1 ,\end{aligned} \tag{126}$$

is written in matrix notation as

$$\begin{bmatrix} x_2 \\ y_2 \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \begin{bmatrix} x_1 \\ y_1 \end{bmatrix} . \qquad (127)$$

Suppose that a further transformation to the variables x_3 , y_3 is given by

$$\begin{bmatrix} x_3 \\ y_3 \end{bmatrix} = \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix} \begin{bmatrix} x_2 \\ y_2 \end{bmatrix} = \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix} \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \begin{bmatrix} x_1 \\ y_1 \end{bmatrix} . \qquad (128)$$

One can verify by eliminating x_2 , y_2 that the matrix product of equation (128) is a matrix such that

$$\begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix} \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} = \begin{bmatrix} b_{11}\ a_{11} + b_{12}\ a_{21} & b_{11}\ a_{12} + b_{12}\ a_{22} \\ b_{21}\ a_{11} + b_{22}\ a_{21} & b_{21}\ a_{12} + b_{22}\ a_{22} \end{bmatrix} . \qquad (129)$$

Equation (129) gives the rule for multiplication. In performing the multiplication $[b_{kl}] \times [a_{mn}]$, to obtain the element in row i and column j of the product matrix take the scalar product of the i^{th} row of matrix $\underline{b}$, and the j^{th} column of matrix $\underline{a}$. The continued product of any number of matrices can be performed by repeating the rule. Multiplication is not commutative, i. e. $[b_{kl}] \times [a_{mn}] \neq [a_{mn}] \times [b_{kl}]$.

21.4.2.3 Returning to equations (125), we observe that in matrix notation

$$\begin{bmatrix} r_{\nu-1} \\ \tau_{\nu-1} \end{bmatrix} = \frac{\mathcal{M}_\nu}{2\ M_{\nu-1}} \begin{bmatrix} r_\nu \\ \tau_\nu \end{bmatrix} \qquad (130)$$

where $\mathcal{M}_\nu$ denotes the <u>square matrix</u>

$$\mathcal{M}_\nu = \begin{bmatrix} (M_{\nu-1} + M_\nu)\ e^{i\frac{\beta\nu}{2}} & (M_{\nu-1} - M_\nu)\ e^{-i\frac{\beta\nu}{2}} \\ (M_{\nu-1} - M_\nu)\ e^{i\frac{\beta\nu}{2}} & (M_{\nu-1} + M_\nu)\ e^{-i\frac{\beta\nu}{2}} \end{bmatrix} . \qquad (130a)$$

Therefore

$$\begin{bmatrix} r_{\nu-1} \\ \tau_{\nu-1} \end{bmatrix} = \frac{\mathcal{M}_\nu}{2\ M_{\nu-1}} \frac{\mathcal{M}_{\nu+1}}{2\ M_\nu} \begin{bmatrix} r_{\nu+1} \\ \tau_{\nu+1} \end{bmatrix} , \qquad (131)$$

whence,

$$\begin{bmatrix} r_{\nu-1} \\ \tau_{\nu-1} \end{bmatrix} = \frac{1}{2^{N-\nu+1}} \prod_{j=\nu-1}^{N-1} \frac{1}{M_j} \prod_{j=\nu}^{N-1} \mathcal{M}_j \begin{bmatrix} r_N \\ \tau_N \end{bmatrix} , \qquad (131a)$$

and

$$\begin{bmatrix} r_o \\ \tau_o \end{bmatrix} = \frac{1}{2^N} \prod_{j=0}^{N-1} \frac{1}{M_j} \prod_{j=1}^{N} \mathcal{M}_j \begin{bmatrix} r_N \\ \tau_N \end{bmatrix} . \qquad (131b)$$

But r_N / τ_N is the Fresnel coefficient of reflection at the last interface, so that

$$r_N = \tau_N\ W_{N+1} = \tau_N \frac{M_N - M_{N+1}}{M_N + M_{N+1}} . \qquad (132)$$

Furthermore, τ_{N+1}/τ_N is the Fresnel coefficient of transmission of the last interface. Hence,

$$\tau_N = \frac{M_N + M_{N+1}}{2\ M_N} \tau_{N+1} ; \qquad (132a)$$

$$r_N = \frac{M_N - M_{N+1}}{2\ M_N} \tau_{N+1} ; \qquad (132b)$$

$$\begin{bmatrix} r_N \\ \tau_N \end{bmatrix} = \frac{1}{2\ M_N} \begin{bmatrix} (M_N - M_{N+1})\ \tau_{N+1} \\ (M_N + M_{N+1})\ \tau_{N+1} \end{bmatrix} = \frac{\tau_{N+1}}{2\ M_N} \begin{bmatrix} (M_N - M_{N+1}) \\ (M_N + M_{N+1}) \end{bmatrix} . \qquad (132c)$$

Finally, from equations (131b) and (132c), one obtains

$$\begin{bmatrix} r_o \\ \tau_o \end{bmatrix} = \frac{\tau_{N+1}}{2^{N+1}} \prod_{j=0}^{N} \frac{1}{M_j} \prod_{j=1}^{N} \mathcal{M}_j \begin{bmatrix} (M_N - M_{N+1}) \\ (M_N + M_{N+1}) \end{bmatrix} . \qquad (133)$$

In this equation the unknowns are usually r_o and τ_{N+1}. τ_o is a complex number that specifies the amplitude and phase of the incident electric vector at $x = 0$ at the left hand side of the first interface $z = 0$, Figure 21.6. One may set $\tau_o = 1$. r_o specifies the amplitude and phase of the reflected vector at $x = 0$ at the left hand side of the first interface of the multilayer. τ_{N+1} specifies the amplitude and phase of the transmitted electric vector at $x = 0$ at the right hand side of the last interface $z = L_N$, Figure 21.6. N is the number of layers.

21.4.2.4 One important advantage of this matrix method is that it enables the exploration of the effects of changing the thickness and refractive index of the ν^{th} layer without recomputing the entire matrix product beyond the $(\nu-1)^{th}$ layer. Changing thickness of the νth layer alters only the matrix $\mathcal{M}_\nu$. Because

$$\prod_{j=1}^{N} \mathcal{M}_j = \prod_{j=1}^{\nu-1} \mathcal{M}_j \times \mathcal{M}_\nu \times \prod_{j=\nu+1}^{N} \mathcal{M}_j , \qquad (134)$$

the first and third matrix products in the right hand member can be computed as matrices that remain fixed during the exploration of the effect of changing thickness. Changing refractive index of the ν^{th} layer alters both $\mathcal{M}_\nu$ and $\mathcal{M}_{\nu+1}$. In this case one utilizes instead of equation (134),

$$\prod_{j=1}^{N} \mathcal{M}_j = \prod_{j=1}^{\nu-1} \mathcal{M}_j \times \mathcal{M}_\nu \mathcal{M}_{\nu+1} \times \prod_{j=\nu+2}^{N} \mathcal{M}_j . \qquad (135)$$

This case illustrates also the manipulation of sub-products of matrices where these sub-products may remain fixed or may be varied. A particular fixed sub-product may be shifted to different positions in the complete matrix product -- corresponding to shifting a group of layers in the multilayer. The use of the matrix method for designing multilayers with periodic structure has been discussed by W. Weinstein.[5]

21.4.2.5 A further interpretation of the matrices $\mathcal{M}_\nu$ is appropriate. It is not a trivial fact that the matrix $\mathcal{M}_\nu$ of equation (130a) can be expressed as the matrix product

$$\mathcal{M}_\nu = \begin{bmatrix} M_{\nu-1} + M_\nu & M_{\nu-1} - M_\nu \\ M_{\nu-1} - M_\nu & M_{\nu-1} + M_\nu \end{bmatrix} \begin{bmatrix} e^{i\frac{\beta\nu}{2}} & 0 \\ 0 & e^{-i\frac{\beta\nu}{2}} \end{bmatrix} , \qquad (136)$$

$$= \mathcal{S}_\nu \mathcal{J}_\nu \qquad (136a)$$

wherein $\mathcal{S}_\nu$ is the first right hand matrix in equation (136) and $\mathcal{J}_\nu$ is the second right hand matrix. $\mathcal{S}_\nu$ and $\mathcal{J}_\nu$ are matrices that depend, respectively, upon the optical constants, M_ν, and the interference path, β_ν, of the νth layer. With reference to equation (133), one may write when desired

$$\prod_{j=1}^{N} \mathcal{M}_j = \prod_{j=1}^{N} \mathcal{S}_j \mathcal{J}_j \qquad (136b)$$

21.4.2.6 Let us now consider solution in terms of the amplitude-sums. With respect to equations (124), let

$$\begin{aligned} A_\nu &= r_\nu + \tau_\nu ; \\ B_\nu &= M_\nu (r_\nu - \tau_\nu) . \end{aligned} \qquad (137)$$

Then

$$\begin{aligned} r_\nu &= \frac{1}{2}\left(A_\nu + \frac{B_\nu}{M_\nu} \right) ; \\ \tau_\nu &= \frac{1}{2}\left(A_\nu - \frac{B_\nu}{M_\nu} \right) . \end{aligned} \qquad (137a)$$

By eliminating r_ν and τ_ν from equations (124) and (124a) with the aid of equation (137a), one finds that

$$\begin{aligned} A_{\nu-1} &= A_\nu \cos(\beta_\nu/2) + B_\nu \frac{i \sin(\beta_\nu/2)}{M_\nu} ; \\ B_{\nu-1} &= A_\nu \; i M_\nu \sin(\beta_\nu/2) + B_\nu \cos(\beta_\nu/2) . \end{aligned} \qquad (137b)$$

[5] Walter Weinstein, Vacuum, 4, 3-18 (1954).

Hence the amplitude-sums, A_ν and B_ν , obey the relation

$$\begin{bmatrix} A_{\nu-1} \\ B_{\nu-1} \end{bmatrix} = \begin{bmatrix} \cos(\beta_\nu/2) & \dfrac{i\sin(\beta_\nu/2)}{M_\nu} \\ i\,M_\nu \sin(\beta_\nu/2) & \cos(\beta_\nu/2) \end{bmatrix} \begin{bmatrix} A_\nu \\ B_\nu \end{bmatrix} . \tag{137c}$$

From equation (132)

$$A_N = r_N + \tau_N = \tau_{N+1} ;$$
$$B_N = M_N \,(r_N - \tau_N) = - M_{N+1}\, \tau_{N+1} . \tag{137d}$$

Therefore

$$\begin{bmatrix} A_o \\ B_o \end{bmatrix} = \tau_{N+1} \prod_{\nu=1}^{N} \begin{bmatrix} \cos(\beta_\nu/2) & \dfrac{i\sin(\beta_\nu/2)}{M_\nu} \\ i\,M_\nu \sin(\beta_\nu/2) & \cos(\beta_\nu/2) \end{bmatrix} \begin{bmatrix} 1 \\ -M_{N+1} \end{bmatrix} . \tag{137e}$$

The method of computation based upon the use of equations (137e) and (137a) is especially advantageous in dealing with non-absorbing multilayers and substrates; for then all M_ν and β_ν are real. Computation of the matrix product of equation (137e) becomes relatively simple. It should be noted that the determinant of each matrix is unity. As pointed out by W. Weinstein,[6] the determinant of the product of matrices is the product of their determinants. Therefore the determinant of products of these matrices is unity -- a valuable fact for the purpose of checking calculations. The complex amplitudes, r_o and τ_{N+1}, are computed at the end, with τ_o assigned a convenient value such as unity. The reflectance, $\rho_o = r_o / \tau_o$, and the transmittance, τ_{N+1} / τ_o, become known.

21.4.3 Matrix methods; the H-vector perpendicular to the plane of incidence.

21.4.3.1 When the magnetic vector is perpendicular to the plane of incidence, the tangential components of E and H are given by equations (84). As in equation (123), we form the total tangential components consisting of the transmitted and reflected waves. Application of the continuity condition for the total tangential components of E and H at the interface $z = L_{\nu-1}$ leads, as in equation (124), to the result

$$R_{\nu-1} + T_{\nu-1} = R_\nu \, e^{i\frac{\beta\nu}{2}} + T_\nu \, e^{-i\frac{\beta\nu}{2}} ; \tag{138}$$

$$R_{\nu-1} - T_{\nu-1} = \frac{M_\nu \, m_{\nu-1}^2}{M_{\nu-1}\, m_\nu^2} \left[R_\nu \, e^{i\frac{\beta\nu}{2}} - T_\nu \, e^{-i\frac{\beta\nu}{2}} \right] . \tag{138a}$$

Therefore

$$2\,R_{\nu-1} = \frac{R_\nu \, e^{i\frac{\beta\nu}{2}}}{M_{\nu-1}\, m_\nu^2} \left(M_{\nu-1}\, m_\nu^2 + M_\nu \, m_{\nu-1}^2\right) + \frac{T_\nu \, e^{-i\frac{\beta\nu}{2}}}{M_{\nu-1}\, m_\nu^2} \left(M_{\nu-1}\, m_\nu^2 - M_\nu \, m_{\nu-1}^2\right) \tag{139}$$

$$2\,T_{\nu-1} = \frac{R_\nu \, e^{i\frac{\beta\nu}{2}}}{M_{\nu-1}\, m_\nu^2} \left(M_{\nu-1}\, m_\nu^2 - M_\nu \, m_{\nu-1}^2\right) + \frac{T_\nu \, e^{-i\frac{\beta\nu}{2}}}{M_{\nu-1}\, m_\nu^2} \left(M_{\nu-1}\, m_\nu^2 + M_\nu \, m_{\nu-1}^2\right) . \tag{139a}$$

Hence,

$$\begin{bmatrix} R_{\nu-1} \\ T_{\nu-1} \end{bmatrix} = \frac{\mathcal{M}_\nu}{2\,M_{\nu-1}\, m_\nu^2} \begin{bmatrix} R_\nu \\ T_\nu \end{bmatrix} , \tag{140}$$

in which $\mathcal{M}_\nu$ is the matrix

$$\mathcal{M}_\nu . = \begin{bmatrix} \left(M_{\nu-1}\, m_\nu^2 + M_\nu \, m_{\nu-1}^2\right) e^{i\frac{\beta\nu}{2}} & \left(M_{\nu-1}\, m_\nu^2 - M_\nu \, m_{\nu-1}^2\right) e^{-i\frac{\beta\nu}{2}} \\ \left(M_{\nu-1}\, m_\nu^2 - M_\nu \, m_{\nu-1}^2\right) e^{i\frac{\beta\nu}{2}} & \left(M_{\nu-1}\, m_\nu^2 + M_\nu \, m_{\nu-1}^2\right) e^{-i\frac{\beta\nu}{2}} \end{bmatrix} . \tag{140a}$$

[6] ibid p 8

Therefore

$$\begin{bmatrix} R_o \\ T_o \end{bmatrix} = \frac{1}{2^N} \prod_{j=0}^{N-1} \frac{1}{M_j\, m_{j+1}^2} \prod_{j=1}^{N} \mathcal{M}_j \begin{bmatrix} R_N \\ T_N \end{bmatrix}, \qquad (140b)$$

however,

$$T_N = \frac{m_{N+1}^2 M_N + m_N^2 M_{N+1}}{2 M_N\, m_{N+1}^2} \quad T_{N+1}, \qquad (140c)$$

and

$$R_N = F_N \quad T_N = \frac{m_{N+1}^2 M_N - m_N^2 M_{N+1}}{2 M_N\, m_{N+1}^2} \quad T_{N+1}, \qquad (140d)$$

with F_N given by equation (21). Hence,

$$\begin{bmatrix} R_o \\ T_o \end{bmatrix} = \frac{T_{N+1}}{2^{N+1}} \prod_{j=0}^{N} \frac{1}{M_j\, m_{j+1}^2} \prod_{j=1}^{N} \mathcal{M}_j \begin{bmatrix} m_{N+1}^2 M_N - m_N^2 M_{N+1} \\ m_{N+1}^2 M_N + m_N^2 M_{N+1} \end{bmatrix}, \qquad (141)$$

in which the matrices $\mathcal{M}_j$ are given by equation (140a). The complex amplitudes R_ν and T_ν retain the same physical significance as in 21.2.10.

21.4.3.2 Equation (141) serves to determine T_{N+1} from T_o and R_o from T_o and T_{N+1}. In most problems, one may set $T_o = 1$. Finally, the complex reflectance γ_o is computed from its definition $\gamma_o = R_o/T_o$. It will be observed that equations (133) and (141) determine the complex transmittances, τ_{N+1} and T_{N+1}, of the multilayer quite directly without necessitating additional multiplications such as those of equation (54) or (70).

21.4.3.3 The solution in terms of the amplitude sums can be obtained as follows. Let

$$\begin{aligned} C_\nu &= R_\nu + T_\nu ; \\ D_\nu &= \frac{M_\nu}{m_\nu^2} (R_\nu - T_\nu) . \end{aligned} \qquad (142)$$

Then

$$\begin{aligned} R_\nu &= \frac{1}{2} \left(C_\nu + \frac{m_\nu^2}{M_\nu} D_\nu\right) ; \\ T_\nu &= \frac{1}{2} \left(C_\nu - \frac{m_\nu^2}{M_\nu} D_\nu\right) . \end{aligned} \qquad (142a)$$

By eliminating R_ν and T_ν from equation (138) with the aid of equation (142a), one obtains

$$\begin{aligned} C_{\nu-1} &= C_\nu \cos \frac{\beta_\nu}{2} + i\, D_\nu \frac{m_\nu^2}{M_\nu} \sin \frac{\beta_\nu}{2} ; \\ D_{\nu-1} &= i\, C_\nu \frac{M_\nu}{m_\nu^2} \sin \frac{\beta_\nu}{2} + D_\nu \cos \frac{\beta_\nu}{2} . \end{aligned} \qquad (142b)$$

Therefore

$$\begin{bmatrix} C_{\nu-1} \\ D_{\nu-1} \end{bmatrix} = \begin{bmatrix} \cos \frac{\beta_\nu}{2} & i \frac{m_\nu^2}{M_\nu} \sin \frac{\beta_\nu}{2} \\ i \frac{M_\nu}{m_\nu^2} \sin \frac{\beta_\nu}{2} & \cos \frac{\beta_\nu}{2} \end{bmatrix} \begin{bmatrix} C_\nu \\ D_\nu \end{bmatrix} . \qquad (142c)$$

From equations (140c) and (140d),

$$C_N = R_N + T_N = T_{N+1};$$
$$D_N = \frac{M_N}{m_N^2}(R_N - T_N) = -\frac{M_{N+1}}{m_{N+1}^2} T_{N+1}. \qquad (142d)$$

Therefore, upon forming $\begin{bmatrix} C_o \\ D_o \end{bmatrix}$ from equation (142c), one obtains

$$\begin{bmatrix} C_o \\ D_o \end{bmatrix} = T_{N+1} \prod_{\nu=1}^{N} \begin{bmatrix} \cos\frac{\beta_\nu}{2} & i\,\frac{m_\nu^2}{M_\nu}\sin\frac{\beta_\nu}{2} \\ i\,\frac{M_\nu}{m_\nu^2}\sin\frac{\beta_\nu}{2} & \cos\frac{\beta_\nu}{2} \end{bmatrix} \begin{bmatrix} 1 \\ -M_{N+1}/m_{N+1}^2 \end{bmatrix}, \qquad (142e)$$

a result that should be compared with equation (137e). The remarks in the paragraph following equation (137e) apply again to equation (142e).

21. 5 QUATERNION METHODS

21. 5. 1 Introduction. A computing program based upon quarternions instead of matrices has been introduced by Dr. Gordon L. Walker * for analyzing thin films with the aid of automatic computers. The relative advantages of matrices and quaternions depend mainly upon circumstances at the location of the automatic calculator. For example, wherever a programmed matrix formulation is available, it may be simpler to adapt a matrix method. As has been seen in Section 21. 4, it is both natural and direct to state the solution to the problems of thin films in matrix notation. Because many solutions in matrix form have been given, we shall restrict our considerations to showing how any matrix solution can be transformed into the corresponding quaternion form.

10. 5. 2 Quaternions. A quaternion Q is the sum of a scalar and a vector. Thus,

$$Q = q_o\sigma_o + q_1\sigma_1 + q_2\sigma_2 + q_3\sigma_3, \qquad (143)$$

in which σ_1, σ_2 and σ_3 are unit vectors and $\sigma_o = 1$. Coefficients q_o, q_1, q_2 and q_3 are scalars that can be complex imaginary. With respect to summation,

$$P + Q = Q + P = (p_o + q_o)\sigma_o + (p_1 + q_1)\sigma_1 + (p_2 + q_2)\sigma_2 + (p_{-} + q_3)\sigma_3. \qquad (143a)$$

If b is a scalar,

$$bQ = Qb = bq_o\sigma_o + bq_1\sigma_1 + bq_2\sigma_2 + bq_3\sigma_3. \qquad (143b)$$

In forming the product of two quaternions, P and Q, one observes the following rules of operation with respect to the unit vectors.

$$\sigma_1^2 = \sigma_2^2 = \sigma_3^2 = -1; \qquad \sigma_o^2 = 1;$$
$$\sigma_1\sigma_2 = -\sigma_2\sigma_1 = \sigma_3;$$
$$\sigma_2\sigma_3 = -\sigma_3\sigma_2 = \sigma_1;$$
$$\sigma_3\sigma_1 = -\sigma_1\sigma_3 = \sigma_2. \qquad (143c)$$

It follows that $PQ \neq QP$. Let p_ν and q_ν be the coefficients of quaternions P and Q, respectively. Set

$$PQ = R = R_o\sigma_o + R_1\sigma_1 + R_2\sigma_2 + R_3\sigma_3. \qquad (143d)$$

*This unpublished scheme has been used by Drs. G. L. Walker, H. Jupnik and A. Traub at the American Optical Company. A method that resembles the method of quaternions in several respects has been discussed by M. Andre Herpin, Comptes Rendus Acad. Sci., 225, 182-183 (1947).

Then,

$$\begin{aligned} R_o &= p_o q_o - p_1 q_1 - p_2 q_2 - p_3 q_3 \,; \\ R_1 &= p_o q_1 + q_o p_1 + p_2 q_3 - p_3 q_2 \,; \\ R_2 &= p_o q_2 + q_o p_2 + p_3 q_1 - p_1 q_3 \,; \\ R_3 &= p_o q_3 + q_o p_3 + p_1 q_2 - p_2 q_1 \,. \end{aligned} \qquad (143e)$$

21.5.3 Corresponding matrices and quaternions. Let

$$\mathcal{M} = \begin{bmatrix} m_{11} & m_{12} \\ m_{21} & m_{22} \end{bmatrix} \qquad (144)$$

be a given matrix. Let $\tilde{\mathcal{M}}$ denote the corresponding quaternion. Then

$$\tilde{\mathcal{M}} = m_o \sigma_o + m_1 \sigma_1 + m_2 \sigma_2 + m_3 \sigma_3 \,, \qquad (144a)$$

wherein

$$m_o = \frac{m_{11} + m_{22}}{2} \,; \qquad m_2 = \frac{m_{12} - m_{21}}{2} \,;$$

$$m_1 = i \, \frac{m_{11} - m_{22}}{2} \,; \qquad m_3 = i \, \frac{m_{12} + m_{21}}{2} \,. \qquad (144b)$$

On the other hand, let $\tilde{\mathcal{M}}$ be the given quaternion. In order to obtain the corresponding matrix, one replaces $\sigma_o = 1$ and the unit vectors by the matrices

$$\sigma_o = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} ; \qquad \sigma_2 = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix} ;$$

$$\sigma_1 = \begin{bmatrix} -i & 0 \\ 0 & i \end{bmatrix} ; \qquad \sigma_3 = \begin{bmatrix} 0 & -i \\ -i & 0 \end{bmatrix} ; \qquad (144c)$$

in which σ_o is a unit matrix. Thus with $\tilde{\mathcal{M}}$ regarded as the given quaternion, the corresponding matrix is

$$\begin{aligned} \mathcal{M} &= m_o \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} + m_1 \begin{bmatrix} -i & 0 \\ 0 & i \end{bmatrix} + m_2 \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix} + m_3 \begin{bmatrix} 0 & -i \\ -i & 0 \end{bmatrix} ; \\ &= \begin{bmatrix} m_o - i m_1 & m_2 - i m_3 \\ -(m_2 + i m_3) & m_o + i m_1 \end{bmatrix} . \end{aligned} \qquad (144d)$$

The matrices of equations (144c) satisfy the requirements of equations (143c). The quaternion corresponding to the product of two matrices taken in a specified order is the product of the quaternions corresponding to each of the two matrices taken in the same order. Thus

$$\text{Quaternion } (\mathcal{M}_1 \, \mathcal{M}_2) = \tilde{\mathcal{M}}_1 \tilde{\mathcal{M}}_2 \,. \qquad (144e)$$

Similarly,

$$\text{Matrix } (\tilde{\mathcal{M}}_1 \, \tilde{\mathcal{M}}_2) = \mathcal{M}_1 \, \mathcal{M}_2 \,. \qquad (144f)$$

For the purposes of this text, equation (144e) is far more important than equation (144f). Repeated application of equation (144e) shows that

$$\text{Quaternion } \left(\prod_{\nu=1}^{N} \mathcal{M}_\nu \right) = \prod_{\nu=1}^{N} \tilde{\mathcal{M}}_\nu \,, \qquad (144g)$$

in which $\tilde{\mathcal{M}}_\nu$ is the quaternion that corresponds to matrix $\mathcal{M}_\nu$.

21.5.4 Replacements for matrix equations. Let $\mathcal{M}$ be the square matrix

$$\mathcal{M} = \begin{bmatrix} m_{11} & m_{12} \\ m_{21} & m_{22} \end{bmatrix}, \tag{145}$$

that connects the quantities r, τ, A and B according to the law

$$\begin{bmatrix} r \\ \tau \end{bmatrix} = \begin{bmatrix} m_{11} & m_{12} \\ m_{21} & m_{22} \end{bmatrix} \begin{bmatrix} A \\ B \end{bmatrix}, \tag{145a}$$

as in equations (126) and (127). One can verify almost directly from equations (126) to (129) that it is permissible to set

$$\begin{bmatrix} r \\ \tau \end{bmatrix} = \begin{bmatrix} r & 0 \\ \tau & 0 \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} A & 0 \\ B & 0 \end{bmatrix} = \begin{bmatrix} A \\ B \end{bmatrix} .$$

Hence equation (145a) can be written in the form

$$\begin{bmatrix} r & 0 \\ \tau & 0 \end{bmatrix} = \begin{bmatrix} m_{11} & m_{12} \\ m_{21} & m_{22} \end{bmatrix} \begin{bmatrix} A & 0 \\ B & 0 \end{bmatrix} . \tag{145b}$$

The quaternion form of matrix equation (145b) is now obtained by replacing each of the three square matrices by its corresponding quaternion with the aid of equation (144b). Let U denote the quaternion corresponding to the left hand member of equation (145b), i.e. let

$$U = \text{Quaternion} \begin{bmatrix} r & 0 \\ \tau & 0 \end{bmatrix} . \tag{145c}$$

By applying the correspondence rules of equation (144b) to equation (145c), one finds that

$$U = \frac{r}{2} \sigma_o + i \frac{r}{2} \sigma_1 - \frac{\tau}{2} \sigma_2 + i \frac{\tau}{2} \sigma_3 . \tag{145d}$$

Similarly, with respect to

$$S_o = \text{Quaternion} \begin{bmatrix} A & 0 \\ B & 0 \end{bmatrix} ,$$

$$S_o = \frac{A}{2} \sigma_o + i \frac{A}{2} \sigma_1 - \frac{B}{2} \sigma_2 + i \frac{B}{2} \sigma_3 . \tag{145e}$$

Matrix equation (145b) is therefore replaced by the quaternion equation

$$U = \tilde{\mathcal{M}} S_o , \tag{145f}$$

in which

$$\begin{aligned} \tilde{\mathcal{M}} &= \text{Quaternion} \begin{bmatrix} m_{11} & m_{12} \\ m_{21} & m_{22} \end{bmatrix} \\ &= m_o \sigma_o + m_1 \sigma_1 + m_2 \sigma_2 + m_3 \sigma_3 , \end{aligned} \tag{145g}$$

as in equations (144a) and (144b). It is instructive to examine the quaternion product $\tilde{\mathcal{M}}$ S_o. One finds from equations (143d), (143e), (145e) and (145g) that with

$$U = \tilde{\mathcal{M}} S_o = \sigma_o U_o + U_1 \sigma_1 + U_2 \sigma_2 + U_3 \sigma_3 , \tag{145h}$$

$$\begin{aligned} U_o &= \frac{A}{2} (m_o - i m_1) + \frac{B}{2} (m_2 - i m_3) ; \\ U_2 &= \frac{A}{2} (m_2 + i m_3) - \frac{B}{2} (m_o + i m_1) ; \end{aligned} \tag{145i}$$

together with

$$U_1 = i U_o ; \quad U_3 = - i U_2 . \tag{145j}$$

The relations between U_o, U_2 and m_{kl} are obtained from equations (144b) and (145i). One finds that

$$U_o = \frac{A}{2} m_{11} + \frac{B}{2} m_{12} ;$$
$$U_2 = \frac{A}{2} m_{21} - \frac{B}{2} m_{22} . \qquad (145k)$$

21.5.5 Quaternion solutions for r and τ. The solutions for r and τ are found by comparing equations (145d) and (145h). These equations require that

$$r = 2 U_o ;$$
$$\tau = -2 U_2 ; \qquad (146)$$

in which U_o and U_2 are the coefficients of σ_o and σ_2 in the quaternion product $\tilde{\mathfrak{M}} S_o$ of equation (145f). Equation (146) forms a simple way of computing r and τ once U_o and U_2 have been found.

21.5.6 A check on the quaternion method. Equations (146) and (145k) can be combined to form a simple check on the correctness of the quaternion method. One finds directly that

$$r = m_{11} A + m_{12} B ;$$
$$\tau = m_{21} A + m_{22} B . \qquad (147)$$

These solutions for r and τ are those of the matrix equation (145a). If therefore equation (145a) is correct, equation (146) is correct.

21.5.7 A more useful statement of the formulation. In stating the matrix that corresponds to a multilayer, it is rarely convenient to specify the matrix in the form of equation (145a). Instead, it is convenient to express the matrix in the form

$$\begin{bmatrix} r \\ \tau \end{bmatrix} = F \begin{bmatrix} m_{11} & m_{12} \\ m_{21} & m_{22} \end{bmatrix} \begin{bmatrix} A \\ B \end{bmatrix} \qquad (148)$$

in which factor F is a scalar. For example, in the matrix of equation (137e) the factor $F = \tau_{N+1}$. Re-examination of the argument leading from equation (145a) to (146) shows that one obtains instead of equation (146) the slightly modified result

$$r = 2 F U_o ;$$
$$\tau = -2 F U_2 ; \qquad (149)$$

in which U_o and U_2 are the coefficients of σ_o and σ_2, respectively, of the quaternion corresponding to the matrix product of equation (148).

21.5.8 Special conditions. Depending upon the nature of the multilayer and upon the manner in which the corresponding matrix problem has been solved, special conditions may exist among the matrix elements m of the matrix $\mathfrak{M}$ associated with the multilayer. For example, with respect to the matrix

$$\mathfrak{M} = \prod_{\nu=1}^{N} \mathfrak{M}_\nu = \begin{bmatrix} m_{11} & m_{12} \\ m_{21} & m_{22} \end{bmatrix} , \qquad (150)$$

of equation (133), it follows that at normal incidence upon a system of non-absorbing layers

$$m_{22} = \bar{m}_{11} ; \qquad m_{21} = \bar{m}_{12} ; \qquad (151)$$

because the diagonal elements of $\mathfrak{M}_\nu$ of equation (130a) are then conjugate complex numbers. Correspondingly, from equation (144b)

$$m_o = R_e (m_{11}) ; \qquad m_2 = i \mathcal{I}_m (m_{12}) ;$$
$$m_1 = -\mathcal{I}_m (m_{11}) ; \qquad m_3 = i R_e (m_{12}) . \qquad (152)$$

It will not, however, be a purpose of this discussion to enumerate the various special conditions together with their consequences.

21.5.9 Application to the matrix solution of equation (133). The matrix solution of equation (133) applies to cases in which the electric vector is polarized so as to vibrate at right angles to the plane of incidence. This example will illustrate procedures that may be followed in converting any matrix solution for thin films into a solution in terms of quaternions. With respect to efficiency, it is undoubtedly preferable to choose as factor F the product

$$\frac{\tau_{N+1}}{2^{N+1}} \prod_{\nu=0}^{N} \frac{1}{M_\nu}$$

of equation (133). Let us demand, however, that the Fresnel coefficients W_ν of equation (20) shall appear as parameters. This can be accomplished by writing $\mathcal{M}_\nu$ of equation (130a) in the form

$$\mathcal{M}_\nu = (M_{\nu-1} + M_\nu) \begin{bmatrix} e^{i\frac{\beta\nu}{2}} & W_\nu\, e^{-i\frac{\beta\nu}{2}} \\ W_\nu\, e^{i\frac{\beta\nu}{2}} & e^{-i\frac{\beta\nu}{2}} \end{bmatrix} . \qquad (153)$$

Let also the factor $M_N + M_{N+1}$ be removed from the second right hand matrix of equation (133). Then,

$$\begin{bmatrix} r_o \\ \tau_o \end{bmatrix} = F \prod_{\nu=1}^{N} \begin{bmatrix} e^{i\frac{\beta\nu}{2}} & W_\nu\, e^{-i\frac{\beta\nu}{2}} \\ W_\nu\, e^{i\frac{\beta\nu}{2}} & e^{-i\frac{\beta\nu}{2}} \end{bmatrix} \begin{bmatrix} W_{N+1} \\ 1 \end{bmatrix} ; \qquad (154)$$

in which

$$F = \frac{\tau_{N+1}}{2^{N+1}} \; \frac{M_N + M_{N+1}}{M_o} \prod_{\nu=1}^{N} \frac{M_\nu + M_{\nu-1}}{M_\nu} . \qquad (155)$$

We take $\tilde{\mathcal{M}}_\nu$ as the quaternion corresponding to the ν^{th} matrix of the product from $\nu = 1$ to $\nu = N$ of equation (154). Explicitly, we take

$$\tilde{\mathcal{M}}_\nu = \cos\frac{\beta\nu}{2}\,\sigma_o - \sin\frac{\beta\nu}{2}\,\sigma_1 - i\,W_\nu \sin\frac{\beta\nu}{2}\,\sigma_2 + i\,W_\nu \cos\frac{\beta\nu}{2}\,\sigma_3 . \qquad (156)$$

The quaternion S_o of the last matrix of equation (154) is obtained from equation (145e) by setting

$$A = W_{N+1} ; \qquad B = 1 . \qquad (157)$$

We form and compute the quaternion

$$\tilde{\mathcal{M}} = \prod_{\nu=1}^{N} \tilde{\mathcal{M}}_\nu \qquad (158)$$

with $\tilde{\mathcal{M}}_\nu$ given in terms of the physical properties β_ν and W_ν of the ν^{th} layer by equation (156). Next, as in equation (145f), we compute the quaternion U as the product

$$\begin{aligned} U &= \frac{1}{2}\,\tilde{\mathcal{M}} \left[W_{N+1}\,\sigma_o + i\,W_{N+1}\,\sigma_1 - \sigma_2 + i\,\sigma_3 \right] ; \qquad (159) \\ &= \prod_{\nu=1}^{N} \tilde{\mathcal{M}}_\nu \; S_o \end{aligned}$$

In making the last computation of equation (159), it is necessary to compute only U_o and U_2, the coefficients of σ_o and σ_2, respectively. With U_o and U_2 thus determined, it follows from equation (149) that

$$\begin{aligned} r_o &= 2\,F\,U_o ; \\ \tau_o &= -2\,F\,U_2 ; \end{aligned} \qquad (160)$$

in which F is given by equation (155).

21.5.10 Determination of the complex reflectance and transmittance of the multilayer. The complex reflectance of the multilayer is given by the ratio $\rho_o = r_o / \tau_o$. From equation (160)

$$\rho_o = -\,U_o / U_2 , \qquad (161)$$

a result that is independent of the factor F and that is evaluated at the right hand boundary of the medium of incidence with the electric vector perpendicular to the plane of incidence. The complex transmittance of the multilayer is given by the ratio τ_{N+1}/τ_o. From equations (160) and (155)

$$\tau_{N+1}/\tau_o = -2^{N+1}\frac{M_o}{M_N+M_{N+1}}\prod_{\nu=1}^{N}\frac{M_\nu}{M_{\nu-1}+M_\nu}\;\frac{1}{2U_2}, \qquad (162)$$

a result evaluated at the point of entry into the last medium with the electric vector perpendicular to the plane of incidence. Computation of the complex reflectance and transmittance is relatively simple when the coefficients U_o and U_2 of the composite quaternion U of the multilayer have been calculated.

21.5.11 Application to the monolayer. The uninitiated reader will find it a useful exercise to verify from equations (158) and (159) that for the monolayer (the case N = 1)

$$U_o = \frac{W_2}{2}e^{i\frac{\beta_1}{2}} + \frac{W_1}{2}e^{-i\frac{\beta_1}{2}};$$
$$-U_2 = \frac{1}{2}e^{-i\frac{\beta_1}{2}} + \frac{W_1W_2}{2}e^{i\frac{\beta_1}{2}}. \qquad (163)$$

It is then shown easily from equations (161), (162) and (163) that

$$\rho_o = \frac{W_2\,e^{i\beta_1} + W_1}{1 + W_1\,W_2\,e^{i\beta_1}}; \qquad (164)$$

$$\frac{\tau_{N+1}}{\tau_o} = \frac{\tau_2}{\tau_o} = \frac{4M_oM_1}{(M_o+M_1)\,(M_1+M_2)}\;\frac{e^{i\frac{\beta_1}{2}}}{1+W_1W_2\,e^{i\beta_1}}. \qquad (165)$$

Equation (164) agrees, for example, with the result of equations (50) and (51) for cases N = 1. Likewise, equation (165) agrees with the result of equation (54). Matrix methods and quaternion methods do not possess advantages over recursion methods such as those of Sections 21.2.8 and 21.2.9 until the number N of layers in the multilayer exceeds 2. In fact, the methods of equations (164) and (165) are to be preferred as regards simplicity and convenience for the monolayer.

21.5.12 Comments. In the interesting method constructed by Gordon L. Walker, the quaternion corresponding to S_o of equations (145) is rendered incomplete in that the coefficients of the unit vectors σ_1 and σ_2 are zero. The method presented here does not involve more quaternion multiplication than the method due to Walker and requires slightly less algebra at the last steps for computing the complex reflectance ρ_o and the complex transmittance τ_{N+1}/τ_o of the multilayer. Furthermore, Walker has preferred to apply the quaternion method to factored matrices $\mathcal{M}_\nu$ of the type described by equations (136). For example with respect to the matrices of equations (133) and (136), one can take

$$F = \frac{\tau_{N+1}}{2^{N+1}}\prod_{\nu=0}^{N}\frac{1}{M_\nu}; \qquad (166)$$

$$A = M_N - M_{N+1}; \quad B = M_N + M_{N+1}; \qquad (166a)$$

and compute the quaternion

$$U = \prod_{\nu=1}^{N}\tilde{\mathcal{S}}_\nu\,\tilde{\mathcal{J}}_\nu\;S_o, \qquad (166b)$$

wherein $\tilde{\mathcal{S}}_\nu$ and $\tilde{\mathcal{J}}_\nu$ are the quaternions corresponding to matrices $\mathcal{S}_\nu$ and $\mathcal{J}_\nu$, respectively, of equation (136a) and S_o is determined from equation (145a). The complex reflectance ρ_o can be computed from equation (161) and the complex transmittance τ_{N+1}/τ_o can be computed from equation (160). It will be found from equations (136) and (144b) that the quaternions $\tilde{\mathcal{S}}_\nu$ and $\tilde{\mathcal{J}}_\nu$ are incomplete. The product $\tilde{\mathcal{S}}_\nu\,\tilde{\mathcal{J}}_\nu$ is, however, a complete quaternion.

21. 6 MONOLAYER COATINGS

21. 6. 1 Introduction. Monolayer coatings serve many specialized purposes. One broad class of monolayers consists of dielectric substances deposited or formed upon absorbing or non-absorbing substrates. A second broad class consists of metallic or semiconducting films on absorbing or non-absorbing substrates. We shall not be concerned with that class of monolayers whose sole function is to alter the mechanical, chemical or electrical properties of a surface. Monolayers may occur naturally as, for example, in the tarnishing of silver by a layer of silver sulphide. The practical range in thickness of a layer may vary from molecular dimensions to centimeters or even meters depending mainly upon the wavelength of the radiation involved. This radiation may extend from the ultraviolet into radar. Not all monolayers should be regarded as homogeneous but a large group of monolayers can be considered homogeneous in constructing an approximate theory for interpreting their "optical" behavior. The approximations thus afforded are often in excellent agreement with experiment.

21. 6. 2 Methods of computation.

21. 6. 2. 1 Wherever automatic calculators have been programmed on the basis of matrix or quaternion methods, this program can serve for computing monolayers although such programs become unduly elaborate. The following method is one of the useful and accurate methods for desk calculators. This method will be presented for the case in which the incident electric vector is perpendicular to the plane of incidence. When the electric vector vibrates in the plane of incidence, it is necessary to replace the Fresnel coefficients, W_ν, of equation (20) by the Fresnel coefficients, F_ν, of equation (21) as shown in Section 21. 2. 10. By setting $\nu = 1$ in equation (57) and noting that $\rho_1 = W_2$ from equation (51), one finds that the complex reflectance, ρ_o, of the monolayer is given by

$$\rho_o = \frac{W_2 \; e^{i\beta_1} + W_1}{1 + W_1 \; W_2 \; e^{i\beta_1}} , \tag{167}$$

as in equation (164). Likewise, by setting $N = 1$ in equation (54) one obtains the complex transmittance in the form

$$\frac{\tau_2}{\tau_o} = \frac{4 \; M_o \; M_1}{(M_o + M_1)(M_1 + M_2)} \quad \frac{e^{i\frac{\beta_1}{2}}}{1 + W_1 \; W_2 \; e^{i\beta_1}} , \tag{168}$$

as in equation (165). Explicitly,

$$W_\nu = \frac{M_{\nu-1} - M_\nu}{M_{\nu-1} + M_\nu} \; ; \; \nu = 1, \; 2; \tag{169}$$

$$\beta_1 = \frac{4\pi}{\lambda} \; M_1 \; d_1 , \tag{169a}$$

in which M_ν is defined by equation (19). ρ_o and τ_2 are evaluated by equations (167) and (168) at the left hand side of the first interface and at the right hand side of the second interface, respectively, as indicated in Figure 21. 10. Marked simplification occurs at normal incidence for which $i_o = 0$ so that $p_o = \sin i_o = 0$.

$$M_\nu = m_\nu = n_\nu \; (1 + i \, K_\nu) \; ; \; i_o = 0 . \tag{169b}$$

Furthermore, at normal incidence no distinction need be made among the directions of polarization of the electric vector when the film and substrate are isotropic. Equations (167) and (168) then serve for all states of polarization.

21. 6. 2. 2 Phase changes that occur on reflection and transmission are given as phase retardations by arg (ρ_o) and arg (τ_2/τ_o), respectively. When these quantities are wanted, ρ_o and τ_2/τ_o must be evaluated as complex numbers by the operations indicated by equations (167) and (168).

21. 6. 2. 3 The parameter β_1 is awkward to handle whenever m_1 is complex $(K_1 \neq 0)$. This awkwardness is increased when the angle of incidence $i_o \neq 0$. At normal incidence

$$\beta_1 = \frac{4\pi}{\lambda} \; n_1 \; d_1 \; (1 + i \, K_1) , \tag{170}$$

and

$$e^{i\beta_1} = e^{-\frac{4\pi}{\lambda} n_1 K_1 d_1} \; e^{i \frac{4\pi}{\lambda} n_1 d_1} , \tag{170a}$$

in which $4\pi \; n_1 \; d_1 / \lambda$ is twice the optical path of the film and the exponent, $-4\pi \; n_1 \; K_1 \; d_1 / \lambda$, is an attenuation

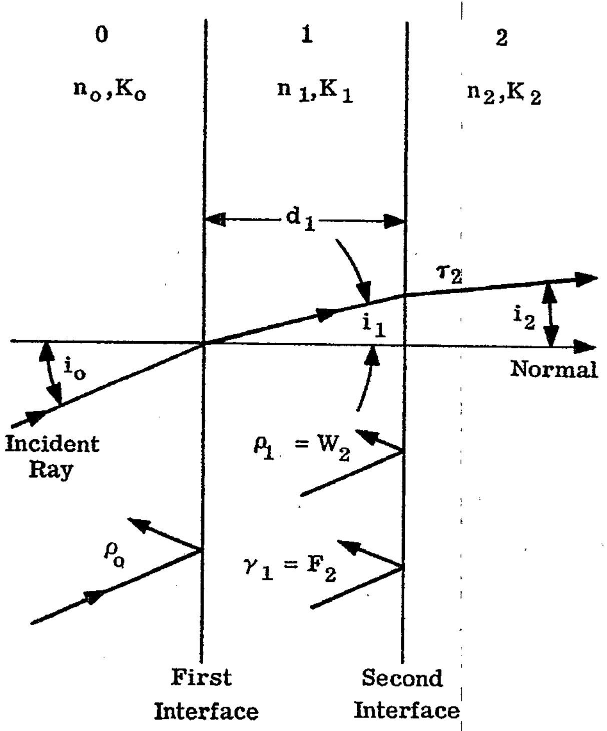


Figure 21. 10- Convention and notation with respect to the monolayer (case N = 1).

factor. When the medium of incidence and the monolayer do not absorb, and when $n_1^2 > n_o^2 p_o^2$,

$$M_1 = \left[n_1^2 - n_o^2 p_o^2\right]^{1/2} = n_1 \cos i_1 ,$$

where i_1 is the angle of refraction in the film. Correspondingly,

$$\beta_1 = \frac{4\pi}{\lambda} n_1 d_1 \cos i_1 ; \quad K_o = K_1 = 0. \tag{170b}$$

21.6.2.4 The energy reflectance R of the monolayer is given by $R = |\rho_o|^2$ and is most easily computed from equation (167) in the form

$$R = |\rho_o|^2 = \frac{|W_2 e^{i\beta_1} + W_1|^2}{|1 + W_1 W_2 e^{i\beta_1}|^2} . \tag{171}$$

Suppose that the system is non-absorbing. In such cases (a class of wide interest) the Fresnel coefficients W_1 and W_2 are real and β_1 is given by equation (170b). Correspondingly,

$$R = \frac{W_1^2 + 2 W_1 W_2 \cos \beta_1 + W_2^2}{1 + 2 W_1 W_2 \cos \beta_1 + W_1^2 W_2^2} . \tag{172}$$

Differentiation of R with respect to β_1 in equation (172) shows that the extreme values of $R(\beta_1)$ occur when

$$\beta_1 = \nu \pi ; \quad \nu \text{ an integer.} \tag{173}$$

Hence when the electric vector is perpendicular to the plane of incidence, when $n_1 > n_o p_o$ and when $K_o = K_1 = K_2 = 0$, the maxima and minima, R_m, are given by

$$R_m = \frac{(W_1 \pm W_2)^2}{(1 \pm W_1 W_2)^2} . \tag{174}$$

The values of W_1 and W_2 (and hence R_m) depend upon the angle of incidence. But at fixed angles of incidence, $R_{max.}$ and $R_{min.}$ remain fixed and R is a periodic function of the thickness d_1. Departures from periodicity can occur when β_1 is altered by changing wavelength because the refractive indices (and hence W_1 and W_2) are dispersive with λ. It is not difficult to see that $R(\beta_1)$ cannot be periodic when the film absorbs. Equation (170a) shows that the $\exp(i\beta_1)$ approaches zero as d_1 approaches infinity. Consequently, from equation (171) R oscillates with increasing β_1 such that

$$R \rightarrow |W_1|^2 , \qquad (175)$$

the Fresnel coefficient of reflectance of the first interface, Figure 21.10.

21.6.2.5 When the system contains no absorbing media, the energy transmittance, T_e, of the monolayer can be found in terms of the energy reflectance, R, from the law of conservation of energy, namely,

$$T_e + R = 1 , \qquad (176)$$

irrespective of the refractive indices of the first and last media. When absorption occurs, one can compute the complex transmittance τ_2 / τ_o from equation (168) and the energy transmittance from equation (58). If only the film is absorbing, $(n_a)_2 = n_2$ so that equation (58) yields the result,

$$T_e = \frac{n_2 \cos i_2}{n_o \cos i_o} \left| \frac{\tau_2}{\tau_o} \right|^2 , \qquad (177)$$

for cases in which the electric vector is perpendicular* to the plane of incidence.

21.6.3 Non-absorbing systems; normal incidence.

21.6.3.1 The behavior of the energy reflectances, $|\rho_o|^2$, for non-absorbing monolayers on non-absorbing substrates is illustrated in Figure 21.11 for cases in which the refractive index $n_o = 1$. When no absorption occurs, reversal of the direction of incidence leaves $|\rho_o|^2$ unchanged. For reasons explained in Section 21.2.14, the monolayer is an absentee layer at points $\beta_1 = \nu 2\pi$ where $\nu = 0$ or an integer. At these points the energy reflectance is that of the uncoated glass.

21.6.3.2 With respect to cases $n_o < n_1 < n_2$, it follows from equation (169) that $W_1 < 0$ and $W_2 < 0$. Correspondingly from equation (174),

$$R_m = R_{min.} ,$$

when

$$\beta_1 = \mu \pi ; \quad \mu \text{ an odd integer.} \qquad (178)$$

By introducing

$$W_1 = \frac{n_o - n_1}{n_o + n_1} ; \quad W_2 = \frac{n_1 - n_2}{n_1 + n_2}$$

into equation (174), one finds that

$$R_{min.} = |\rho_o|^2_{min.} = \left(\frac{n_o n_2 - n_1^2}{n_o n_2 + n_1^2} \right)^2 . \qquad (179)$$

In particular, $R_{min.} = 0$ when β_1 obeys equation (178) and n_1 is chosen so that

$$n_1 = \sqrt{n_o n_2} . \qquad (180)$$

Equations (178) and (180) agree with the more general equation (115) and (116) at normal incidence. Equation (179) is important for two reasons. First, it enables one to estimate the minimum value of the energy reflectance. Secondly, it enables one to compute the refractive index n_1 of the film from the measured**value of $R_{min.}$ and the known values of n_o and n_2. The refractive indices of evaporated monolayers are usually less than those of the bulk materials and vary with the conditions of evaporation.

* See equations (70) and (71) for cases in which the electric vector vibrates in the plane of incidence.

**The spectral energy reflectance, R_p, of a coated plate is ordinarily obtained with a spectrophotometer. Suppose, for example, that the absorption of the plate is negligible and that the energy reflectance of the back surface is B. On account

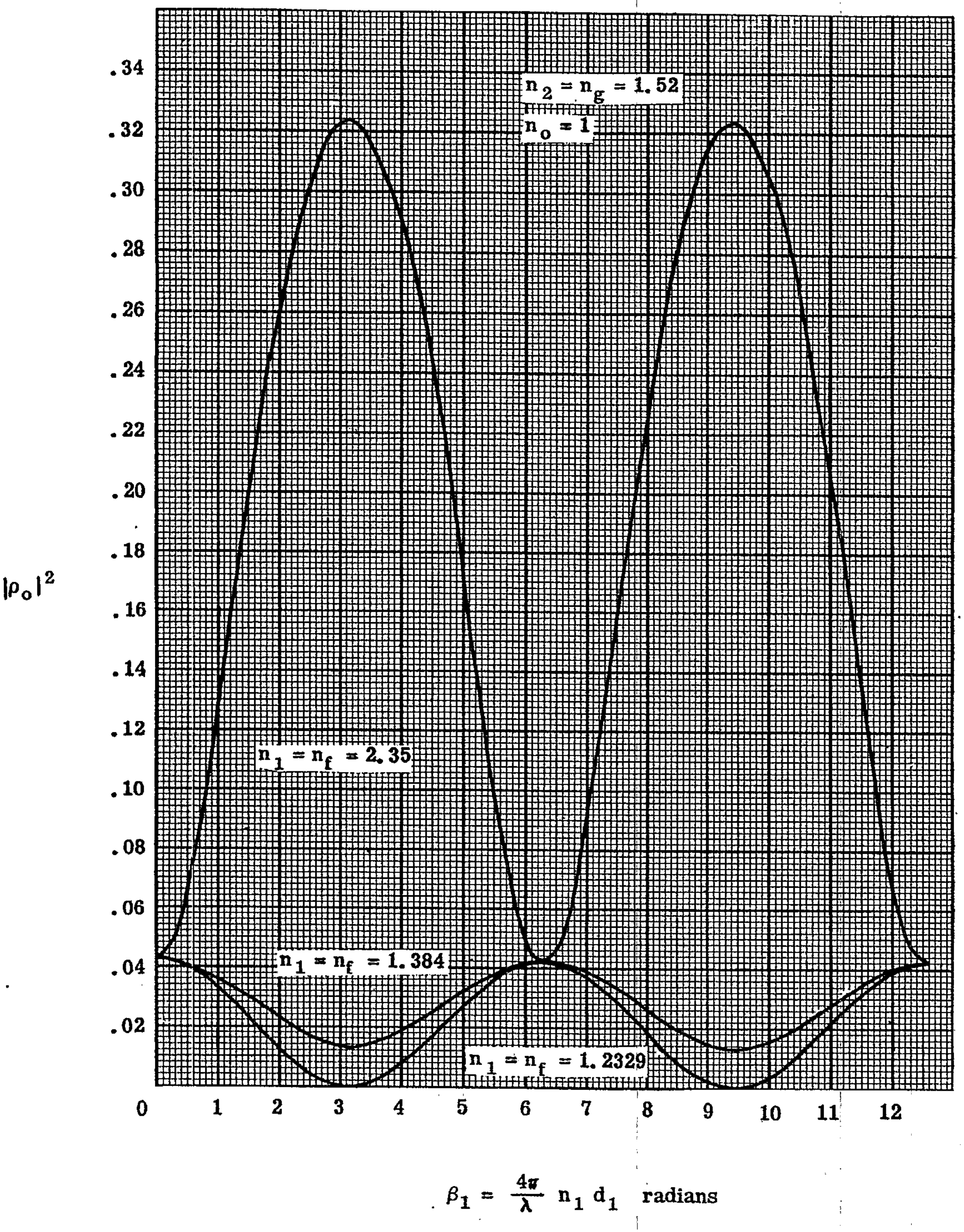


Figure 21. 11- Energy reflectances $|\rho_o|^2$ vs β_1 in radians. These curves illustrate the periodic behavior of nonabsorbing monolayers on nonabsorbing substrates. β_1 is now twice the optical path of the monolayer. The curves drawn for $n_f = 2.35$ and $n_f = 1.384$ with $n_g = 1.52$ correspond to zinc sulphide and magnesium fluoride, respectively, on spectacle crown glass. The curves illustrate the important characteristic that $|\rho_o|^2 \gtrless$ reflectance of the uncoated substrate according as $n_f \gtrless n_g$ at points $\beta_1 \neq \nu\pi$ where $\nu = 0$ or an integer. The curve for $n_f = 1.2329 = \sqrt{n_g} = \sqrt{1.52}$ illustrates how zero reflectance is achieved at points $\beta_1 = \mu\pi$ where μ is an odd integer.

of the interreflections that occur within the plate, it follows that

$$R_p = \frac{R + B - 2RB}{1 - RB} , \tag{1}$$

where R is the surface reflectance $R = \left|\rho_o\right|^2$. Hence

$$R = \left|\rho_o\right|^2 = \frac{R_p - B}{1 + R_pB - 2B} . \tag{2}$$

If the plate is a glass plate in air with its back surface uncoated, $B = \left[(n_g - 1)/(n_g + 1)\right]^2$. where n_g is the refractive index of the glass plate.

21. 6. 3. 3 When $n_o < n_1 > n_2$, equation (169) shows that Fresnel's coefficients W_1 and W_2 have opposite sign. At points $\beta_1 = \mu\pi$ (μ odd), the extreme reflectances R_m of equation (174) are now maxima with the choice of the negative sign. Also $\left|\rho_o\right|^2_{max.}$ is therefore given by the right hand member of equation (179). $\left|\rho_o\right|^2_{max.}$ approaches unity with increasing n_1 .

21. 6. 3. 4 In Figure 21. 12, $\left|\rho_o\right|^2$ is plotted against λ in the visible region for the indicated values of $n_1 = n_f$ and $n_2 = n_g$ with $n_o = 1$. The family of curves illustrates the effectiveness of various monolayers in reducing surface reflectances of spectacle crown glass. The thicknesses of the monolayers are chosen so that the optical path of each monolayer is one-fourth wavelength at 0. 550 microns.

21. 6. 3. 5 Energy reflectances $\left|\rho_o\right|^2$ are plotted against wavelength in the infrared region in Figure 21. 13 for $n_o = 1$, $n_2 = n_s = 3.450$ and for the indicated values of $n_1 = n_f$. This family of curves illustrates the effectiveness of various monolayers in reducing the reflectance of a surface of silicon. As applied to silicon, the effects of absorption by the substrate are not included in Figure 21. 13. Absorption of silicon is low in the infrared region. The curve for $n_f = 1.52$ has been added to illustrate what occurs when n_f is smaller than the value required for producing zero reflectance, i. e. when $n_f < \sqrt{n_o n_2}$. Let n_f and n_f' denote the refractive indices of two different non-absorbing monolayers such that $n_f > \sqrt{n_o n_2}$ and $n_f' < \sqrt{n_o n_2}$. It is not difficult to show that when the refractive indices are independent of wavelength, the two monolayers produce the same spectral reflectance curves provided that n_f and n_f' are chosen in accordance with the relation

$$n_f \; n_f' = n_1 \; n_1' = n_o \; n_2 \; . \tag{181}$$

For example, a monolayer having the refractive index $n_f' = 1.57$ is equivalent to the monolayer having the refractive index $n_f = 2.200$ when n_o and n_2 have been chosen as in Figure 21. 13.

21. 6. 4 Dielectric monolayers on opaque, metallic substrates; normal incidence. Dielectric layers of hard materials such as magnesium fluoride and silicon monoxide are often deposited upon surfaces of aluminum, silver and other metals for the purpose of protecting these surfaces from abrasion. The effects of monolayers of $M_g F_2$ and SiO upon the surface reflectance and phase change on reflection are illustrated in Figure 21. 14 for opaque substrates of aluminum. Appreciable losses in reflectance can occur when the monolayers are so thin that their optical paths $n_1 d_1 < \lambda/4$. For maximum reflectance, the optical path of the monolayer should be slightly less than $\lambda/2$. This maximum reflectance can exceed that of the uncoated surface provided that the refractive index n_1 of the monolayer is high enough. With $M_g F_2$[7] and SiO,[8] one obtains a maximum reflectance that departs imperceptibly from that of the uncoated substrate. The phase change on reflection varies markedly with the optical path of the monolayer. This property is of importance to interferometry.

21. 6. 5 Absorbing monolayers on opaque, metallic substrates; normal incidence. The effects of absorbing monolayers upon the surface reflectance of opaque, metallic substrates are illustrated in Figure 21. 15 by a series of monolayers that have a fixed and relatively high refractive index $n_1 = 4.0$. The curve for the non-absorbing film $n_1 K_1 = 0$ shows that the maximum reflectance can exceed * that of the uncoated surface appreciably when n_1 becomes high. Increasing the absorption of the monolayer within the range of $n_1 K_1$ of Figure 21. 15 reduces both the maximum and the minimum reflectances. The minimum reflectances occur for optical paths p_1 near 45° or $\lambda/8$. The family of curves suggest that tarnishing of silver is due to the formation of a monolayer.

[7] Monolayers of MgF_2 have become valuable for increasing the reflectance of aluminum in the ultraviolet. For a discussion of the region around 1216A, see P.H. Berning, G. Hass and R. P. Madden; Paper T51 given during 44th Annual Meeting of OSA, Ottawa, October 1959.

[8] The refractive indices and absorption of so called SiO films depend markedly upon conditions of evaporation. Slow deposition in the presence of air or oxygen produces "SiO" films that can have refractive indices near 1.7 with low absorption. See G. Hass, Vacuum, 2, p 338 (1952).

* Compare with Figure 21.14

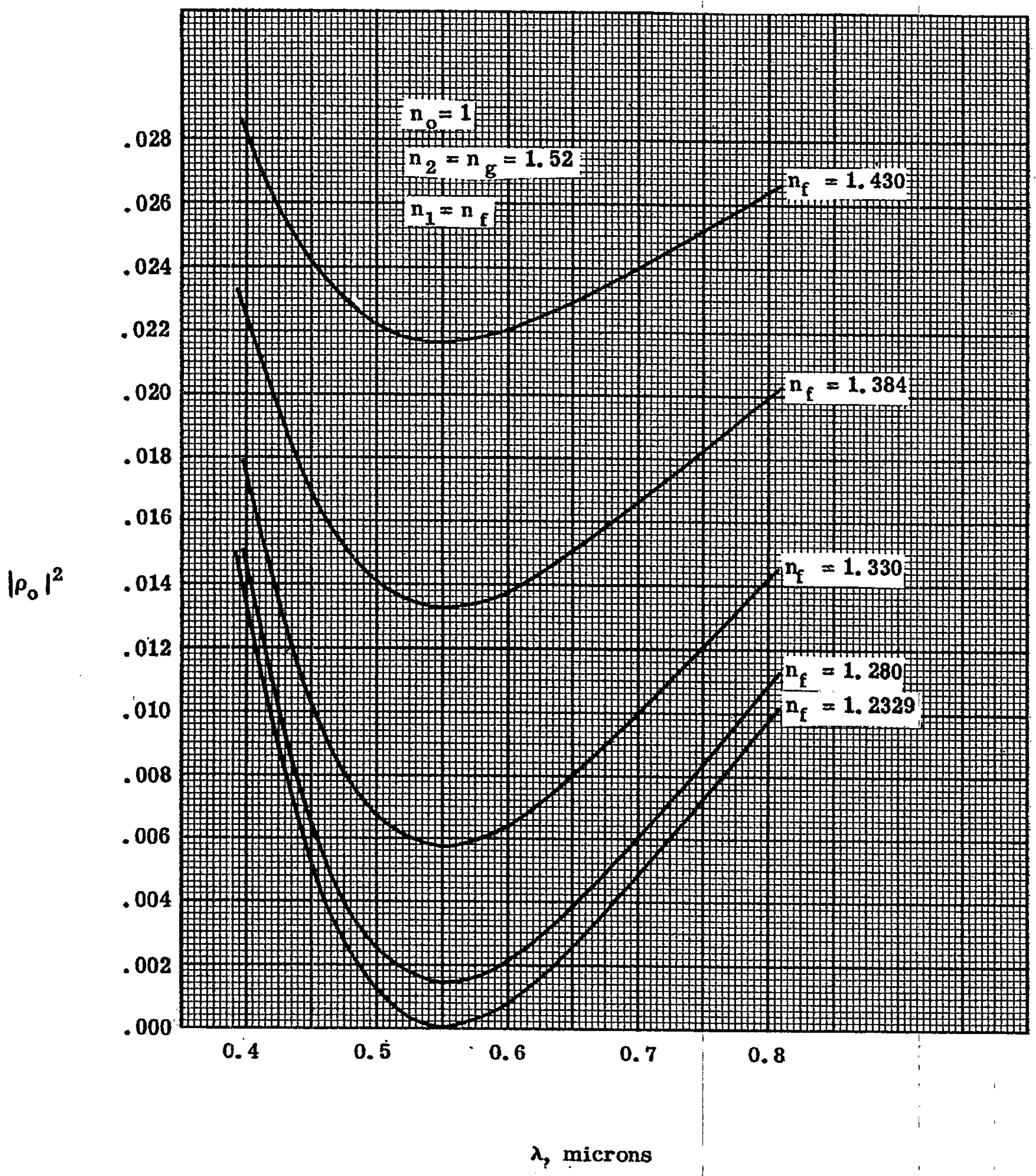


Figure 21. 12- Plot of energy reflectances vs wavelength in microns for various low reflecting monolayers on spectacle crown glass. Each monolayer has the optical path $\lambda/4$ at $\lambda = 0.550$ microns.

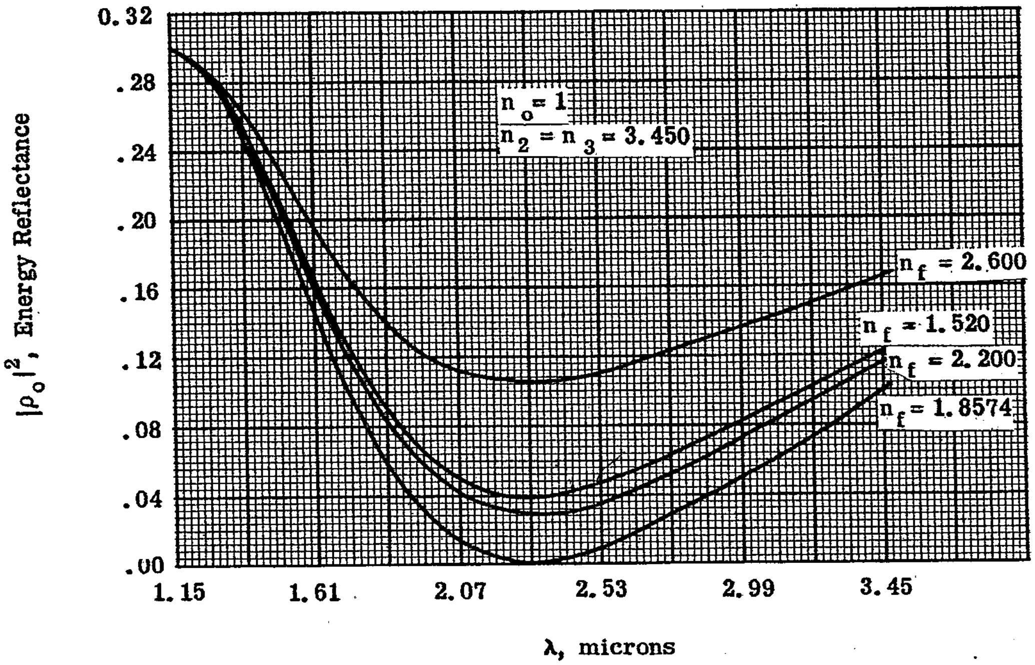


Figure 21. 13- Low reflecting monolayers on substrates having high refractive index.

21. 6. 6 Metallic monolayers on glass; normal incidence. The curves of Figure 21. 16 a and b illustrate the optical behavior of monolayers of silver on non-absorbing substrates such as glass. Layers of silver differ from layers of other metals mainly in that layers of silver have remarkably low * absorption although the nK -value is moderately high. Silver is sensibly opaque of thicknesses near 0. 15λ. It should be observed that the reflectance at the glass to film interface is less than that at the air to film interface -- a characteristic of most metals. As the thickness of the silver layer is increased, the phase changes on reflection pass into the third quadrant for reasons discussed following equation (16) in Section 21. 2. 3. The reflectance curve for reflection from glass to silver exhibits theoretically a minimum at a thickness between zero and 0. 01λ.

21. 6. 7 Non-absorbing systems; oblique incidence. The manner in which low reflecting, non-absorbing monolayers modify the reflectances $|\rho_o|^2$ and $|\gamma_o|^2$ of surfaces of non-absorbing substrates is illustrated ** by Figures 21. 17, 21. 18, and 21. 19 for the 45° angle of incidence. The spectral reflectance curves of Figure 21. 17 for $|\rho_o|^2$ and $|\gamma_o|^2$ correspond to electric vectors that vibrate respectively perpendicular and parallel to the plane of incidence for a monolayer of $M_g\ F_2$ on a surface of spectacle crown glass. Whereas a monolayer of $M_g\ F_2$ is effective in reducing both $|\rho_o|^2$ and $|\gamma_o|^2$ for spectacle crown glass, the reflectance $|\rho_o|^2$ is still quite high. Figure 21. 18 shows that the refractive index of the monolayer must approach the value $n_1 = 1.2$ in order to reduce $|\rho_o|^2$ below 1% for spectacle crown glass. Rugged films having refractive indices below 1. 30 are not available. On the other hand, Figure 21. 19 shows that one can choose a relatively high and available value of n_1 for reducing both $|\rho_o|^2$ and $|\gamma_o|^2$ to values below 1. 2% when the refractive index n_2 of the substrate is markedly higher than that of spectacle crown glass. Figure 21. 19 illustrates also the fact that $|\rho_o|^2_{min.} = |\gamma_o|^2_{min.}$, when $n_1 = \sqrt{n_o n_2}$. The effect of increasing or decreasing the angle of incidence from 45° is to raise or to lower, respectively, the reflectance at the crossing point C of Figure 21. 19. It is to be expected that the reduction of $|\rho_o|^2$ presents the more formidable problem; for $|\gamma_o|^2$ is automatically zero at Brewster's angle of incidence.

*The amount of absorption is altered considerably by the conditions of evaporation. L. G. Schultz gives the values $n = 0.055$ and $nK = 3.32$ for silver at $\mu = 0.55$.

**As with many other subjects of this text, the possible number of instructive illustrations is restricted in the interests of conserving space.

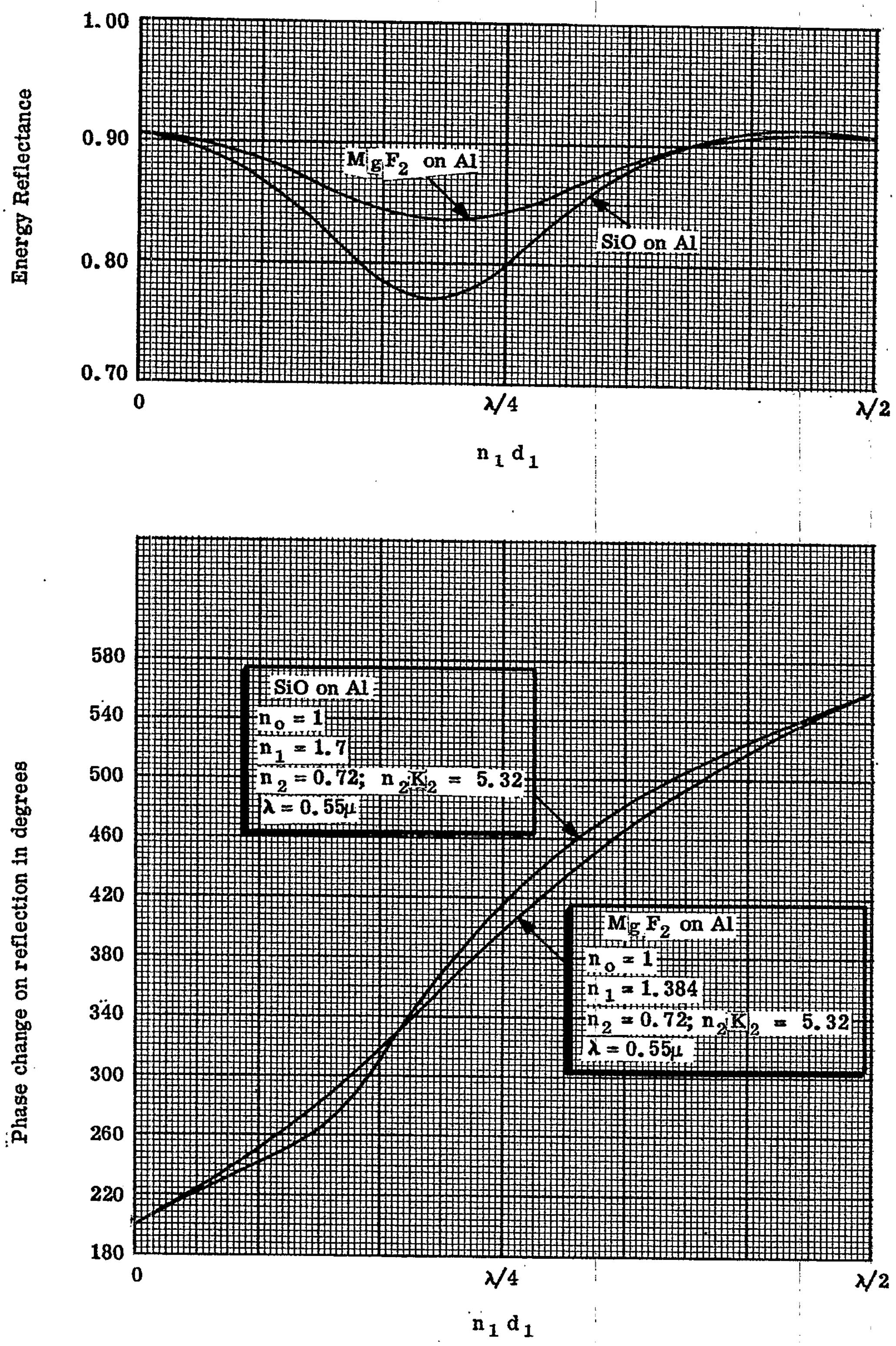


Figure 21. 14- Energy reflectances and phase changes on reflection vs optical path of the film in wavelengths for MgF_2 and SiO on opaque substrates of aluminum. Phase changes on reflection appear as phase retardations. Absorption by SiO monolayers has been neglected. The optical constants of aluminum are those given by L. G. Schultz, J. Opt. Soc. Amer., 44, 357-368 (1954).

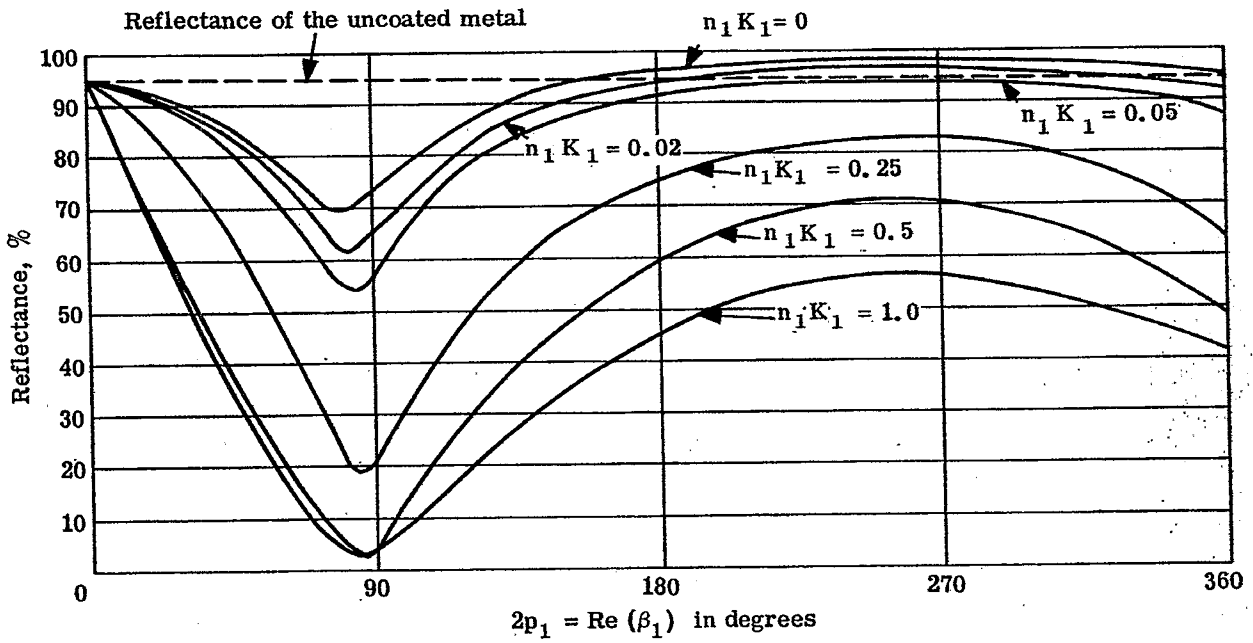


Figure 21. 15- Reflectances from various absorbing monolayers on a metallic substrate as functions of $R_e\,(\beta_1) = 2p_1$ where p_1 is the optical path through the monolayer. The metallic substrate corresponds to silver with the optical constants $n_2 = 0.15$ and $n_2K_2 = 3.28$. The monolayers have the high refractive index $n_1 = 4.0$ and the indicated values of n_1K_1. The case $n_1 K_1 = 0$ is the nonabsorbing monolayer.

21. 7 BILAYER COATINGS

21. 7. 1 Introduction. Bilayers possess advantages over monolayers for decreasing or increasing the reflectance of surfaces of dielectrics or metals. As examples, zero reflectance can be obtained at an assigned wavelength by suitable choice of the thickness ratio of the two layers and a marked degree of control over the distribution of spectral reflectance or transmittance becomes possible. Bilayers are superior to monolayers for beam splitters and for control of phase changes that occur on reflection or transmission. Whereas absorbing bilayers are used for such purposes as controlling the transmittance of sunglasses, the most important bilayers are predominantly dielectric. Low reflecting bilayers fall into well defined groups whose characteristics will be discussed.

21. 7. 2 Methods of computation. Matrix and quaternion methods have been treated in Sections 21. 4 and 21. 5. The reader who is inclined toward watching the interfacial reflectances and transmittances and toward using recursion formulae may choose the method of Sections 21. 2. 8 to 21. 2. 11. The convention and notation with respect to bilayers is shown in Figure 21. 20. Suppose that the electric vector is perpendicular to the plane of incidence. We obtain the complex reflectance, ρ_o, in the form

$$\rho_o = \frac{\rho_1\, e^{i\beta_1} + W_1}{1 + W_1\, W_2\, e^{i\beta_1}}, \tag{182}$$

wherein

$$\rho_1 = \frac{W_3\, e^{i\beta_2} + W_2}{1 + W_2\, W_3\, e^{i\beta_2}}. \tag{182a}$$

The Fresnel coefficients of reflection W_ν are defined, as usual, by equations (19) and (20). β_ν ($\nu = 1, 2$) is defined by equation (56). The complex transmittance, τ_3, corresponding with ρ_o is obtained by setting $N = 2$, $\tau_o = 1$ and $\rho_2 = W_3$ in equation (54). The complex reflectance, γ_o, for cases in which the electric vector vibrates in the plane of incidence can be computed from equation (182), after replacing ρ_ν by γ_ν and W_ν by F_ν as defined by equation (21). The corresponding complex transmittance is denoted by T_3 and is obtained by setting $N = 2$, $T_o = 1$ and $\gamma_2 = F_3$ in equation (70). At normal incidence, $\gamma_o = \rho_o$ and $T_3 = \tau_3$. Closed formulae are available for the energy reflectances. These formulae simplify markedly

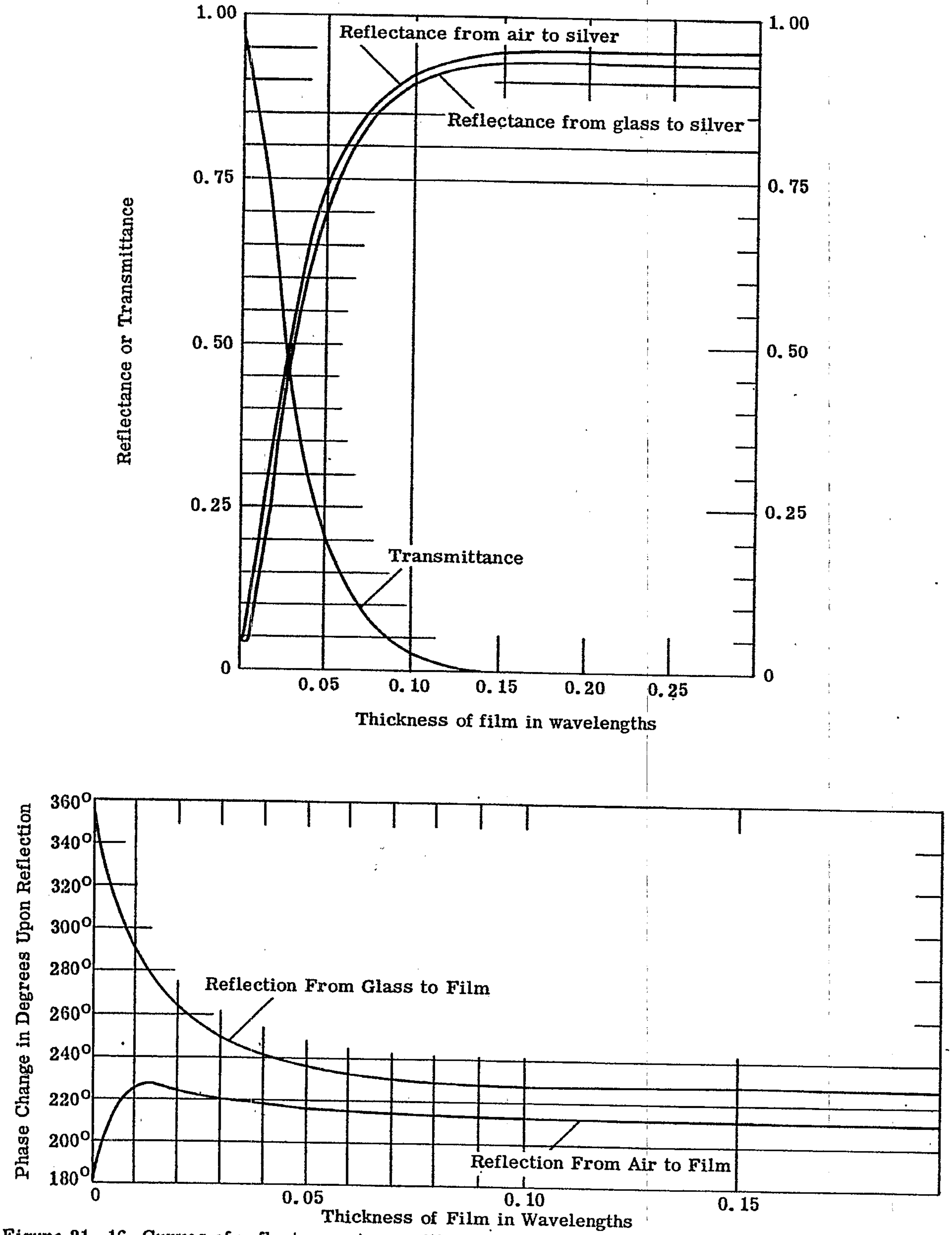


Figure 21. 16- Curves of reflectance, transmittance and phase changes on reflection vs thickness of the monolayer in wavelengths for silver films on spectacle crown glass. These curves have been computed for glass having the refractive index 1.52 and for an absorbing film having the optical constants $n_f = 0.15$ and $n_f K_f = 3.28$. These optical constants predate those measured for silver by L. G. Schultz and belong to a wavelength near 0.55 microns. The phase changes of Figure 21.16b are retardations. Phase retardations of 230° can be regarded, if desired, as phase advances of 130°.

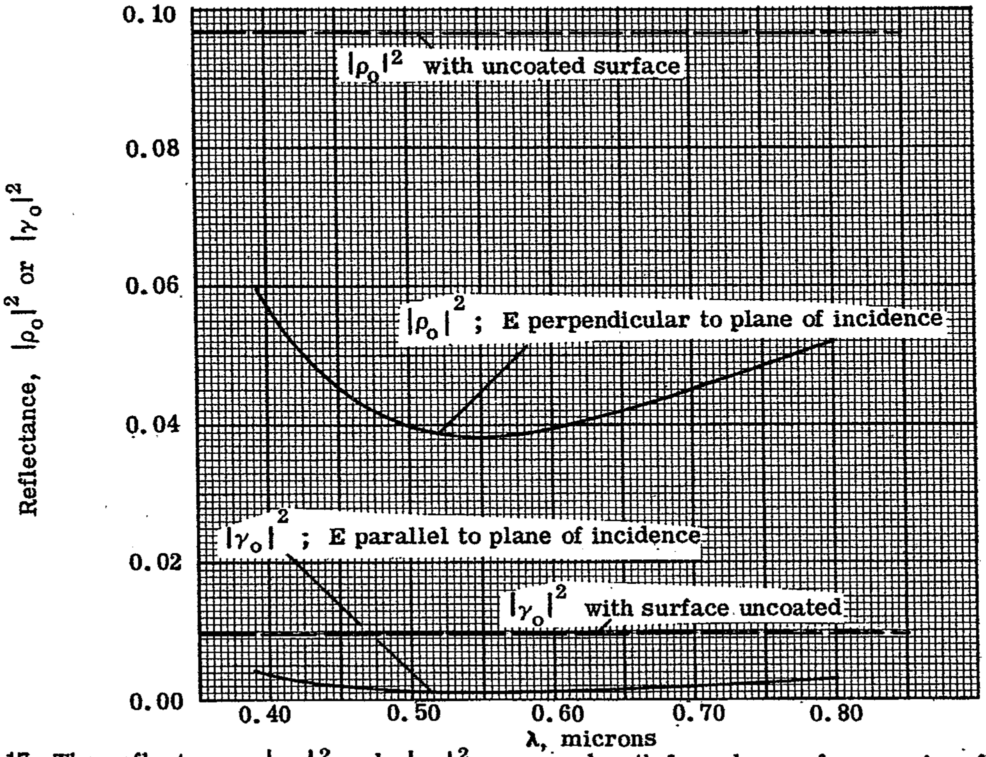


Figure 21. 17- The reflectances $|\rho_o|^2$ and $|\gamma_o|^2$ vs wavelength for a layer of magnesium fluoride on spectacle crown glass. The refractive indices are $n_o = 1$; $n_1 = 1.384$ and $n_2 = 1.52$ and the angle of incidence is 45°. The thickness of the film has been chosen so that $\beta_1 = 4\pi n_1 d_1 \cos i_1/\lambda = \pi$ at $\lambda = 0.55\mu$.

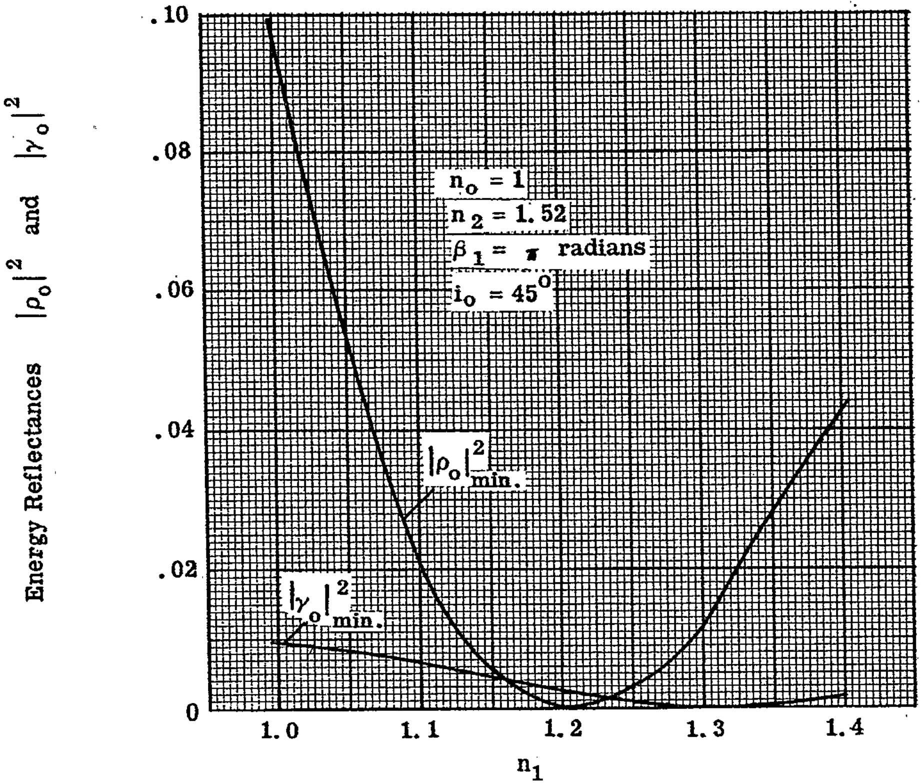


Figure 21. 18- Plot of the minimum reflectances $|\rho_o|^2_{min.}$ and $|\gamma_o|^2_{min.}$ against the refractive indices n_1 of the monolayers on spectacle crown glass at 45° angle of incidence.

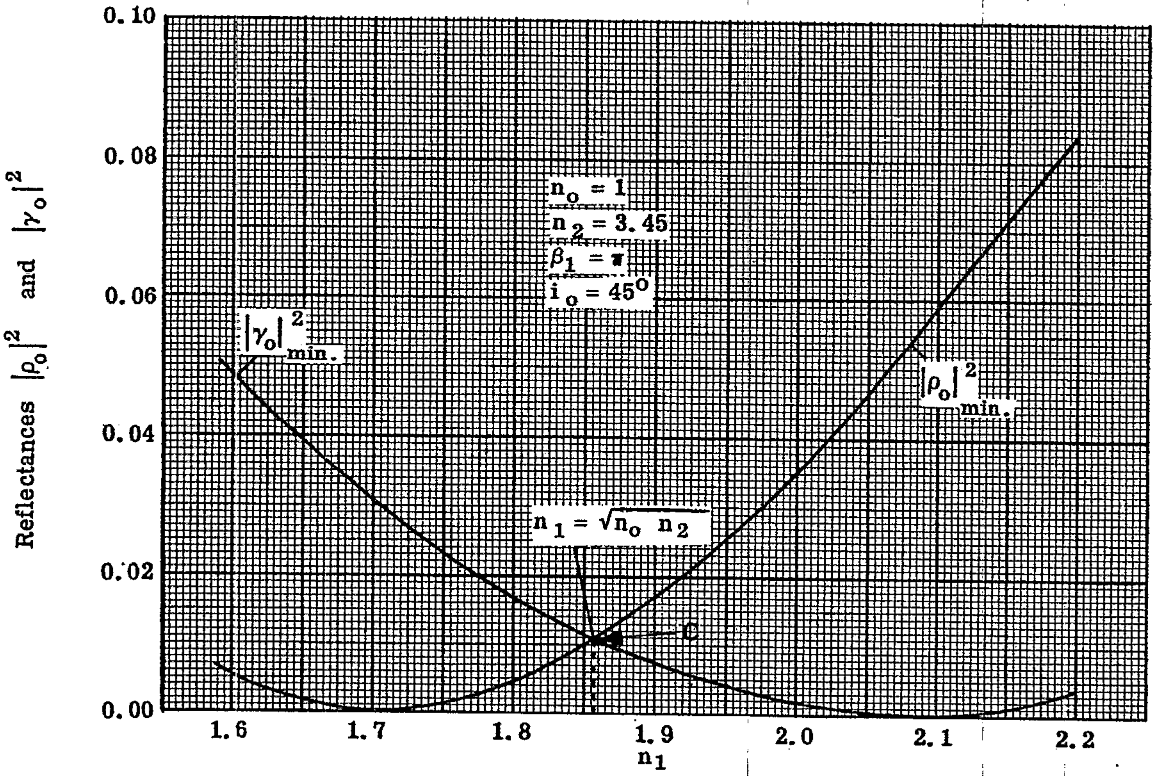


Figure 21. 19- Plot of the minimum reflectances $|\rho_o|^2_{min.}$ and $|\gamma_o|^2_{min.}$ against n_1 for monolayers on a nonabsorbing substrate of high refractive index n = 3.45 at the fixed angle 45° of incidence.

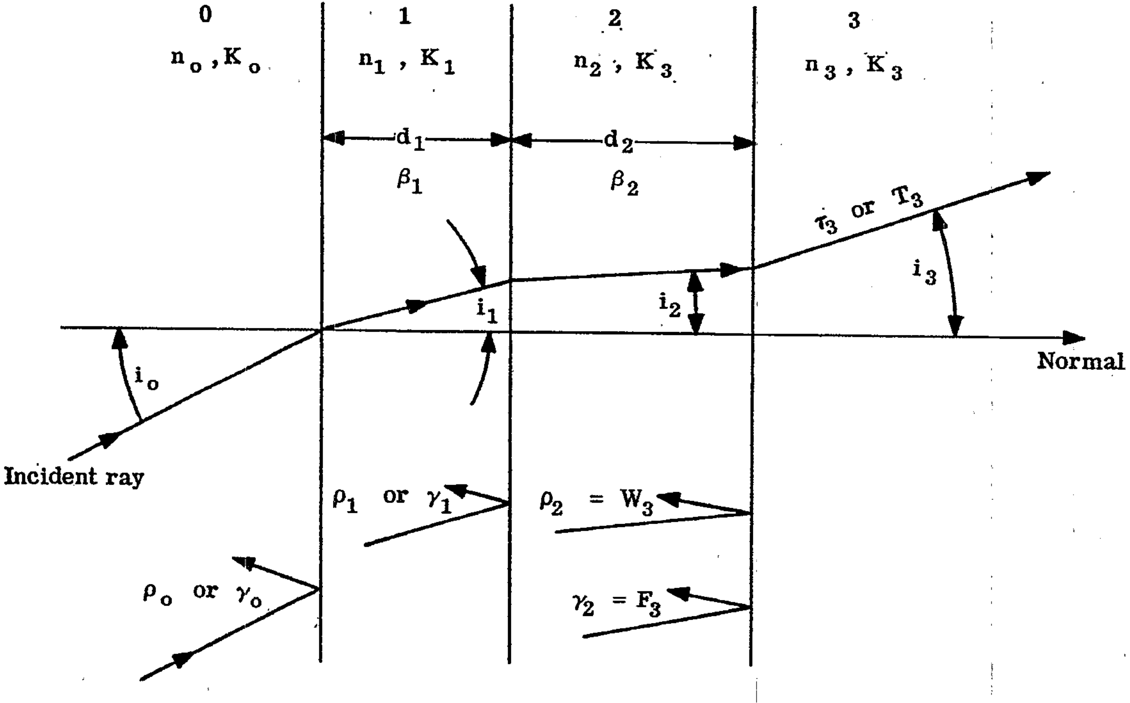


Figure 21. 20- Notation and convention with respect to bilayers (case N = 2).

only when all of the media and layers are non-absorbing and when total internal reflectance need not be considered. Thus,

$$|\rho_o|^2 = 1 - \frac{(1-W_1^2)(1-W_2^2)(1-W_3^2)}{D}, \tag{183}$$

where

$$D = 1 + W_1^2 W_2^2 + W_1^2 W_3^2 + W_2^2 W_3^2 + 2 W_1 W_2 (1 + W_3^2) \cos \beta_1 + 2 W_2 W_3 (1 + W_1^2) \cos \beta_2 + 2 W_1 W_3 \cos (\beta_1 + \beta_2) + 2 W_1 W_3 W_2^2 \cos (\beta_1 - \beta_2). \tag{183a}$$

W_ν and β_ν are now real with

$$\beta_\nu = \frac{4\pi}{\lambda} n_\nu d_\nu \cos i_\nu ; \quad \nu = 1, 2. \tag{183b}$$

Equation (183) can be used to compute $|\gamma_o|^2$ by replacing ρ_o by γ_o and W_ν by F_ν. Because absorption has been excluded, the sum of the energy reflectance and the energy transmittance must be unity when

$$|\tau_o| = |T_o| = 1 .$$

21.7.3 The simplest low-reflecting bilayer. Monolayers such as $M_g F_2$ on glass fail to produce zero reflectance because $n_1 > \sqrt{n_o n_2}$. This class of monolayers is characterized by the fact that $n_o < n_1 < n_2$. Consequently, $W_1 < 0$ and $W_2 < 0$ at normal incidence. The following conclusions are not difficult to ascertain from equation (169) for normal incidence.

$$|W_2| < |W_1| \quad \text{when} \quad n_1 > \sqrt{n_o n_2} ; \tag{184}$$

$$|W_2| > |W_1| \quad \text{when} \quad n_1 < \sqrt{n_o n_2} . \tag{184a}$$

Hence the absolute value of the Fresnel coefficient W_2 at the film-to-glass interface is too small to satisfy the condition for zero reflectance when $n_1 > \sqrt{n_o n_2}$; for $W_1 = W_2$ at the point of zero reflectance. These considerations suggest that it should be possible to decrease the energy reflectance from monolayers for which $n_1 > \sqrt{n_o n_2}$ by depositing either a thin dielectric film that has high refractive index or a thin film of highly reflecting metal upon the interface between media no. 1 and 2 as illustrated in Figure 21.21. The author[9] has shown that metals can be used to augment the interfacial reflectance for obtaining zero reflectance. An advantage of this method is that a wide range of dielectric materials can be used as the monolayer by making suitable choices of the thickness of the thin metallic film. A disadvantage of employing metallic or absorbing films for augmenting the interfacial reflectance is that the energy reflectance for incidence from glass to film can be quite high even when the energy reflectance in the opposite direction from air to film has been made zero. This disadvantage as regards the irreversibility of the reflectance can be avoided by using dielectric materials of high refractive index as the reflectance augmenting layer in the manner described by Osterberg, Kashdan and Pride.[10] The use of dielectrics having high refractive index instead of metal yielded some unexpected advantages that will now be discussed.

21.7.3.1 Summary of the theory. In summarizing the results of an unpublished theory of the author, let us utilize the convention of Figure 21.22 in which layer no. 1 is to be the thin dielectric layer of high refractive index. All of the media are non-absorbing and the incidence is to be normal. One can show from the zero condition ($\rho_o = 0$) of equation (98) that the value of β_1 required for zero reflectance is given by

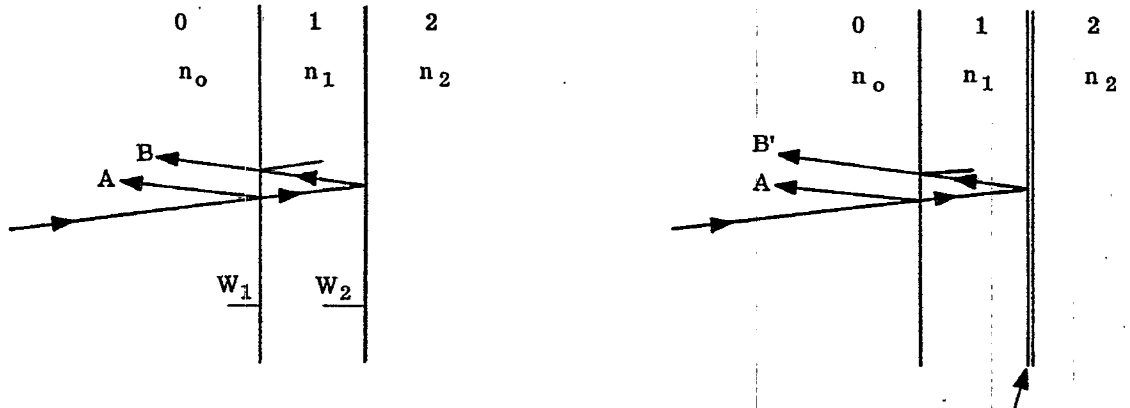
$$\cos \beta_1 = \frac{W_1^2 + W_2^2 - W_3^2 (1 + W_1^2 W_2^2)}{2 W_1 W_2 (W_3^2 - 1)}, \tag{185}$$

in which

$$\beta_1 = \frac{4\pi}{\lambda} n_1 d_1 \text{ radians}. \tag{185a}$$

[9] U. S. Patent 2,366,687 (Jan. 2, 1945).

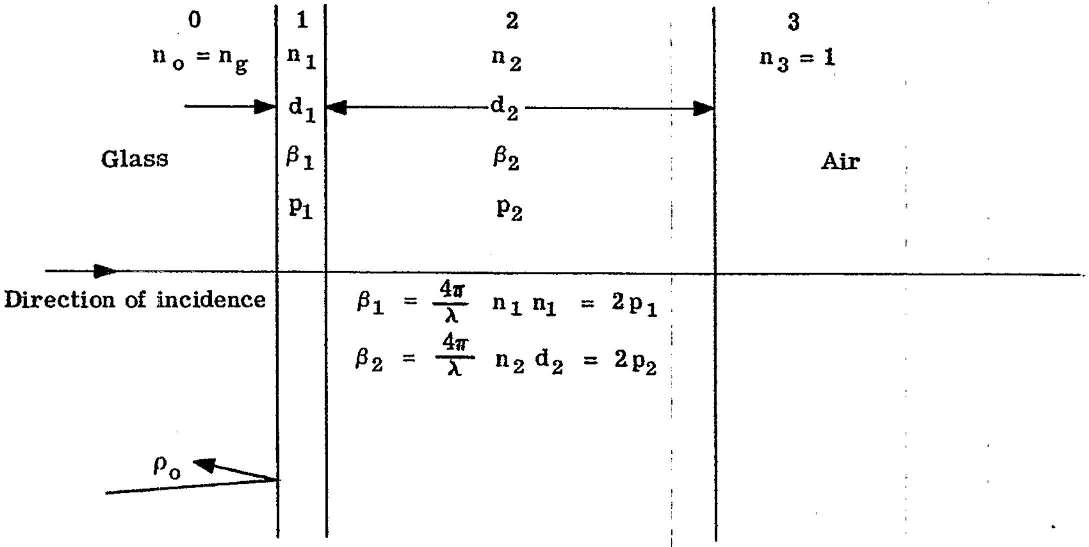
[10] See abstracts of papers published in J. Opt. Soc. Amer., 42, p. 291 (1952).



If $n_1 > \sqrt{n_o n_2}$, then $|W_2| < |W_1|$ and $|B| < |A|$.

Augment the reflectance of this interface so that |vector B'| = |vector A| .

Figure 21. 21- Illustration of the principle of augmenting one of the interfacial reflectances of a monolayer so as to obtain zero or decreased reflectance by depositing a thin, highly reflecting layer upon the interface.



$$\beta_1 = \frac{4\pi}{\lambda} n_1 n_1 = 2p_1$$

$$\beta_2 = \frac{4\pi}{\lambda} n_2 d_2 = 2p_2$$

Figure 21. 22- Notation with respect to the simplest low reflecting bilayer. Layer no. 1 is a thin film of high refractive index n_1 . p_1 and p_2 are the optical paths of the two layers.

Correspondingly, in terms of the refractive indices

$$\cos \beta_1 = 1 + \frac{2n_1^2 (n_g - 1)\ (n_g - n_2^2)}{(n_1^2 - n_g^2)\ (n_1^2 - n_2^2)}, \qquad (185b)$$

for cases in which $n_3 = n_{air} = 1$. For solutions to exist, it is necessary that

$$\cos \beta_1 \leqq 1, \qquad (185c)$$

a condition that places restrictions upon the refractive indices of equation (185b). Of these restrictions, the one of greatest interest is

$$n_1^2 > n_2^2\ n_g > \left\{ \begin{matrix} n_2^2 \\ n_g^2 \end{matrix} \right. > n_g\,. \qquad (185d)$$

If equation (185b) has a solution $\beta_1 = \beta_{10}$, it has other solutions that differ from β_{10} by integral multiples of 2π. We are to choose the smallest possible solution for which

$$0 < \beta_1 < \pi. \qquad (185e)$$

Correspondingly, the optical path p_1 of the thin layer will be less than $\pi/2$. It can be much less than $\pi/2$, as we shall see. Having computed β_1, one can compute the associated value of β_2 from β_1 in the following manner. Let

$$\theta_1 = \tan^{-1} \frac{-W_1 \sin \beta_1}{-(W_1 \cos \beta_1 + W_2)}; \qquad (186)$$

$$\theta_2 = \tan^{-1} \frac{W_1 W_2 \sin \beta_1}{1 + W_1 W_2 \cos \beta_1}. \qquad (186a)$$

Then

$$-\beta_2 = \theta_1 - \theta_2; \quad (\theta_2 > \theta_1). \qquad (186b)$$

There exists no ambiguity as to the quadrant in which θ_1 and θ_2 fall because

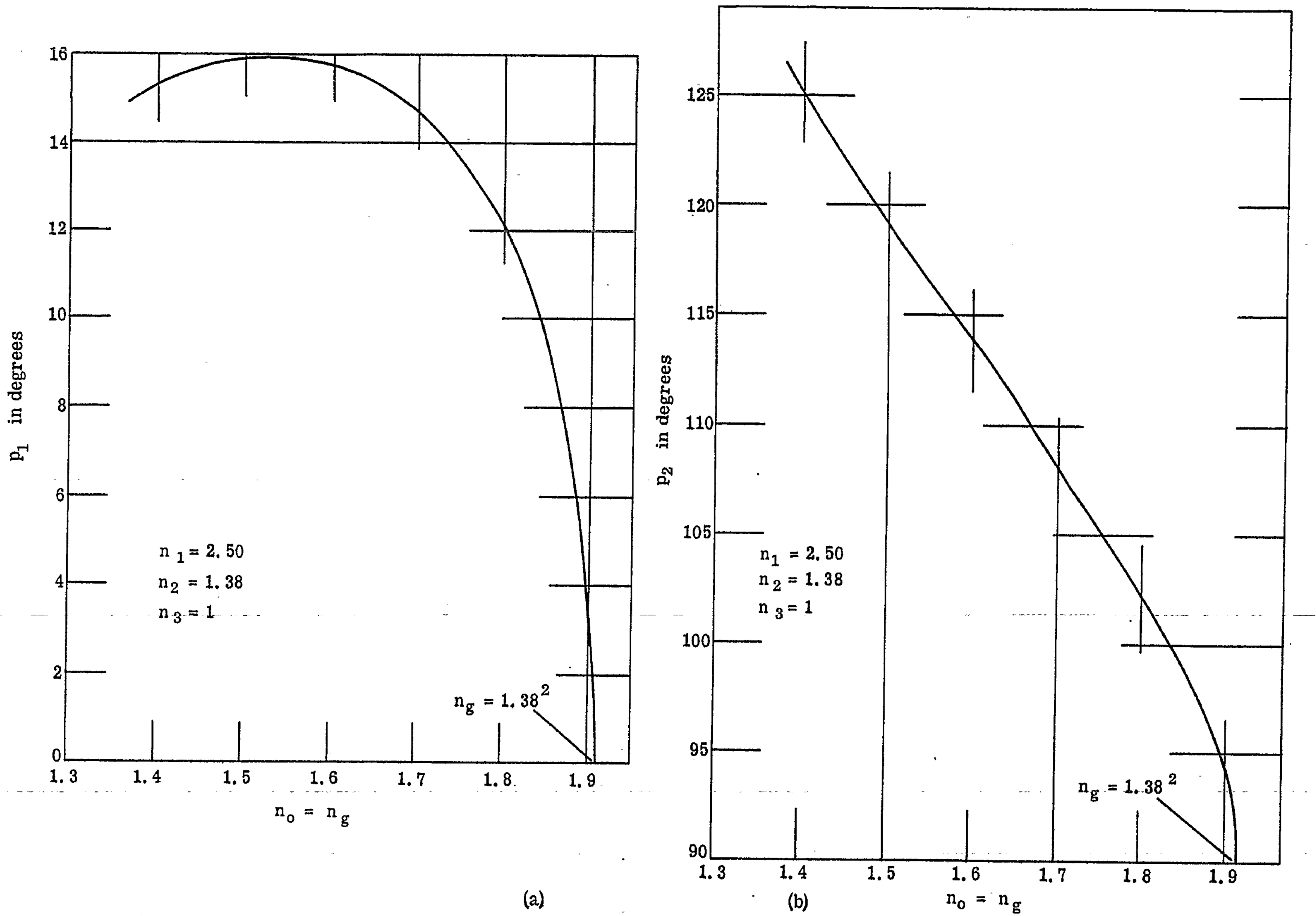
$$\begin{aligned} \sin \theta_1 &: -W_1 \sin \beta_1; & \sin \theta_2 &: W_1 W_2 \sin \beta_1; \\ \cos \theta_1 &: -(W_1 \cos \beta_1 + W_2); & \cos \theta_2 &: 1 + W_1 W_2 \cos \beta_1; \end{aligned} \qquad (186c)$$

in which the factor of proportionality is greater than zero. Equations (185) and (186) thus enable one to compute * the doubled optical paths β_1 and β_2, Figure 21.12, of the two layers when n_1 has been suitably chosen with respect to n_g and n_2.

21.7.3.2 Let us now consider the case: $n_1 = 2.50$; $n_2 = 1.38$; $n_3 = 1$. This case applies to bilayers comprised of a thin inner layer of TiO_2 and an outer layer of $M_g F_2$ under conditions at which the refractive indices of TiO_2 and $M_g F_2$ are 2.50 and 1.38, respectively. The optical paths p_1 and p_2 required for vanishing reflectance have been computed from equations (185) and (186) and plotted as functions of n_g in Figure 21.23 (a) and (b). Curiously, the required optical path p_1 of the inner layer with refractive index $n_1 = 2.50$ is substantially 15.5° or 0.043λ for the range n_g of ordinary glasses. This fact has been confirmed experimentally and means that a TiO_2 layer of fixed thickness will serve for practically all glasses. This thickness is only $0.043/2.5 = 0.0172\lambda$. The required thickness approaches zero as n_g approaches $n_2^2 = 1.38^2$. The corresponding optical path p_2 of the outer layer must be chosen with some care since the slope of the curve of Figure 21.23(b) is quite marked. The optical path, p_2, exceeds the quarter wave condition considerably in the range, n_g, for ordinary glasses. However, the main effect of altering the optical path of the outer layer is to alter the wavelength at which minimum reflectance occurs. This class of bilayer is therefore relatively easy to produce. It is extremely durable. Cerium oxide is to be preferred to titanium dioxide as the inner layer because cerium oxide can be evaporated as a dielectric material without need for subsequent heating and oxidation. Equations (185) and (186) should be applied to redetermine p_1 and p_2 when cerium oxide is used. Comparisons of the bilayer (theoretical and experimental) with a monolayer of

*Frequently, one can be quite certain about the quadrant in which β_2 must fall. Under such circumstances it is simpler to compute β_2 from the formula

$$\cos \beta_2 = \frac{W_2^2 + W_3^2 - W_1^2\ (1 + W_2^2\ W_3^2)}{2 W_2\ W_3\ (W_1^2 - 1)}.$$



(a) (b)

Figure 21.23- Plot of the optical paths p_1 and p_2 against the refractive index n_g of the substrate. p_1 and p_2 are the optical paths required to obtain zero reflectance for the layer of high refractive index $n_1 = 2.50$ and the layer of low refractive index n_2, respectively.

$M_g F_2$ on spectacle crown glass of refractive index $n_g = 1.52$ are made in Figure 21.24. The theoretical and experimental curves for the bilayer are in good agreement and exhibit much lower reflectances than monolayers of $M_g F_2$ near the wavelength for minimum reflectance. The dispersion of the reflectance is, however, markedly greater for the bilayer.

21.7.3.3 With reference to white light reflectance of the bilayer, T. Sawaki and H. Kubota investigated bilayers having low white light reflectance and found that those bilayers which simulate monolayers having augmented interfacial reflectance rank among the bilayers exhibiting the lowest white light reflectances.

21.7.4 Achromatized bilayers.

21.7.4.1 Curves of the computed * spectral reflectances at normal incidence for various achromatized bilayers on spectacle crown glass are illustrated in Figures 21.25 and 21.26. The outer layer is exemplified by $M_g F_2$. It has a refractive index n_1 which fails to meet the zero condition, $n_1 = \sqrt{n_o n_2}$, for monolayers. The outer and inner layers have the optical path $\lambda/4$ and $\lambda/2$, respectively, at a chosen ** wavelength, λ_m. The inner layer of refractive index n_2 is therefore an absentee layer at $\lambda = \lambda_m$. Consequently, the reflectance R_m at $\lambda = \lambda_m$ is due to the outer layer and the substrate of refractive index $n_3 = n_g$. When $n_2 > n_1$, the reflectance R_m is a maximum for wavelengths near λ_m. Achromatization occurs because the energy reflectance must drop before it can rise as λ is varied on either side of λ_m. With respect to the curves of Figure 21.25, the minimum reflectances on each side of point R_m are equal. Bilayers exhibiting this property will be classified as isoachromatic. The two minima of the spectral reflectance curve for the case $n_1 = 1.38$, $n_2 = 1.86$ and $n_2 = 1.52$ are zeros. Isoachromatic bilayers displaying zero minima will be called null-isoachromatic bilayers. The publications by A. F. Turner [11] and A. Vasicek [12] deal with null-isoachromatic bilayers.

21.7.4.2 The spectral reflectance curve for $n_2 = 1.52$, i.e. for the comparison monolayer of $M_g F_2$, of Figures 21.25 and 21.27 lies above the other spectral reflectance curves except at extreme values of λ and β_1. To substitute any one of the isoachromatic bilayers of Figures 21.25 and 21.27 in place of the monolayer of refractive index n_1 will therefore reduce the white light reflectance. The class of bilayer discussed in Section 10.7.3 and its subsections is, however, much superior for reducing white light reflectance. Except for specialized applications, achromatic bilayers are to be preferred when increased neutrality of spectral reflectance becomes important. For this purpose, the null-isoachromatic bilayers are not as suitable as the isoachromatic bilayers exemplified by the curve for $n_2 = 1.55$ of Figure 21.27. The flatness of this curve is quite remarkable. Comparison of Figures 21.25 and 21.26 shows that a wide range of distributions of spectral reflectances can be attained by means of achromatized bilayers. Figure 21.28 has been included to show how the separation of the minima at points A and B can be reduced by choosing n_1 nearer to the value $n_1 = \sqrt{n_o n_g}$.

21.7.4.3 The theoretical design of null-achromatic bilayers will now be developed algebraically in a manner that illustrates one use of the method of admittances. From equations (81a) and (81b), one obtains for normal incidence the results,

$$Y_3 = -n_3; \tag{187}$$

$$Y_1 = n_1 \frac{Y_2 - i n_1 \tan p_1}{n_1 - i n_2 \tan p_1}; \quad p_1 = \frac{\beta_1}{2}; \tag{187a}$$

$$Y_2 = n_2 \frac{Y_3 - i n_2 \tan p_2}{n_2 - i Y_3 \tan p_2}; \quad p_2 = \frac{\beta_2}{2}. \tag{187b}$$

The condition for zero reflectance is $\rho_o = 0$ or, from equation (79a),

$$Y_1 = -M_o = -n_o = -1 \tag{188}$$

since we shall suppose that the medium of incidence is vacuum. By eliminating the admittances Y_1, Y_2 and Y_3, one obtains quite directly the condition for zero reflectance in its general form for non-absorbing bilayers, namely,

$$n_1 \frac{1 - i n_1 \tan p_1}{n_1 - i \tan p_1} = n_2 \frac{n_3 + i n_2 \tan p_2}{n_2 + i n_3 \tan p_2}; \tag{188a}$$

* The materials of this discussion have been taken from unpublished research notes of the author.

** λ_m is often chosen as 0.55μ when the bilayer is intended for the visible region.

[11] A. F. Turner, J. de Phys., 11, 444 (1950).

[12] A. Vasicek, Optica Acta, May 1951, special issue, pp. 20-25.

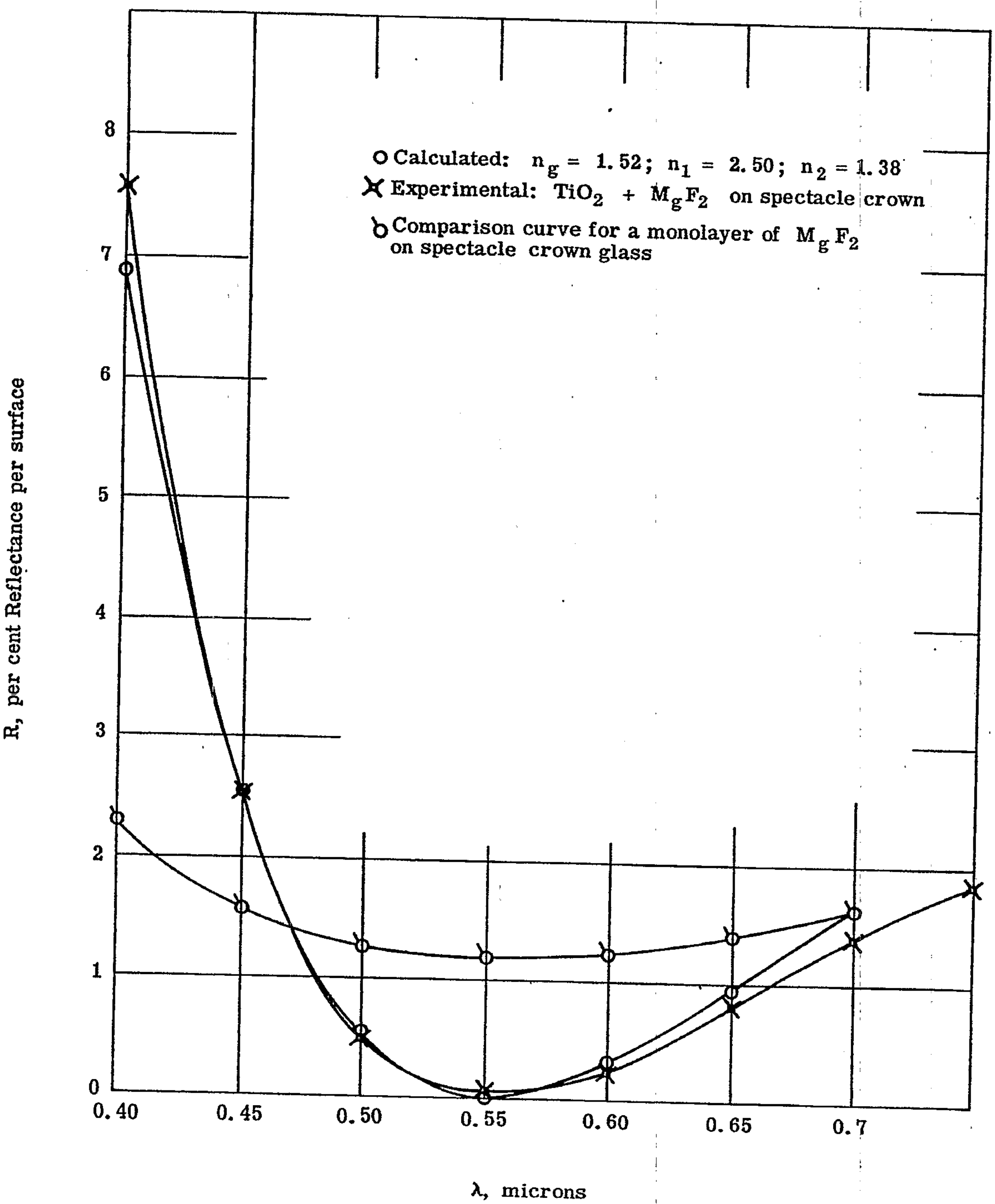


Figure 21.24- Comparison of the computed and experimental reflectances from bilayers of $TiO_2 + M_gF_2$ on spectacle crown glass of refractive index 1.52. In making the computations it has been assumed that the refractive indices remain fixed at the values $n_g = 1.52$; $n_1 = 2.50$ and $n_2 = 1.38$. A spectral reflectance curve for monolayers of $M_g F_2$ on spectacle crown glass is included for further comparison.

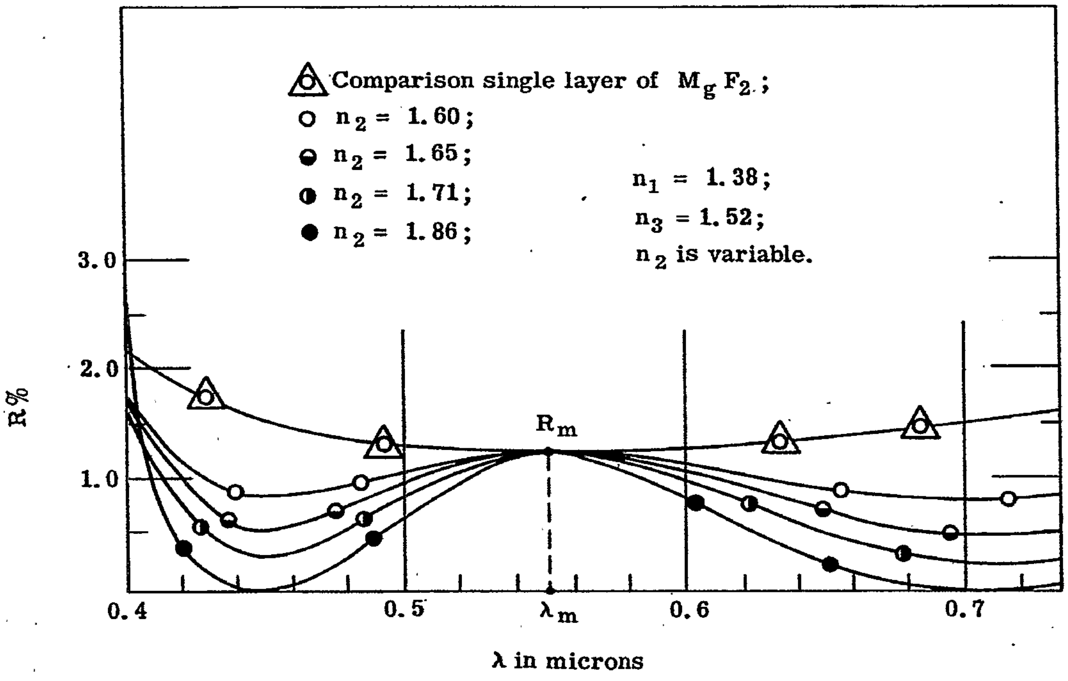


Figure 21.25- Spectral reflectance curves for a family of isoachromatic bilayers on spectacle crown glass of refractive index 1.52.

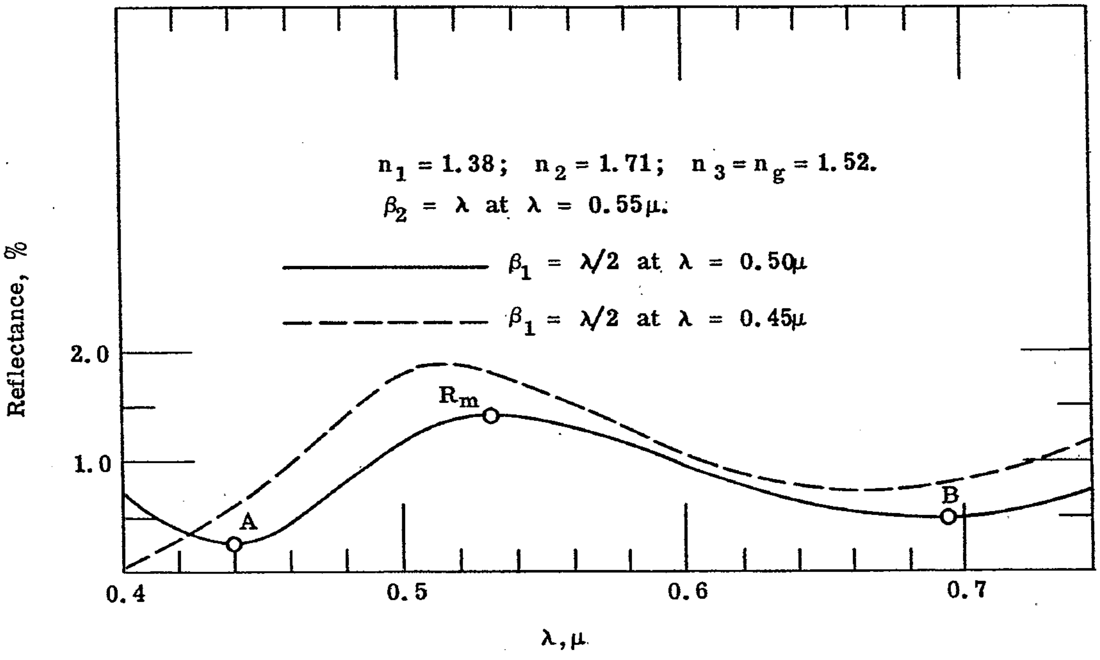


Figure 21.26- Spectral reflectance curves for two achromatic but not isoachromatic bilayers on spectacle crown glass. These curves illustrate the trend wherein the minimum reflectances at A and B become unlike when the outer layer of low refractive index and the inner layer of higher refractive index are not quarter wave and half wave, respectively, in optical path at the same wavelength.

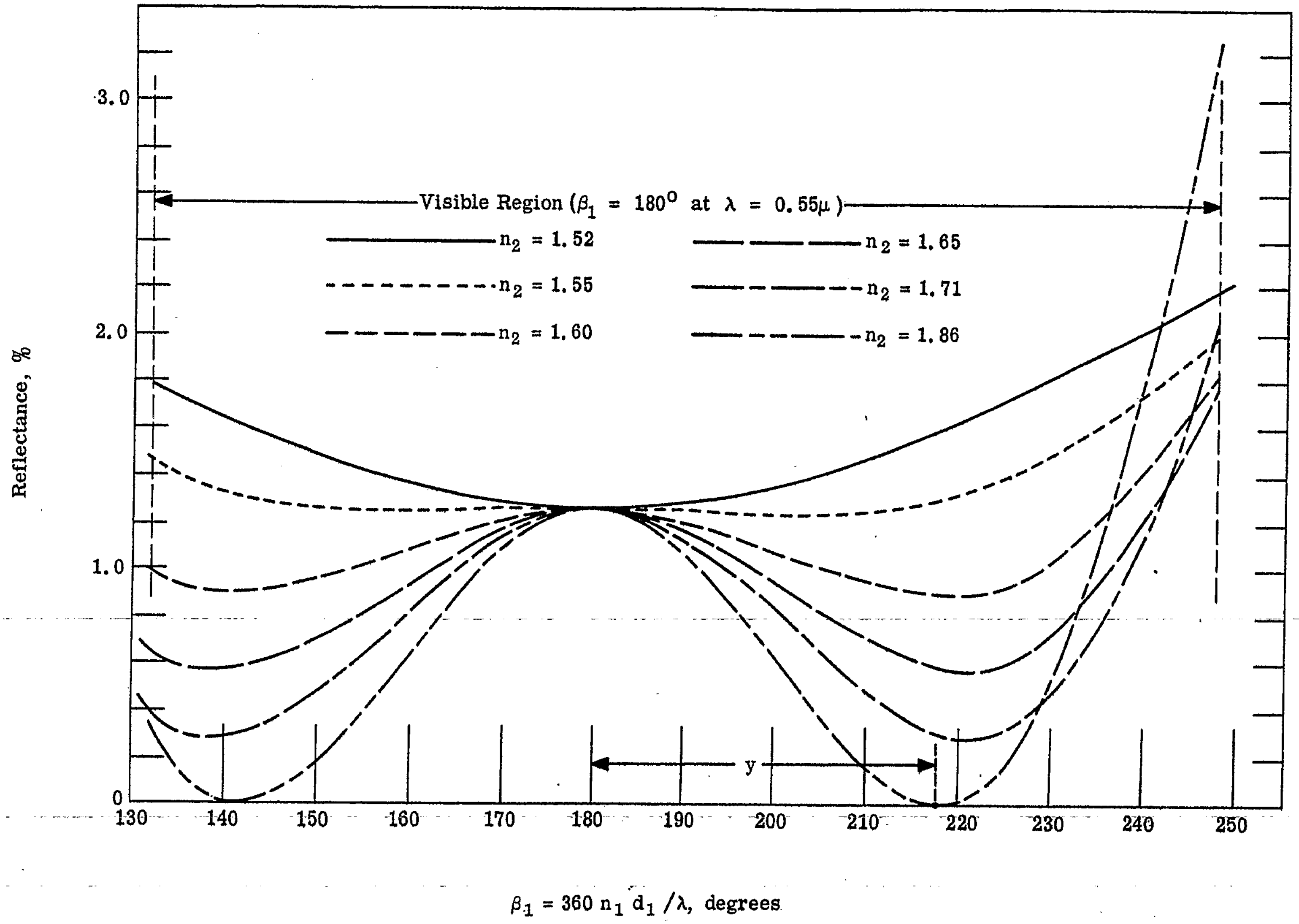


Figure 21.27- Curves illustrating the symmetry of the energy reflectances of isoachromatic bilayers about the point $\beta_1 = 180°$ where β_1 is twice the optical path of the outer layer of low refractive index $n_1 = 1.38$ (corresponding to MgF_2). The refractive index of the substrate is 1.52 as in Table 21.14.

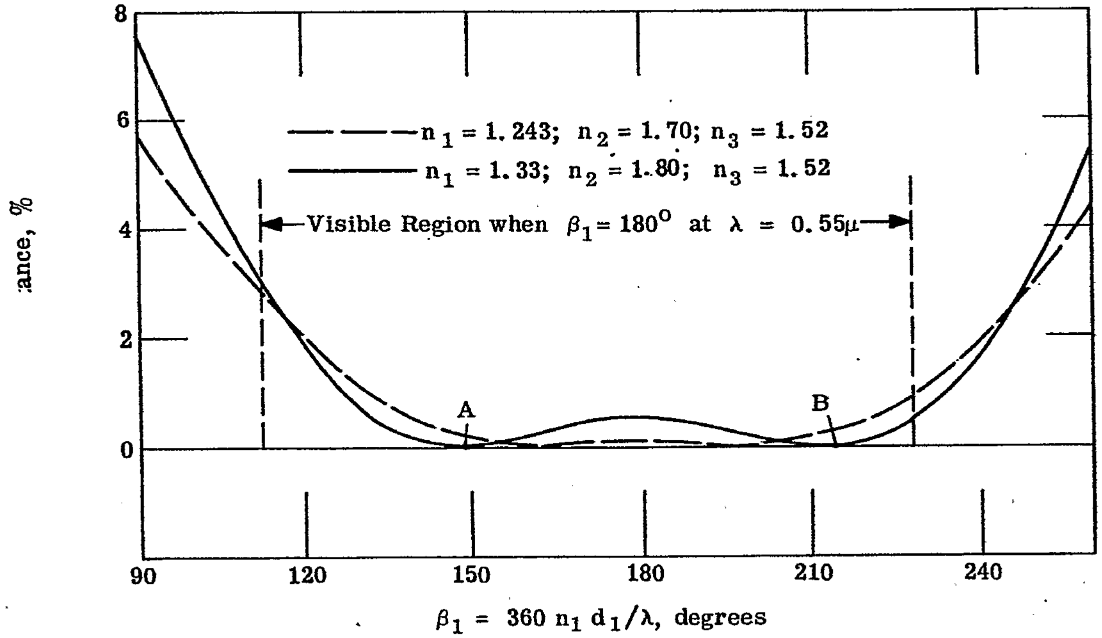


Figure 21.28 - Curves showing the effects of reducing the refractive index n_1 of the outer layer at fixed values of n_3.

in which

$$p_1 = 2\pi n_1 d_1/\lambda; \quad p_2 = 2\pi n_2 d_2/\lambda. \tag{188b}$$

We suppose at this point that the ratio n_2/n_1 does not depend upon wavelength and introduce

$$p_2 = f p_1, \tag{188c}$$

in which f can be assigned any desired value. Investigation shows that null-isoachromatic bilayers are possible mathematically for the choice $f = 1$, but require values of n_1 that are usually too small to obtain physically. The choice $f = 2$ leads to the null-isoachromatic bilayers that are of interest to this section. Equating real and imaginary parts on each side of equation (188a), after clearing fractions and introducing

$$p_2 = 2p_1 = 2p, \tag{189}$$

one obtains the zero condition in the form

$$n_1(1 - n_3) + \left[\frac{2n_1^2 n}{n_2} - 2n_2 - n_1(1 - n_3)\right] \tan^2 p = 0; \tag{189a}$$

$$\frac{2n_1 n_3}{n_2} - 2n_1 n_2 + n_3 - n_1^2 - (n_3^2 - n_1^2) \tan^2 p = 0; \tag{189b}$$

By eliminating $\tan^2 p$, one obtains the result

$$n_2 n_3 (n_1^2 + 1)(n_1 + n_2) - 2n_1 (n_2^3 + n_1 n_3^2) = 0, \tag{190}$$

an equation for determining n_2 from n_1 and n_3 or n_2 from n_1 and n_3. One obtains also

$$\cos 2p = \cos \beta_1 = n_1 \frac{n_2^2 - n_3}{n_1(n_3 - n_2^2) + n_2(n_3 - n_1^2)}. \tag{190a}$$

Introduce

$$2p = \beta_1 = 180^{\circ} \pm y \ , \qquad (190b)$$

where y is indicated in Figure 21.27. Then the solution for y becomes

$$\cos y = n_1 \frac{n_2^2 - n_3}{n_1 (n_2^2 - n_3) + n_2 (n_1^2 - n_3)} \ . \qquad (190c)$$

The required refractive indices and the optical paths, p_1 and p_2, of the two members of the bilayer thus become known algebraically.

21.7.5 Quarter wave bilayers.

21.7.5.1 Non-absorbing, quarter wave bilayers on non-absorbing substrates comprise a third, important class of bilayers. In the interests of brevity, the following discussion will be restricted to normal incidence.

21.7.5.2 Let us suppose that layer number two, Figure 21.20, is a quarter wave layer. Then $\beta_2 = \pi$ and ρ_1, equation (182a), will be real for non-absorbing systems. Consequently from equation (182),

$$\left|\rho_o\right|^2 = \frac{\rho_1^2 + W_1^2 + 2\rho_1 W_1 \cos\beta_1}{1 + \rho_1^2 W_1^2 + 2\rho_1 W_1^2 \cos\beta_1} \ . \qquad (191)$$

By differentiating $\left|\rho_o\right|^2$ with respect to β_1, one finds that the condition for maxima and minima is $\sin\beta_1 = 0$ or $\beta_1 = \nu\pi$ where ν is an integer. The choice $\nu = 1$ makes layer number one a quarter wave layer. If, therefore, both members of the bilayer are quarter wave layers at a given wavelength, λ_o, then the energy reflectance $\left|\rho_o\right|^2$ is either a minimum or a maximum for wavelengths in the immediate neighborhood of λ_o. Maxima and minima can occur at wavelengths removed from λ_o but at these wavelengths the bilayer will not be a quarter wave bilayer.

21.7.5.3 Consider now the recursion formula (81a) at wavelength λ_o for which $\beta_1 = \beta_2 = \pi$. One obtains

$$Y_{\nu-1} = \frac{M_{\nu-1}^2}{Y_\nu} = \frac{n_{\nu-1}^2}{Y_\nu} \ . \qquad (192)$$

Whence

$$Y_1 = \frac{n_1^2}{Y_2} \ ; \quad Y_2 = \frac{n_2^2}{Y_3} \ . \qquad (192a)$$

From equation (81b) we note that $Y_3 = -n_3$. Therefore the admittance Y_1 of the quarter wave bilayer is given by

$$Y_1 = -\frac{n_1^2}{n_2^2} n_3 \qquad (192b)$$

at $\lambda = \lambda_o$.

21.7.5.4 Equations (79a) and (192b) give the complex reflectance ρ_o of the non-absorbing bilayer on a non-absorbing substrate at $\lambda = \lambda_o$ and at normal incidence in the form

$$\rho_o = \frac{n_o - \left(\frac{n_1}{n_2}\right)^2 n_3}{n_o + \left(\frac{n_1}{n_2}\right)^2 n_3} \ , \qquad (193)$$

from which the energy reflectance $\left|\rho_o\right|^2$ is either a maximum or a minimum. The condition that must exist among the refractive indices to obtain zero reflectance is

$$n_o n_2^2 = n_3 n_1^2 \ . \qquad (194)$$

Available materials[13] do not ordinarily meet the zero condition of equation (194).

21.7.5.5 Let us now consider the case $n_o = 1$. In most applications, the medium of incidence is air for which n_o may be set at the approximate value unity. Then

$$R = |\rho_o|^2 = \left(\frac{1 - \left(\frac{n_1}{n_2}\right)^2 n_3}{1 + \left(\frac{n_1}{n_2}\right)^2 n_3} \right) ; \quad \lambda = \lambda_o ; \tag{195}$$

where R denotes energy reflectance. The manner in which R depends on the choice of n_1 and n_2 is illustrated in Figure 21.29 for the case $n_3 = n_g = 1.52$. Whereas only restricted generalizations can be made about combinations n_1 and n_2 that produce reflectances less than the reflectance of the uncoated substrate, we may conclude that the reflectance of the bilayer exceeds that of the uncoated substrate whenever $n_1 > n_2$ for the case $n_o = 1$. To obtain a high reflecting bilayer, one should deposit first the layer having the lower refractive index n_2.

21.7.6 Non-quarter wave bilayers.

21.7.6.1 Quarter wave bilayers will rarely satisfy the zero condition (194). However, when equation (194) is not satisfied by the refractive indices n_1 and n_2, it may be possible to choose β_1 and β_2 different from 180° in such a manner as to meet the more general zero condition. The corresponding bilayers belong to a fourth important class. The bilayers of Section 21.7.3 modified by adding $\lambda/2$ to the optical path of the inner layer of high refractive index are examples of this fourth class of bilayers. Of greatest interest are those bilayers for which β_1 and β_2 depart only slightly from 180°. The exact method of equations (185) to (185b) and equations (186) can be applied to find members of this fourth class.

[13] For a discussion of methods of chemical deposition that utilize a mixture of materials having high and low refractive indices in order to obtain films having refractive indices in the approximate range 1.44 to 2.1, see U. S. Pat. 2466119, April 15, 1949 by H. R. Moulton and E. D. Tillyer.

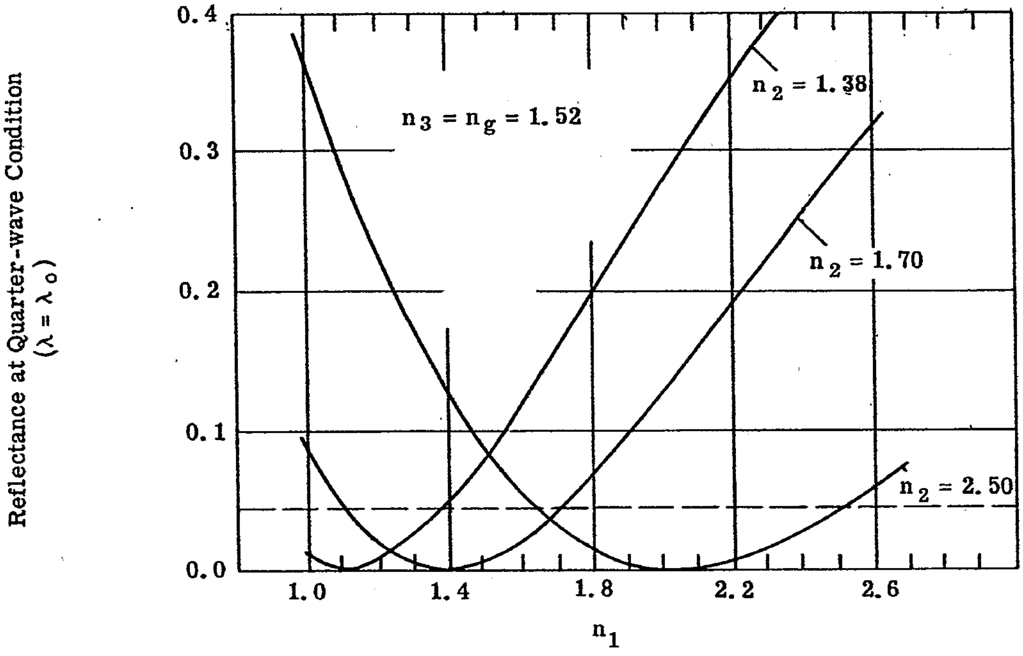


Figure 21.29- Plots of energy reflectances at $\lambda = \lambda_o$ vs the refractive index n_1 of the outer layer for the indicated values of the refractive indices n_2 of the inner layer with $n_3 = n_g = 1.52$. The broken line indicates the reflectance of the uncoated substrate, i.e. reflectance from air to the uncoated glass of refractive index $n_g = 1.52$.

21. 7. 6. 2 The early investigators of thin films utilized a graphical method, involving closed polygons, for finding multilayers having zero reflectance. The approximate method of Section 21. 2. 11 is the algebraic equivalent of the method of closed polygons when applied to the zero condition of equation (98). By inserting ρ_1 from equation (75) into equation (98), one finds that

$$W_3 \ e^{i(\beta_1 + \beta_2)} + W_2 \ e^{i\beta_1} + W_1 = 0. \qquad (196)$$

By equating the real and imaginary parts of the left hand member of equation (196) to zero, one obtains the zero condition in the form

$$\sin(\beta_1 + \beta_2) = - \frac{W_2}{W_3} \sin \beta_1 \qquad (197)$$

$$W_3 \cos(\beta_1 + \beta_2) + W_2 \cos \beta_1 + W_1 = 0. \qquad (197a)$$

Elimination of $(\beta_1 + \beta_2)$ from equation (197a) with the aid of equation (197) yields straightforwardly the result

$$\cos \beta_1 = \frac{W_3^2 - W_1^2 - W_2^2}{2 W_1 W_2} . \qquad (197b)$$

Equations (197) and (197b) determine in a simple manner first β_1 and then β_2. As the first example, consider the case in which $n_o = 1$, $n_1 = 1.38$, $n_2 = 1.72$ and $n_3 = 1.52$. With reference to the curve for $n_2 = 1.70$, Figure 21.29, n_2 is then a little too high to obtain zero reflectance when $\beta_1 = \beta_2 = 180^o$. Equations (197b) and (197) yield four pairs of solutions for β_1 and β_2. These are

$$\beta_1 = 164^o \ 16' \ ; \ \beta_2 = \begin{cases} 224^o \ 36' \\ 346^o \ 52' \end{cases} ; \qquad (198)$$

and

$$\beta_1 = 199^o \ 44' \ ; \ \beta_2 = \begin{cases} 135^o \ 24' \\ 13^o \ 08' \end{cases} . \qquad (198a)$$

The case $\beta_1 = 199^o \ 44'$ and $\beta_2 = 13^o \ 08'$ belongs to the classification of Section 21. 7. 3. In case $\beta_1 = 164^o \ 16'$ and $\beta_2 = 224^o \ 36'$, the optical paths $\beta_1/2$ and $\beta_2/2$ are most nearly equal to 90^o. On the other hand, when one examines the examples $n_o = 1$, $n_2 = 1.68$, $n_3 = 1.52$ with $n_1 = 1.38$ and 1.384, he finds that $|\cos \beta_1| > 1$. Although the physical parameters have been changed only slightly, the method of closed polygons does not admit solution.

21. 7. 7 High reflecting bilayers on metals. Bilayers of non-absorbing films can be used for gaining significant increases in reflectance from metals. For example, a very durable bilayer of silicon monoxide and titanium oxide for increasing the reflectance of an evaporated aluminum mirror has been described by G. Hass.[14]

21. 8 TRILAYERS

21. 8. 1 Introduction. The main interest in trilayers has centered on the possibilities which they provide for obtaining lower and flatter curves of spectral reflectances than is feasible with bilayers. Trilayers can exhibit three minima in spectral reflectance over the visible region. In extreme cases all three of these minima can be zero minima. The so called quarter-half-three quarter wave trilayer [15] is advantageous for achromatization. *

21. 8. 2 Low reflectance trilayers. The behavior of the quarter-half-quarter wave type of low reflecting trilayer is indicated in Figure 21.30. It should be noted that the central layer is a half-wave (and hence absentee) layer at a wavelength λ_o at which the inner and outer layers are quarter wave layers. Consequently, the trilayer behaves in effect as a quarter wave bilayer at $\lambda = \lambda_o$. The condition for zero reflectance can be found by considering equation (194) for quarter wave bilayers. Thus, $n_o^2 \ n_3^2 = n_4 \ n_1^2$ when the refractive indices

[14] Georg Hass, Vacuum, 2, p 339 (1952).

[15] For an example and discussion of this class of trilayer see O. S. Heavens, Optical Properties of Thin Solid Films, Butterworths Scientific Publications, London (1955), pp 213-215.

* The term apochromatization would be more appropriate when the film is designed to have three minima, i.e. is "corrected" at three wavelengths.

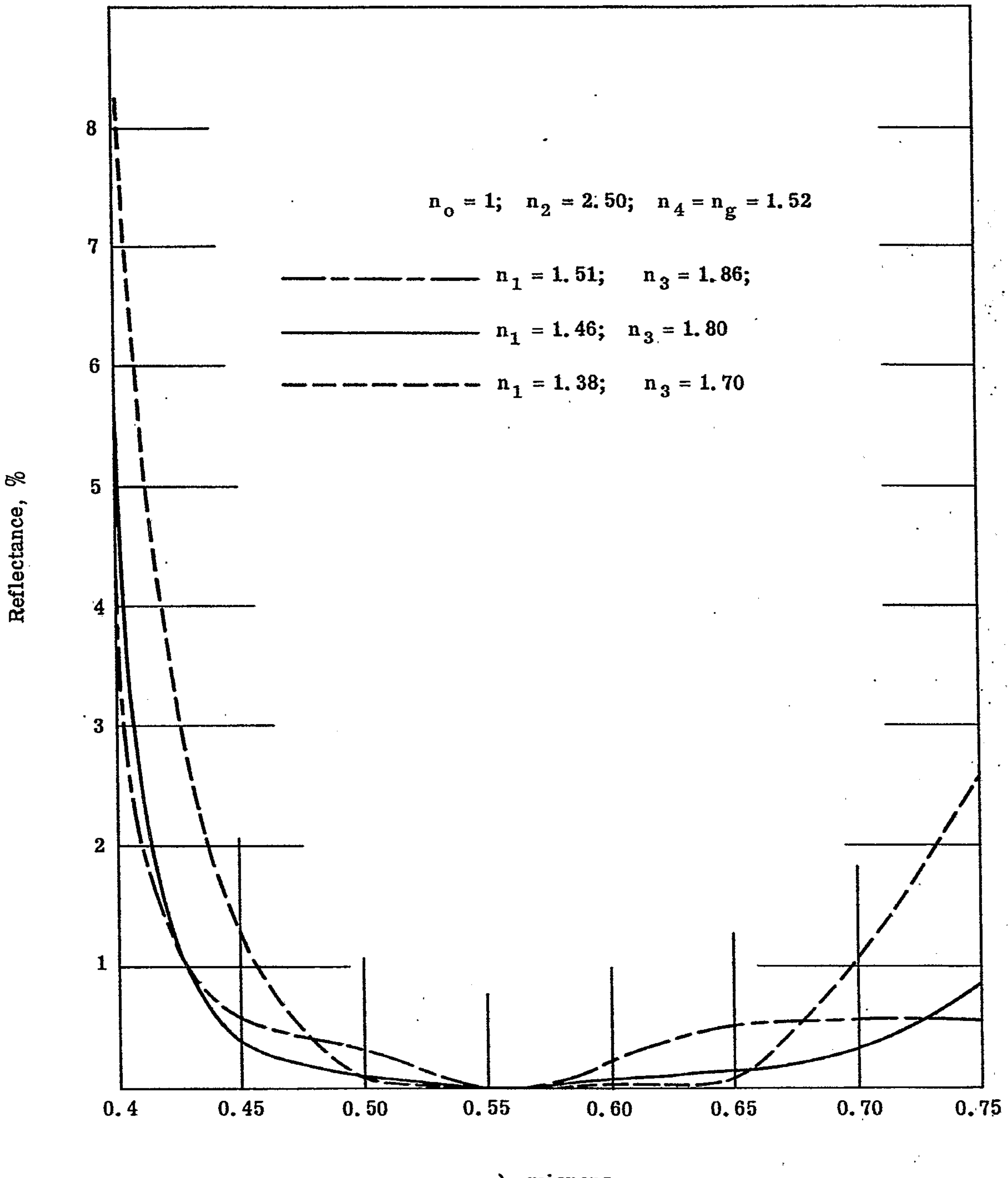


Figure 21.30- Spectral reflectances of quarter-half-quarter-wave trilayers on a substrate of refractive index 1.52. The high index $n_2 = 2.50$ belongs to the central, half-wave layer. The curves illustrate the effects of altering the refractive indices n_1 and n_3 of the quarter wave layers. The curves of this figure have been computed by the approximate method of section 21.2.11.

are properly identified with those of the trilayer. The condition for zero reflectance at $\lambda = \lambda_o$ is therefore

$$n_3 = \frac{n_1}{n_o} \sqrt{n_4} \ . \qquad (199)$$

The refractive indices n_1 and n_3 of Figure 21.30 have been chosen in accordance with equation (199). The curves of Figure 21.30 are quite flat and low from 0.45 to 0.65 microns. They tend toward achromatization but do not succeed. The performances of the quarter-half-quarter and quarter-half-three quarter wave trilayers as anti-reflection films are not far different.

21.8.3 Band pass trilayers. A trilayer consisting of a dielectric layer silvered on both major surfaces is essentially a Fabry-Perot interferometer. Like Fabry-Perot interferometers, such trilayers are readily designed to transmit narrow bands of wavelengths after the manner described and illustrated in Section 16.16. These trilayers are utilized as narrow pass band filters. By forming the dielectric layer as a wedge of gradual taper, a convenient and effective monochromator is achieved. Thin film theory involves the assumption that the number of interreflections within each film is infinite. When regarded as Fabry-Perot interferometers, the trilayers discussed here require the choice of equation 16-(107) rather than 16-(106).

21.9 QUADRILAYERS

21.9.1 A low reflecting quadrilayer. The quarter-half-quarter trilayer discussed in Section 21.8.2 has been modified in a significant manner by Dr. Helen Jupnik so as to achieve a more practical anti-reflection film that has excellent performance. Difficulties occur in making the trilayers of Figure 21.30 because the refractive indices n_1 or n_3 or both are not available as durable, non-absorbing materials. To overcome difficulty due to availability of a material having the most desired refractive index n_3, Dr. Jupnik replaces the corresponding quarter wave layer by an "equivalent" quarter wave bilayer having refractive indices n'_3 and n'_4 as illustrated in Figure 21.31.

21.9.2 The principle of equivalence. The principle of equivalence is so important to the theory of thin films that its application to Jupnik's quadrilayer will be considered in detail. The substituted bilayer shall be a

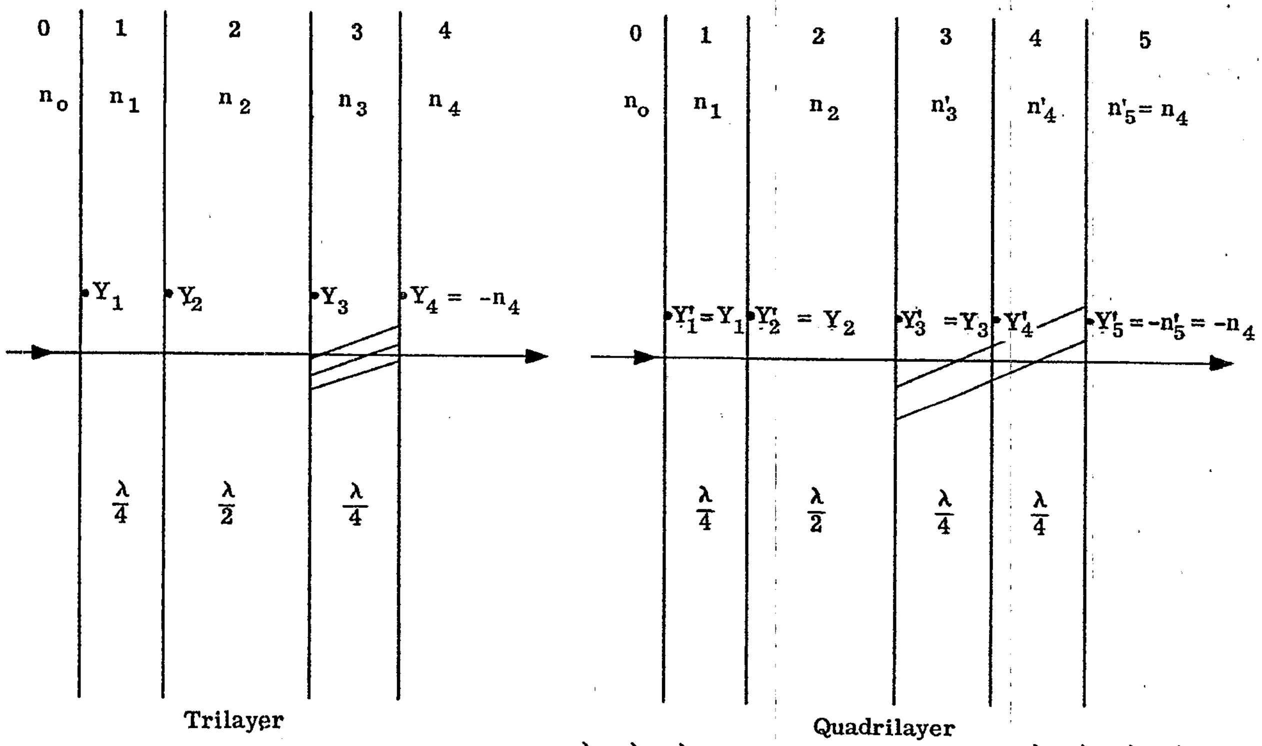


Figure 21.31- Notation with respect to the $\frac{\lambda}{4}-\frac{\lambda}{2}-\frac{\lambda}{4}$ trilayer and H. Jupnik's $\frac{\lambda}{4}-\frac{\lambda}{2}-\frac{\lambda}{4}-\frac{\lambda}{4}$ quadrilayer. The last quarter-wave layer is replaced by an equivalent bilayer. Y_ν denote admittances.

quarter wave bilayer and shall be equivalent to a quarter wave layer of refractive index n_3 at the wavelength λ_o at which the optical path is in fact one-quarter wavelength. The argument is simple on the basis of the admittances Y_ν. Suppose that it is possible to choose n'_3 and n'_4, Figure 21.31, such that $Y'_3 = Y_3$. Because layers number 1 and 2 have not been altered, it must now be true that $Y'_2 = Y_2$ and $Y'_1 = Y_1$, facts that can be checked, if desired, from equation (81a). From equation (79a) the complex reflectance ρ_o of the multilayer is $\rho_o = (n_o + Y_1)/(n_o - Y_1)$ at normal incidence. Hence ρ_o is left unaltered when $Y'_1 = Y_1$ or $Y'_3 = Y_3$. With respect to the trilayer (N = 3), equation (81b) yields

$$Y_4 = -n_4 . \tag{200}$$

Since $\beta = \pi$ for quarter wave layers, equation (81a) yields

$$Y_3 = n_3^2 / Y_4 = -n_3^2 / n_4 . \tag{200a}$$

Similarly, with respect to the last two proposed elements of the quadrilayer, Figure 21.31,

$$Y'_5 = -n'_5 = -n_4 ; \tag{200b}$$

$$Y'_4 = (n'_4)^2 / Y'_5 = -(n'_4)^2 / n_4 ; \tag{200c}$$

$$Y'_3 = (n'_3)^2 / Y'_4 = -n_4 (n'_3)^2 / (n'_4)^2 . \tag{200d}$$

By setting $Y_3 = Y'_3$ from equations (200a and d), one finds almost directly that

$$n'_3 / n'_4 = n_3 / n_4 . \tag{201}$$

The quarter wave bilayer having refractive indices n'_3 and n'_4 that satisfy equation (201) is equivalent to the quarter wave layer of refractive index n_3 for all values of n_3.

21.9.3 Selection of values.

21.9.3.1 We have seen that choosing n_3 in accordance with equation (199) makes $\rho_o = 0$ at $\lambda = \lambda_o$. By introducing n_3 from equation (199) into equation (201), we obtain Dr. Jupnik's selection for n'_3 and n'_4 in the form

$$\frac{n'_3}{n'_4} = \frac{n_1}{n_o \sqrt{n_4}} \; ; \; n_4 = n'_5 = n_g . \tag{202}$$

One is free to assign values to n'_3 or n'_4. In the example of Figure 21.32, n_1 and n'_4 are assigned the value 1.384 corresponding with the choice of $M_g F_2$. Then, with $n_o = 1$ and $n'_5 = 1.52$, one computes $n'_3 = 1.55$* from equation (202).

21.9.3.2 Comparison of Figures 21.30 and 21.32 reveals a number of interesting points. First, the substitution of the quarter wave bilayer serves also to achromatize the multilayer. Secondly, for the choice $n_1 = 1.38$ the spectral reflectance curve of Figure 21.32 is low and flat over a greater range of wavelengths. Thirdly, reflectances less than 0.1% are exhibited over a remarkably long range of wavelengths.

21.9.3.3 The effect of reducing the refractive index, n_2, of the half-wave member of the quadrilayer is illustrated in Figure 21.33. The achromatic points have moved outward to 0.45 and 0.70 microns to produce low reflectances over a greater portion of the visible spectrum than in Figure 21.32. This gain in spectral range is obtained at the cost of a slight increase of the reflectances at the points marked A and B. Further analysis shows also that the effects of increasing the refractive index, n_5, of the substrate are slight even when the refractive indices n_1 to n_4 of Figures 21.32 and 21.33 are left unchanged. From equation (202), a change in the refractive index of the substrate requires that n'_3 be changed correspondingly if one insists that the reflectance shall be zero at $\lambda = \lambda_o$.

21.10 QUARTER WAVE MULTILAYERS

21.10.1 Introduction. The following discussion is restricted to normal incidence upon non-absorbing systems and to the wavelength λ_o at which the optical path of each layer is one quarter wavelength. Equation (81a) shows that at $\lambda = \lambda_o$, where $\beta_\nu = \beta = \pi$,

$$Y_{\nu-1} = n_{\nu-1}^2 / Y_\nu . \tag{203}$$

* The refractive index of thorium oxyfluoride falls near 1.55.

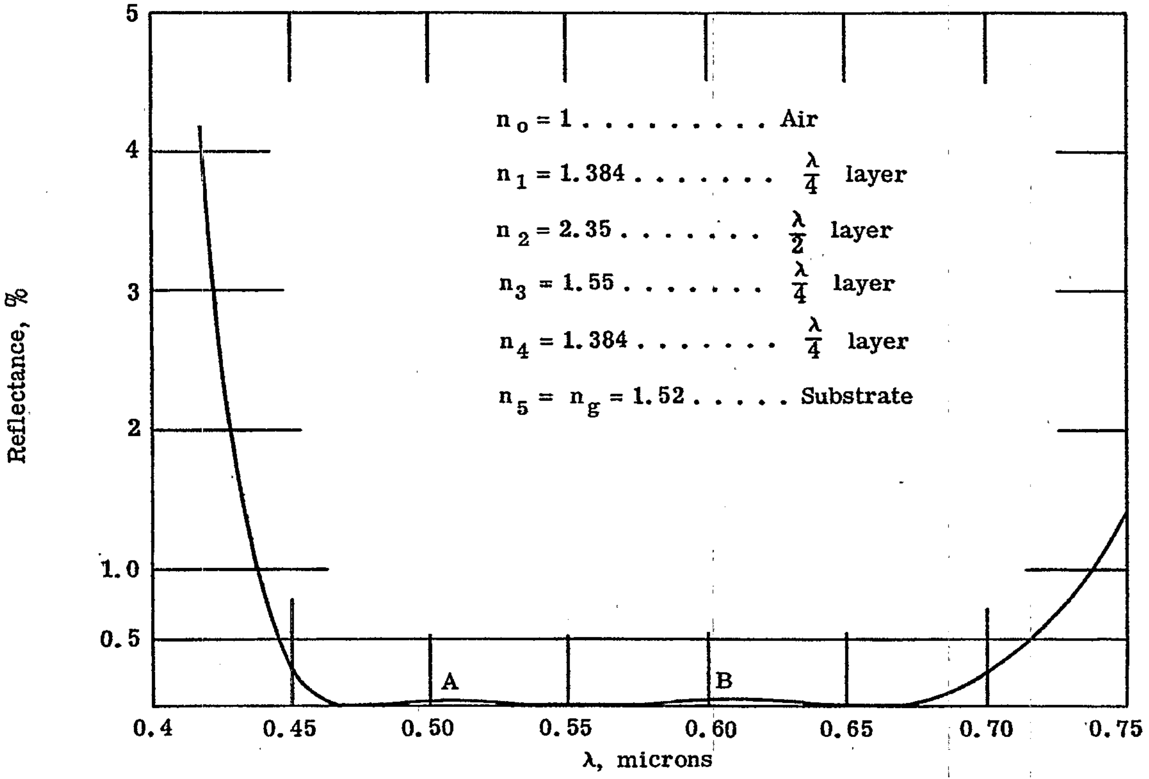


Figure 21.32- Curve of the computed spectral reflectances of a quadrilayer by Dr. H. Jupnik for the indicated refractive indices of the system.

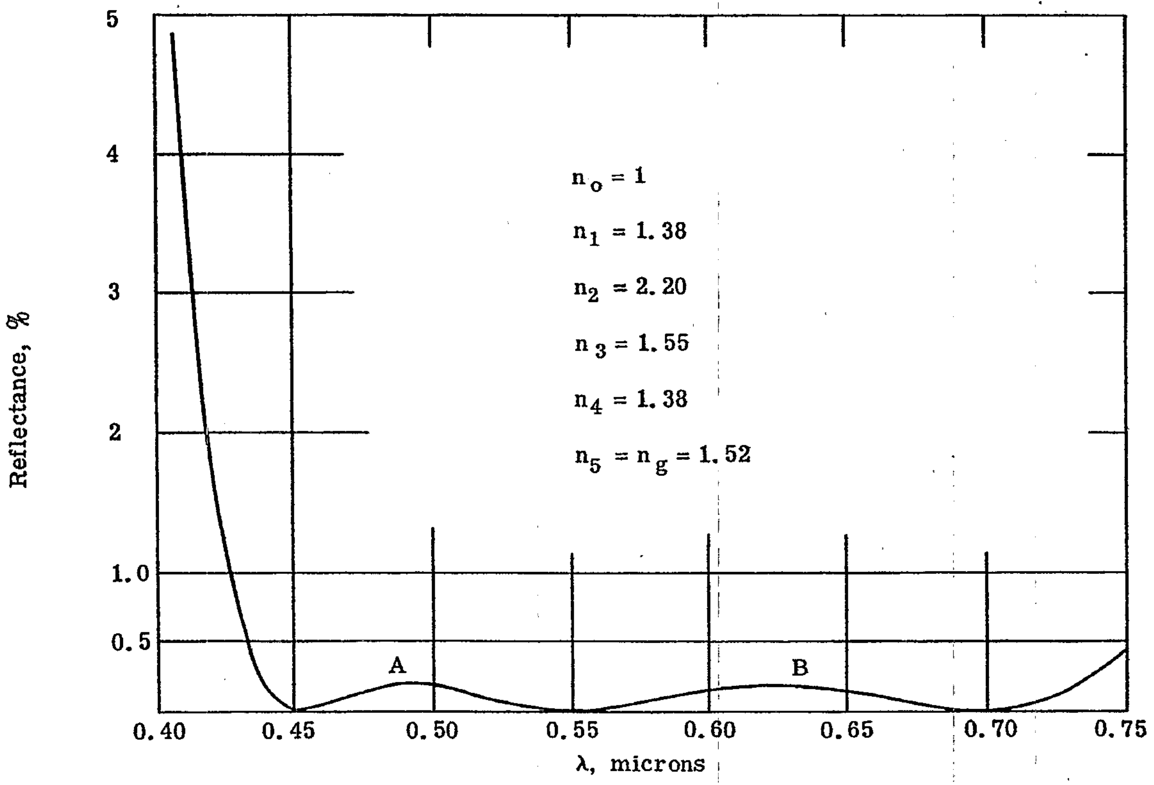


Figure 21.33- Curve of spectral reflectances illustrating the effect of decreasing the refractive index n_2 of the half-wave layer of the quadrilayer of Figure 21.32. The curves of both Figures 21.32 and 21.33 have been computed by an accurate method.

Closer examination of equation (81a) reveals that equation (203) holds whenever $\beta_\nu/2 = \beta/2 = \mu\,\pi/2$ where μ is an odd integer and $\beta/2$ is optical path. Increasing the optical path of any one or any number of the layers by an integral number of half-wavelengths will not alter the following conclusions.

21.10.2 The admittance, Y_1. From equation (203) one finds directly that $Y_2 = n_1^2/Y_1$; $Y_3 = n_2^2/Y_2 = Y_1\,n_2^2/n_1^2$; $Y_4 = n_3^2/Y_3 = n_3^2\,n_1^2/n_2^2\,Y_1$; etc. Hence one concludes by induction that

$$Y_N = \frac{1}{Y_1}\;\frac{n_{N-1}^2 n_{N-3}^2 n_{N-5}^2 \cdot\,\cdots\cdots\, n_1^2}{n_{N-2}^2 n_{N-4}^2 n_{N-6}^2 \cdot\,\cdots\cdots\, n_2^2}\quad;\ N\ \text{even}\,; \tag{204}$$

$$Y_N = Y_1\;\frac{n_{N-1}^2 n_{N-3}^2 \cdot\,\cdots\cdots\, n_2^2}{n_{N-2}^2 n_{N-4}^2 \cdot\,\cdots\cdots\, n_1^2}\quad;\ N\ \text{odd}\,; \tag{204a}$$

where N is the number of layers in the multilayer. However,

$$Y_N = n_N^2/Y_{N+1} = -\,n_N^2/n_{N+1} \tag{204b}$$

since $Y_{N+1} = -\,n_{N+1}$, see equation (81b). Hence, the admittance Y_1 of any quarter-wave multilayer is given by

$$Y_1 = -\,n_{N+1}\;\frac{n_{N-1}^2 n_{N-3}^2 \cdot\,\cdots\cdots\, n_1^2}{n_N^2\; n_{N-2}^2 \cdot\,\cdots\cdots\, n_2^2}\quad;\ N\ \text{even}\,; \tag{205}$$

$$Y_1 = -\,\frac{1}{n_{N+1}}\;\frac{n_N^2\; n_{N-2}^2 \cdot\,\cdots\cdots\, n_1^2}{n_{N-1}^2 n_{N-3}^2 \cdot\,\cdots\cdots\, n_2^2}\quad;\ N\ \text{odd}\,. \tag{205a}$$

With Y_1 thus determined, the corresponding complex reflectance ρ_o of the multilayer can be computed from equation (79a). For normal incidence,

$$\rho_o = \frac{n_o + Y_1}{n_o - Y_1}\;. \tag{206}$$

21.10.3 The zero condition. According to equation (206), the reflectance will be zero at $\lambda = \lambda_o$, whenever $Y_1 = -\,n_o$, i.e. whenever

$$\frac{n_o}{n_{N+1}} = \frac{n_{N-1}^2 n_{N-3}^2 \cdot\,\cdots\cdots\, n_1^2}{n_N^2\; n_{N-2}^2 \cdot\,\cdots\cdots\, n_2^2}\quad;\ N\ \text{even}\,; \tag{207}$$

$$n_o\, n_{N+1} = \frac{n_N^2\; n_{N-2}^2 \cdot\,\cdots\cdots\, n_1^2}{n_{N-1}^2 n_{N-3}^2 \cdot\,\cdots\cdots\, n_2^2}\quad;\ N\ \text{odd}\,. \tag{207a}$$

When $N = 2$, one obtains $n_o/n_3 = n_1^2/n_2^2$, the result of equation (194) for quarter wave bilayers. When $N = 3$, one obtains $n_o n_4 = n_1^2\, n_3^2/n_2^2$, the zero condition for trilayers. Equations (207 and 207a) give one much greater flexibility as regards the choice of materials for obtaining zero reflectance than does the highly specialized zero condition for a monolayer or for a bilayer.

21.10.4 High reflecting multilayers. Equation (206) shows that there are two different ways in which one can achieve the result $|\rho_o| \to 1$. Thus the energy reflectance approaches its highest value unity as

$$Y_1 \to 0\,; \tag{208}$$

$$|Y_1| \to \infty\,. \tag{208a}$$

Suppose with respect to equation (205) that $n_{N-1}/n_N < 1$, $n_{N-3}/n_{N-2} < 1$, ... $n_1/n_2 < 1$. Then $|\rho_o| \to 1$ as $N \to \infty$ on account of equation (208). On the other hand, if one chooses $n_{N-1}/n_N > 1$, ... $n_1/n_2 > 1$, then $Y_1 \to \infty$ and $|\rho_o| \to 1$ as the number N of the layers in the multilayer approaches infinity. Similar observations apply when N is odd as in equation (205a). Therefore many possibilities are open for achieving high energy reflectance by increasing the number of layers in the multilayer.

21.10.5 The periodic system of repeated bilayers. The production of a multilayer is simplified by alternating layers of high and low refractive indices n_h and n_l. When the number N of layers is even, the resulting system of layers is periodic and forms an assembly of repeated bilayers as illustrated in Figure 21.34. There

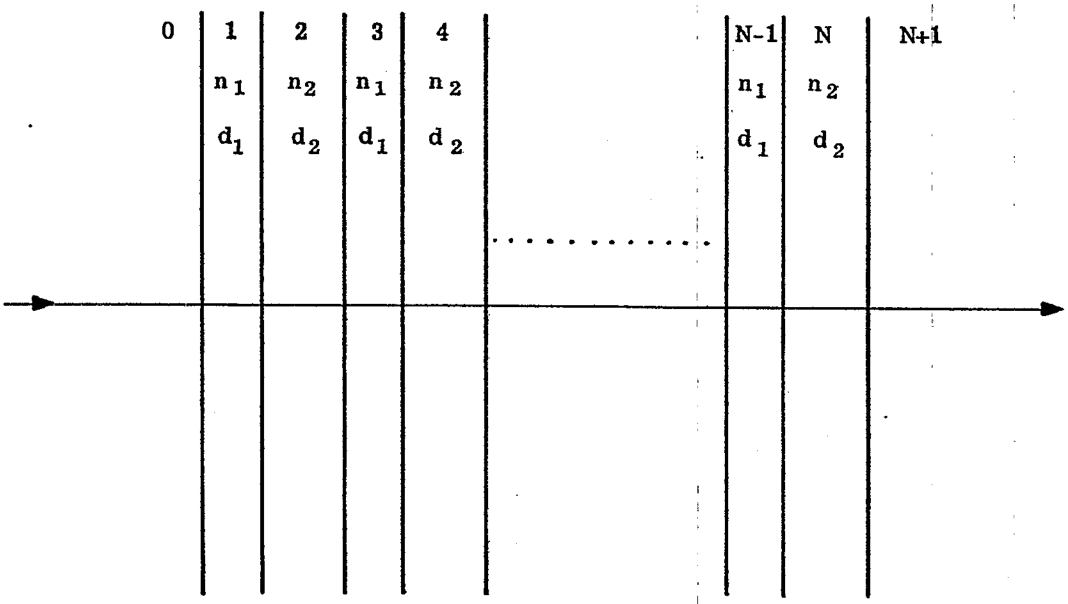


Figure 21.34- A multilayer consisting of repeated bilayers having the refractive indices n_1 and n_2. The thicknesses d_1 and d_2 are likewise repeated and are chosen so that the optical path is one quarter wavelength at an assigned wavelength λ_o. The number N of layers must now be even.

are N/2 even and N/2 odd integers between 0 and N when N is an even integer. Consequently, equation (205) simplifies to the result

$$Y_1 = - n_{N+1}\left(\frac{n_1}{n_2}\right)^N ; \tag{209}$$

and equation (206) assumes the explicit form

$$\rho_o = \frac{n_o - n_{N+1}\left(\frac{n_1}{n_2}\right)^N}{n_o + n_{N+1}\left(\frac{n_1}{n_2}\right)^N} = \frac{\frac{n_o}{n_{N+1}} - \left(\frac{n_1}{n_2}\right)^N}{\frac{n_o}{n_{N+1}} + \left(\frac{n_1}{n_2}\right)^N} . \tag{210}$$

The zero condition requires that

$$\frac{n_o}{n_{N+1}} = \left(\frac{n_1}{n_2}\right)^N . \tag{211}$$

If $n_o < n_{N+1}$, one must choose $n_1 < n_2$ in order to obtain zero reflectance, i.e. one must apply the layer of higher refractive index upon the substrate of refractive index n_{N+1}. On the other hand, equation (210) shows that $|\rho_o|$ can be made high whether one chooses $n_1 < n_2$ or $n_1 > n_2$ provided that N is taken sufficiently large; but that one should choose the alternative $n_1 > n_2$ in order to obtain the highest energy reflectance $|\rho_o|^2$ for a given number N of layers.

21.10.6 Odd number of alternating layers. - An important group of quarter wave multilayers containing an odd number N of layers having alternating refractive indices n_1 and n_2 is illustrated by Figure 21.35. If, for example, N = 5 and $n_1 > n_2$, these facts are indicated by the notation HLHLH or $(HL)^2$ H in which H and L refer to the high and low refractive indices n_1 and n_2, respectively. As a second example, if N = 15 and $n_1 < n_2$, the multilayer is described by writing $(LH)^7$ L. As in section 24.9.4, the complex

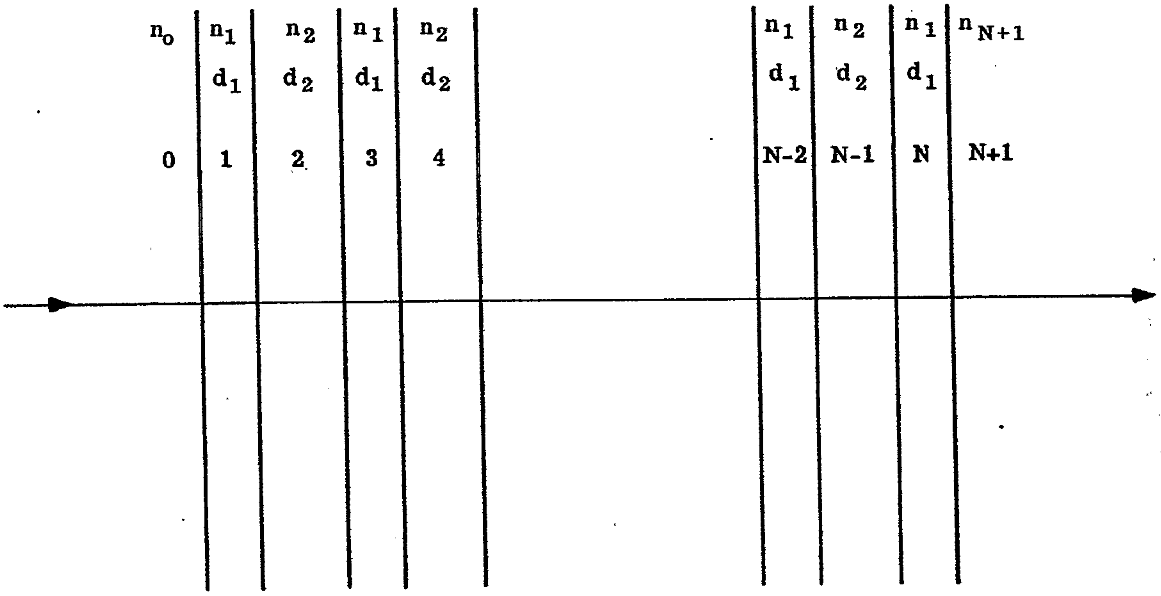


Figure 21.35- A multilayer consisting of an odd number N of quarter wave layers whose refractive indices n_1 and n_2 are alternated. A layer of refractive index n_1 occurs at each end of the multilayer. The optical paths of the layers are $\lambda/4$ at $\lambda = \lambda_o$.

reflectance ρ_o at $\lambda = \lambda_o$ can be found in a simple manner from equations (205a) and (206). The result is

$$\rho_o = \frac{n_o - \dfrac{n_1^2}{n_{N+1}} \left(\dfrac{n_1}{n_2}\right)^{N-1}}{n_o + \dfrac{n_1^2}{n_{N+1}} \left(\dfrac{n_1}{n_2}\right)^{N-1}} , \tag{212}$$

with the energy reflectance $R = |\rho_o|^2$. To obtain the highest value of R with the fewest number of layers, one chooses n_1 as large as possible and n_2 as low as possible. The curve of computed spectral transmittance $T = 1 - |\rho_o(\lambda)|^2$ of the multilayer $(HL)^5 H$ is shown in Figure 21.36 for the indicated values of n_o, n_1, n_2 and $n_{N+1} = n_g$. The layers'are quarter wave layers at $\lambda = \lambda_o = 0.75$ microns at which one computes from equation (212) that $|\rho_o|^2 = 0.9946$ whence $T = 0.0054$. Energy reflectances exceeding those from silver are obtained easily with multilayers.

21.10.7 Achromatization. The spectral transmittance curve of Figure 21.36 illustrates one of the more serious difficulties encountered in the design of thin films. The oscillations in the spectral transmittances exemplified by those occurring between 0.4 and 0.62 microns are usually undesirable. The term achromatization or achromatizing pertains to the minimization of the amplitudes of undesired, rapid oscillations of spectral transmittance or reflectance curves in such a manner that the curves are flattened. This usage of the term achromatization is not entirely consistent with that of Section 21.7.4. Achromatization of bilayers and multilayers differ slightly, we may say, as to the manner in which a curve of spectral reflectance or transmittance is "flattened". An example * of one method of achromatizing multilayers is illustrated in considerable detail by Figures 21.35, 21.36, 21.37, and 21.38. Flattening of the curve of spectral transmittances between 0.4 and 0.62 microns is accomplished without appreciable alteration of the curve between 0.62 and 0.75 microns. As with bilayers, achromatization of multilayers can be achieved also by adding half-wave layers at strategic locations within the multilayer.

* It will not be possible due to lack of time and space to do justice to the able work of many investigators who have contributed to methods of achromatization.

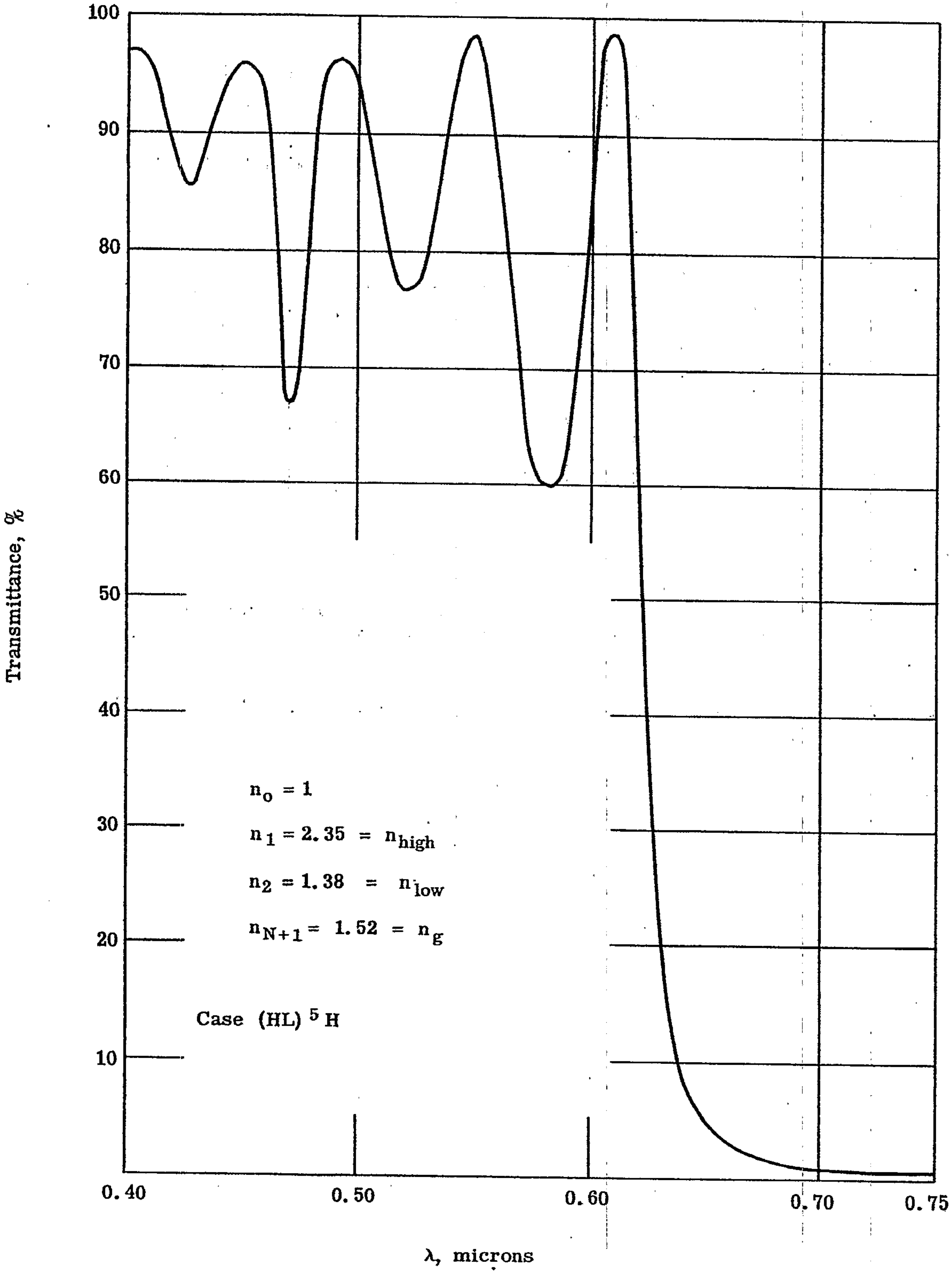


Figure 21.36- Curve of computed spectral transmittances taken from the files of Dr. H. Jupnik for the alternating multilayer $(HL)^5 H$ having layers that are quarter-wave layers at $\lambda_o = 0.75$ microns. The heights and locations of the numerous maxima and minima have not been determined with greatest possible care.

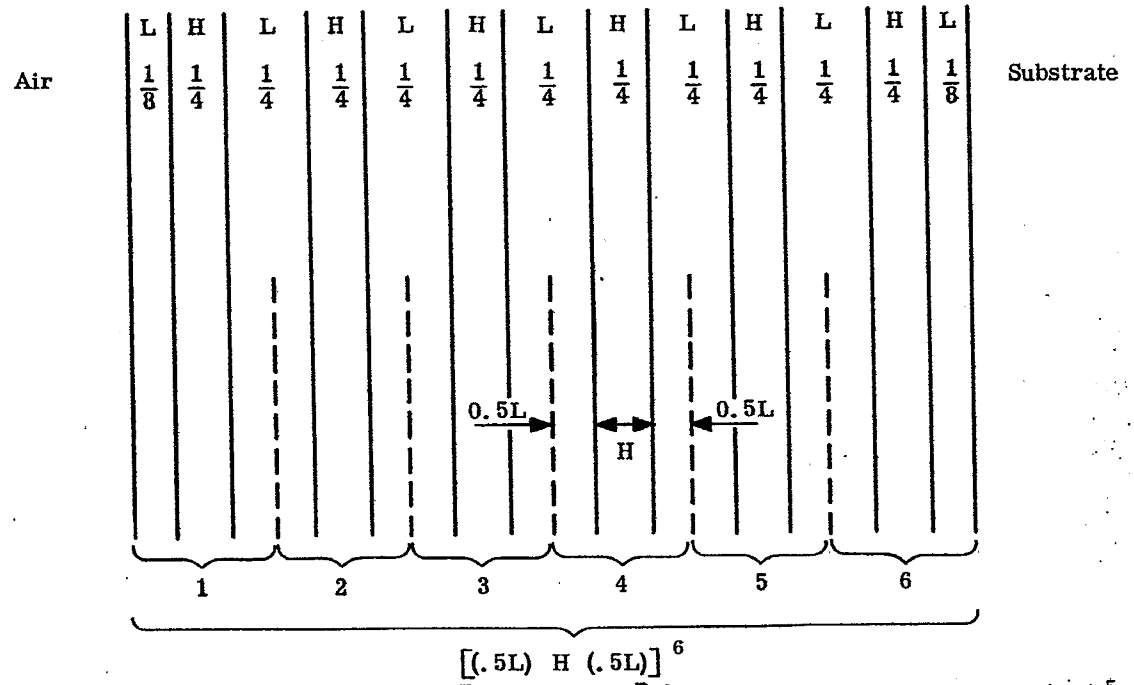


Figure 21.37- Explanation of the notation $[(.5L)\ H\ (.5L)]^6$ with respect to a multilayer $(HL)^5H$ to which has been added at each end an achromatizing layer of low refractive index and of optical path $\lambda/8$ at $\lambda = \lambda_o$. The system consists of six repeated "trilayers," (.5L) H (.5L) or $(\frac{L}{2})\ H\ (\frac{L}{2})$.

21.10.8 Narrow pass band filters.

21.10.8.1 The design of multilayers that are intended for use as narrow pass band filters is based upon the principles of the Fabry-Perot interferometer. The silver coatings on opposite surfaces of the plane parallel plate are replaced by high reflecting multilayers. In turn, the plane parallel plate is usually replaced by a layer that serves as the spacer. High reflecting multilayers are superior to silver coatings * when only small amounts of absorption can be tolerated and when durability becomes an important consideration. One arrangement is illustrated in Figure 21.39. When all the layers of multilayers, B_1 and B_2, are quarter wave layers at $\lambda = \lambda_o$, the optical path of the separating layer, S, should be an integral number, ν, of half waves at λ_o. It is not difficult to see that at normal incidence the transmittance is unity at $\lambda = \lambda_o$ for the idealized system that contains no absorption, scattering or departures from the rigid design of Figure 21.39. First, one notes that the spacer layer S is an absentee layer at λ_o. Layers 1 and 1' are then effectively in contact and comprise a half-wave or absentee layer. This now places layers 2 and 2' effectively in contact to form a third absentee layer. One concludes that all opposing pairs of layers form absentee layers at λ_o -- and, indeed, that the filter becomes an "absentee filter" that must have transmittance unity.

21.10.8.2 The behavior of the filter at $\lambda \neq \lambda_o$ can be appreciated and evaluated as follows from the theory of Fabry-Perot interferometers. With reference to equation 16-(103) we note that the parameter, A, is now given by

$$A = |\rho_o| \quad |\rho'_o| \tag{213}$$

in which ρ_o and ρ'_o are the complex reflectances of the "coated" surfaces of the spacer, S, Figure 21.39, since the spacer has been assumed to be non-absorbing. Let T denote the time averaged energy transmittance of the filter. Then T = W where W is the quantity given by equation 16-(107). From this equation we see that

$$2T = \frac{|\tau \ \tau'|^2}{1 - 2A\cos\alpha + A^2} \tag{214}$$

* Silver coatings having extremely small amounts of absorption can be produced by evaporation; but the required technique is not well known and the silver coatings are not likely to remain low-absorbing.

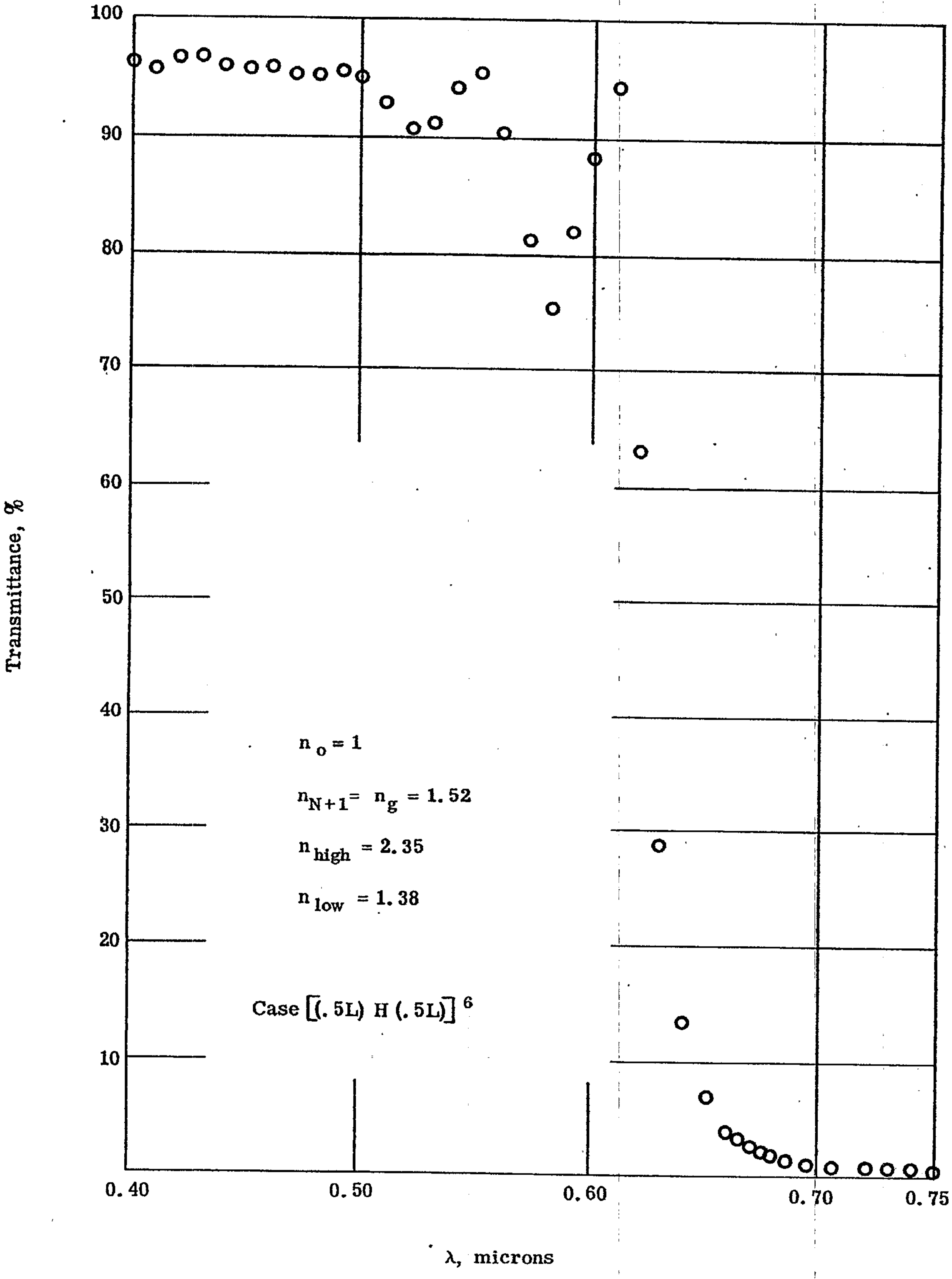


Figure 21.38- Plot of the spectral transmittances obtained by achromatizing the system $(HL)^5H$ of Figure 21.35. by the addition of a λ/8 layer of low refractive index at each end of the multilayer. The notation for the multilayer thus obtained is $[(5L)\ H\ (.5L)]^6$. These plotted data have been taken from the files of Dr. H. Jupnik.

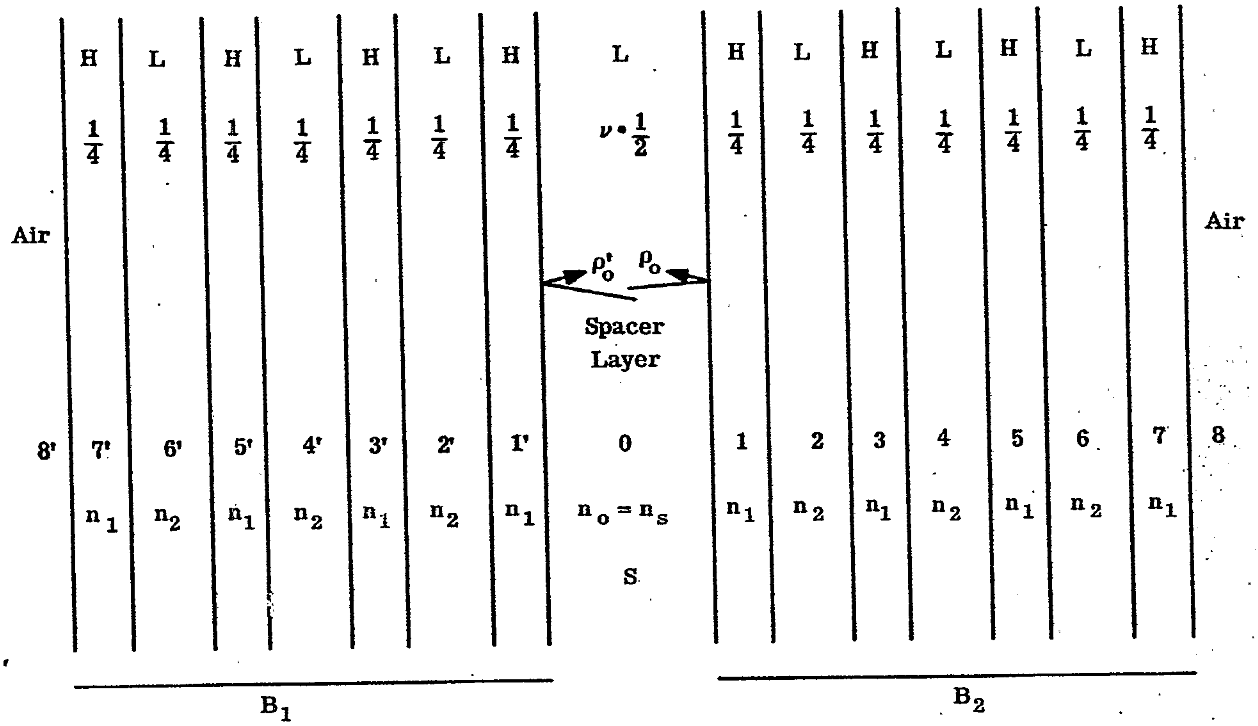

H L H L H L H L H L H L H L H
1/4 1/4 1/4 1/4 1/4 1/4 1/4 ν·1/2 1/4 1/4 1/4 1/4 1/4 1/4 1/4
Air
Air
ρ'o ρo
Spacer Layer
8' 7' 6' 5' 4' 3' 2' 1' 0 1 2 3 4 5 6 7 8
n1 n2 n1 n2 n1 n2 n1 no = ns n1 n2 n1 n2 n1 n2 n1
S
B1
B2

Figure 21.39- A narrow pass band, interference filter consisting of a spacer layer coated with high reflecting, quarter-wave multilayers B_1 and B_2 that consist of seven alternating layers.

in which τ' and τ are the complex transmittances of multilayers B_1 and B_2, Figure 21.39, and

$$\alpha = \frac{4\pi n_s d_s}{\lambda} + \arg(\rho_o) + \arg(\rho'_o) \qquad (215)$$

where n_s and d_s are the refractive index and thickness, respectively, of the spacer. The transmittance T is maximum when

$$\alpha = \nu\, 2\pi \qquad (216)$$

where the integer ν is called the order number of the filter or interferometer. The corresponding wavelength is usually indicated by λ_ν in Fabry-Perot interferometry. By eliminating α from equation (215) with the aid of equation (216), one obtains the result

$$2n_s\, d_s + \frac{\lambda\nu}{2\pi}\left[\arg(\rho_o) + \arg(\rho'_o)\right] = \nu\lambda_\nu\,. \qquad (217)$$

Sharp peaks of transmittance $T = W$ are produced by equation (214), as illustrated by Figure 16.19, when the equivalent equations (216) and (217) are satisfied. These sharp peaks are the narrow pass bands.

21.10.8.3 Suppose now that the numerous conditions of Figure 21.39 are met by the interference filter at $\lambda = \lambda_o$. Then $\rho_o = \rho'_o$ and ρ_o is determined from equation (212) with $N = 7$. Because we have chosen $n_1 > n_2$, ρ_o turns out to be real and negative. We may write

$$\rho_o = \left|\rho_o\right| e^{\pm i\pi}\,. \qquad (218)$$

The phase change on reflection can be taken as either of the two physically indistinguishable values, $\pm\pi$. The simplest way of interpreting equation (217) at $\lambda = \lambda_o$ is now to take $\rho_o = \left|\rho_o\right| e^{i\pi}$ and $\rho'_o = \left|\rho_o\right| e^{-i\pi}$ since these are physically indistinguishable. Correspondingly,

$$\arg(\rho_o) + \arg(\rho'_o) = 0\,;\quad \lambda = \lambda_o\,. \qquad (219)$$

Then, simply,

$$2n_s\, d_s = \nu \lambda_\nu \quad \text{at} \quad \lambda = \lambda_o . \tag{220}$$

Hence we may regard the integers ν of Figure 21.39 and equation (220) as the same integer, i.e. as the spectral order number of the filter. Equation (220) explains why an interference filter can exhibit two pass bands in the visible region. Suppose that λ_o is chosen at the red end of the spectrum. Here, $\lambda_o = \lambda_\nu$. If ν is high enough, it can happen that at a shorter wavelength $\lambda_{\nu+1}$ the order number is increased to $\nu + 1$. In other words, it can happen that

$$2n_s\, d_s = (\nu + 1)\, \lambda_{\nu+1} , \tag{221}$$

where $\lambda_{\nu+1} < \lambda_o$. Because equation (217) and its simplifications are conditions for maxima of T, the wavelength $\lambda_{\nu+1}$ defines the center of a subsidiary pass band. By increasing ν (increasing d_s), one obtains additional pass bands in a specified spectral range such as the visible region.

21.10.8.4 The widths of the narrow pass bands are important properties of the interference filter. These widths can be evaluated as follows. From equation 16-(113),

$$|\Delta \alpha| = \frac{1 - A}{\sqrt{A}} , \tag{222}$$

where $\Delta\alpha$ is a particular departure of α from its value α_ν at the wavelength λ_ν at which T is maximum. This departure of α from α_ν changes T from $T = T_{\text{maximum}}$ to $T_{\text{maximum}}/2$. Let us assume for simplicity that the variation of the phase changes on reflection arg (ρ_o) and arg (ρ'_o) with wavelength can be ignored.* By differentiating α with respect to λ in equation (215), one obtains

$$|\Delta \alpha| = \frac{4\pi n_s d_s}{\lambda^2} |\Delta \lambda| . \tag{223}$$

We evaluate the derivative at the center of the pass band under consideration where $\lambda = \lambda_\nu$. Let $2n_s d_s$ be eliminated from equation (223) with the aid of equation (220). Then

$$|\Delta \alpha| = \nu 2\pi \frac{|\Delta\lambda|}{\lambda_\nu} . \tag{224}$$

By eliminating $\Delta\alpha$ from equation (224) with the aid of equation (222), we obtain the formulae

$$2\,|\Delta\lambda| = \frac{\lambda_\nu}{\nu} \frac{1 - A}{\pi\sqrt{A}} ; \tag{225}$$

$$= \frac{\lambda_\nu}{\nu} \cdot \frac{1 - |\rho_o|\;|\rho_o|}{\pi\sqrt{|\rho_o|\;|\rho'_o|}} ; \tag{225a}$$

$$= \frac{\lambda_\nu}{\nu f} \tag{225b}$$

where the finesse [16] f is defined as

$$f = \pi\sqrt{A}\,/(1 - A) . \tag{226}$$

$|\Delta\lambda|$ is the half-width of the pass band of the filter at the selected order number ν. By definition, $|\Delta\lambda|$ is the departure of λ from λ_ν that causes the transmittance of the filter to drop from $T_{\text{max.}}$ at λ_ν to $T_{\text{max.}}/2$ at $\lambda = \lambda_\nu \pm |\Delta\lambda|$. When $|\rho_o| = |\rho'_o|$,

$$2\,|\Delta\lambda| = \frac{\lambda_\nu}{\nu} \quad \frac{1 - |\rho_o|^2}{\pi\,|\rho_o|} , \tag{227}$$

where $|\rho_o|$ is evaluated at the wavelength λ_ν. At the wavelength $\lambda = \lambda_o$, ρ_o can be calculated from equation (212) provided that the high reflecting multilayer falls in the class governed by equation (212). Equation (227) shows that the half-width is decreased by increasing $|\rho_o|$ or the order number ν. At fixed λ_ν, the order number is increased by increasing the optical path $n_s d_s$ of the spacer.

*In some applications, arg (ρ_o) is deliberately made a rapid function of wavelength in order to utilize the dispersion of the phase changes on reflection for obtaining narrow pass bands.

[16] P. Giacomo, Rev. D'optique, 35, 317 (1956).

21.10.8.5 As an example of the evaluation of a half-width from the theory, let us now consider the type of interference filter of Figure 21.39 with $n_1 = 2.35$, $n_2 = 1.38$ and $n_s = 1.38$. This choice corresponds to the use of ZnS and $M_g F_2$ as the material of high and low refractive index, respectively. We suppose that the multilayers B_1 and B_2 are quarter-wave systems at $\lambda_o = 5550$ Angstroms and that the spacer is a half-wave spacer at λ_o. We compute the half-width at the main transmittance peak located at $\lambda_\nu = \lambda_o$ where $\nu = 1$. In applying equation (212), we may take $n_o = n_2 = 1.38$, $n_{N+1} = 1$, $n_1 = 2.35$ and $N = 7$. Then $|\rho_o| = 0.9797$ and $|\rho_o|^2 = 0.9598$. Then from equation (227), $|\Delta\lambda| = 36.2$ Angstroms. The half-width may be decreased to 12.07 Angstroms by choosing $\nu = 3$. In practice, further reduction of $|\Delta\lambda|$ is accomplished by increasing $|\rho_o|$, i.e. by increasing N or by choosing a pair of dielectric materials for which the ratio n_1/n_2 is higher. We see that half-widths of one Angstrom or less become difficult to attain.

21.11 MATERIALS AND TEXTS

Whereas many scattered publications deal with the optical properties of substances that are suitable for use as thin films, the writer is unaware of a publication that contains an exhaustive summary of the optical properties of the many possible materials from the ultraviolet region into the region of the infrared.

One of the longest tables of the optical and mechanical properties of materials that are used in making thin films will be found in L. Holland, "Vacuum Deposition of Thin Films," John Wiley & Sons, Inc. (1956).

Quite detailed discussion of the optical constants and properties of metallic films is included in O. S. Heavens, "The Optical Properties of Thin Films," Butterworths Scientific Publications, London (1955). This book includes scattered information about the optical constants of other materials such as ZnS, Sb_2O_3, Ge and Te.

A useful list of the optical constants of metals and inorganic compounds appears in most editions of "Handbook of Chemistry and Physics," Chemical Rubber Publishing Co.

A book by W. Lewis, "Thin Films and Surfaces," Temple Press Ltd, London, First Edition is devoted to the structure, properties and production of various thin films. Emphasis is placed upon aluminum and alloys containing aluminum.

The excellent work of Dr. Georg Hass and his associates has contributed information about the optical properties and formation of thin films -- especially the oxides of titanium, silicon, aluminum and rare earths. As one example, see Georg Hass, Vacuum, 2, 331-345 (1952). This publication contains a substantial list of references.

The following texts may be consulted for much additional, valuable discussion relative to thin films.

Auwarter, Max, ed. -- "Ergebnisse der Hochvakuum technik und der Physik dunner Schichten," Stuttgart, Wissen schaftliche Verlagsgessellschaft (1957).

Mayer, Herbert -- "Physik dünner Schichten," Stuttgart, Wissenschaftliche Verlagsgesellschaft, 1950, Volume 1 and 2.

Vasicek, Antonin -- "Optics of thin films," Amsterdam, North-Holland Publishing Co., 1959.

22 INFRARED OPTICAL DESIGN

22.1 INTRODUCTION

22.1.1 General. The basic principles of optical design for the infrared region are the same as those for visible and ultraviolet light. The differences arise mainly from the nature of the materials which must be used, and from the operational and environmental requirements of most of the current applications.

22.2 INFRARED OPTICAL MATERIAL

22.2.1 Image converter tube. Reflecting and refracting materials suitable for use at the various infrared wavelengths, have been discussed in Section 16. In particular, reference was made to the publication by Ballard, McCarthy, and Wolfe, tabulating information on currently available materials. (Development work is active in this field, and the designer should keep abreast of the situation with appropriate journals and other sources of possible information on new materials). Only general comments on materials will be made here, although they will be extended somewhat in the subsequent portions of this section. Radiation in the 0.8 to 1.2μ region is used with night vision devices employing image converter tubes, such as the "Sniperscope". These are ordinarily "active" devices. That is, they are used to look at objects which are illuminated by infrared light from a source which is under control of the user. The light source is usually a tungsten lamp or carbon arc covered by a filter which absorbs the visible light while passing the infrared. The effective wavelength range results from the combination of the spectral characteristics of the source, of the filter, and of the photo-sensitive cathode of the image converter tube.

22.2.2 Infrared imagery. In infrared use, an objective similar to a photographic objective forms an image, in infrared light, on the photocathode of the tube. The illuminated areas of the cathode emit electrons which are accelerated and focussed on a fluorescent screen at the opposite end of the tube, thus forming a visible image which can be viewed by the user with the aid of a magnifier.

22.2.3 Glass for infrared usage. Ordinary optical glass transmits satisfactorily in this region and is always used. However, the dispersion characteristics of the several types become much more nearly alike in the infrared than they are in the visible and, as a consequence, much stronger powers of crowns and flints are required to obtain achromatization. For example, a doublet with a 100mm. focal length, made of light barium crown (1.5725/57.4) and dense flint (1.6170/36.6), and achromatized in the visible, will have a crown with a 36mm. focal length and a flint with a focal length of 57mm. If the same glasses be used to achromatize the doublet in the region from 0.8μ to 1.2μ, the crown must have a focal length of 19mm. and the flint 23mm. The chromatic aberration of a single crown lens in this region is approximately two-thirds to three-quarters of that for a lens of the same glass from the C line to the F line in the visible, and some slight advantage can be taken of this fact. It is still true, however, that in using a basic lens type, e.g., a Petzval, it is sometimes necessary to replace a single crown lens by two, in order to avoid the high-order aberrations which would otherwise result from the strong curves necessary for achromatization.

22.2.4 Optical glass infrared absorption. Radiation in the region from the visible to about 3.5μ is within the range of usefulness of lead sulfide cells. Ordinary optical glass begins to absorb slightly at about 2.0μ, and the absorption becomes very great at approximately 2.6μ to 2.7μ, depending on the type. Consequently, ordinary glass cannot be used for systems requiring performance beyond 2.7μ. Although the usefulness of lead sulfide cells extends to 3.5μ or beyond, it may happen that the combination of (1) the spectral characteristics of the source, (2) any filters in the system, (3) the intervening medium such as air, and (4) the lead sulfide cell itself, will produce a situation under which only a very small proportion of the response of the cell would be lost by using ordinary glass. In such a case, it is worth while to consider carefully whether the loss of a slight amount of response is sufficient to outweigh the advantages of using ordinary glass. It is well to study a number of glasses in this connection, since there is some variability in transmission from type to type, remembering that the flints as a class, transmit slightly better than the crowns.

22.2.5 Materials suitable for wavelengths beyond 2.7μ. The materials available and suitable for use at wavelengths beyond approximately 2.7μ have, for the most part, refractive characteristics quite different from those with which the designer works in the visible and the near infrared. Indices of refraction range from approximately 1.35 (lithium fluoride) to 4.1 (germanium). The range of dispersion characteristics is even more striking. With ordinary glass, the ration of ν values available for achromatization is limited to about 2.4:1 or less. In the infrared, this range may run as high as 46:1, the value for a positive silicon element and a negative element in the 3.5 - 5.5μ region. In spite of this great range of values of optical constants, it usually turns out that for reasons not primarily optical, only a few media are available for a given application. The designer must make his selection from those few, and determine by trial which combination allows him to get the best correction within the limits of allowable complexity of the system. For this reason, among others, reflecting systems are frequently used. The reflectors are completely achromatized, of course, and have nearly constant reflectance over a large range of wavelength. Refracting elements are used with the reflectors to control aberrations.

22. 3 ENVIRONMENTAL REQUIREMENTS

22. 3. 1 Current applications. Currently, most designs of infrared optical systems are intended for use by the Military, and therefore the systems must meet Military requirements for serviceability after long exposure to adverse environmental conditions. In particular, many infrared systems are to be airborne, and must meet stringent requirements for compactness and lightness of weight, as well as the ability to withstand rapid changes in temperature and humidity without damage. These requirements place more rigid limitations on the choice of materials, and on the elaborateness of system, than are encountered in other applications.

22. 3. 2 Choice of materials. The choice of material for the windows of airborne equipment is one which must be based primarily on considerations of this nature. These windows are sometimes flat, but are more frequently dome-shaped, since they are used with scanning equipment. Being exposed on the exterior of the vehicle, they must be resistant to the variations of temperature and humidity to be expected in service. The material must be hard enough to withstand excessive scratching; for example, from dust kicked up during take-off and landing. Particularly when used in supersonic vehicles, the material must be able to withstand the thermal shock resulting from friction with the atmosphere. When heated from such friction, it must not radiate much energy in the infrared region in which the optical system is operating; otherwise a false signal may be generated or a true ore obscured. Of course, a material radiates only in the region in which it absorbs (See Section 16) and therefore, ordinarily a material transmitting well enough to be considered for use as a window will not give trouble in this respect. However, because of the scarcity of suitable materials, it is sometimes necessary to consider those which do have slight absorption in the region of use and the possible effect of such unwanted radiation must be considered. In this connection, it is important to know the transmission characteristics of the material at elevated temperatures, since these characteristics may differ significantly from those at ordinary temperatures.

22. 3. 3 Size limitations. The window material must be obtainable in pieces large enough for the intended use. This requirement frequently rules out a number of otherwise promising materials. Some attempts at getting around this difficulty have been made by using segmented windows made up of a number of small pieces, and by replacing domes by polyhedrons made up of small, flat pieces. Such structures do not seem to have been generally satisfactory, however, probably because of the difficulty of providing adequate strength in combination with freedom from excessive obscuring by the supporting framework.

22. 4 OPERATIONAL REQUIREMENTS

22. 4. 1 Detection of infrared. To be useful, the infrared optical system must feed the energy it collects into a photosensitive device of some type. Fundamentally then, the design of the complete infrared instrument requires similtaneous consideration for the optics, the photosensitive device, and the associated equipment (usually electronic in nature), with respect to the performance requirements to be met. With respect to this discussion, photosensitive devices will only be described for the purposes of orientation.

22. 4. 2 Classification by instruments. Infrared devices are customarily classified by systems engineers as "image forming", or "non-image forming". An "image forming" device is an instrument whose output is a visual pictorial display of the field viewed by the device. An example is the "Sniperscope" previously mentioned. A "non-image forming" device is an instrument whose output is a signal, which is usually electrical. An example is an instrument giving information of the presence of a target of some nature in a particular portion of the field of view. This classification, while logical from the systems engineer's point of view, is not always very significant to the optical designer, since the optical systems of many "non-image forming" devices must actually have an optical image somewhere in the system in order to permit the location of a target within a particular portion of the field of view. Similarly, some "image-forming" devices, which depend on scanning procedures, require optical systems which simply condense the energy from a small field onto a photocell.

22. 4. 3 Classification by wavelength range. For the optical designer, a more useful classification of infrared devices is by the wavelength range which the instrument utilizes. For the purpose of this discussion, the range of the near infrared will include the region from 0.75μ to 3.0μ and beyond the near infrared will include the region from 3.0μ to 1000μ.

22.5 THE NEAR INFRARED REGION

22.5.1 Current applications. There are three types of commonly used systems which work in the near infrared; that is, in the region from the visible to about 1.3μ. All three are image formers, both in the systems sense and in the optical sense. They are the infrared photographic process, the image-converter tube systems, and the triggered radiation system. Ordinary optical glass is suitable at these wavelengths and the optical design is quite similar to that for photographic objectives, within limitations discussed in 17.7.2

22.5.2 Infrared photography. Infrared photography uses plates or films similar to those of photography with visible light, except that the emulsions have been sensitized by the addition of infrared-sensitive dyes.

22.5.3 Infrared image converter systems.

22.5.3.1 The system using an infrared image converter tube has an optical objective which forms an infrared image on the photosensitive cathode of the image converter as shown in Figure 22.1. The cathode emits electrons into the space within the tube, the rate of emission from a given area being proportional to the intensity of illumination of the area. The electrons are focused by electrostatic or electromagnetic means, in order to form an image on an electron-sensitive phosphor at the other end of the tube. The phosphor emits visible light, and the image can be seen by viewing the phosphor with the eye, usually with the aid of a magnifier.

22.5.3.2 Specifications for the components of the system are determined by a compromise between the desired performance characteristics and the state of the tube-maker's and the optical designer's arts. There are usually rather stringent requirements for compactness and portability. The instrument must operate as a telescope of a certain power, usually unity or greater. It is desirable to have a fast objective, since this increases the range at which the instrument is effective. Given the desired field angle, the size of the cathode of the tube determines the focal length of the objective, or vice versa. The electrostatic tubes (which are the sort usually used in this country) operate at a magnification less than unity. The sub-system consisting of objective and tube can then be considered as having an equivalent focal length equal to the focal length of the objective multiplied by the magnification of the image tube. (Both the objective and the electrostatic tube invert the image, so an erect image is presented on the phosphor.) For example, if the system is to have 1-1/2X power, the objective has a focal length of 50mm, and the image converter has a magnification of 0.7X; then the focal length

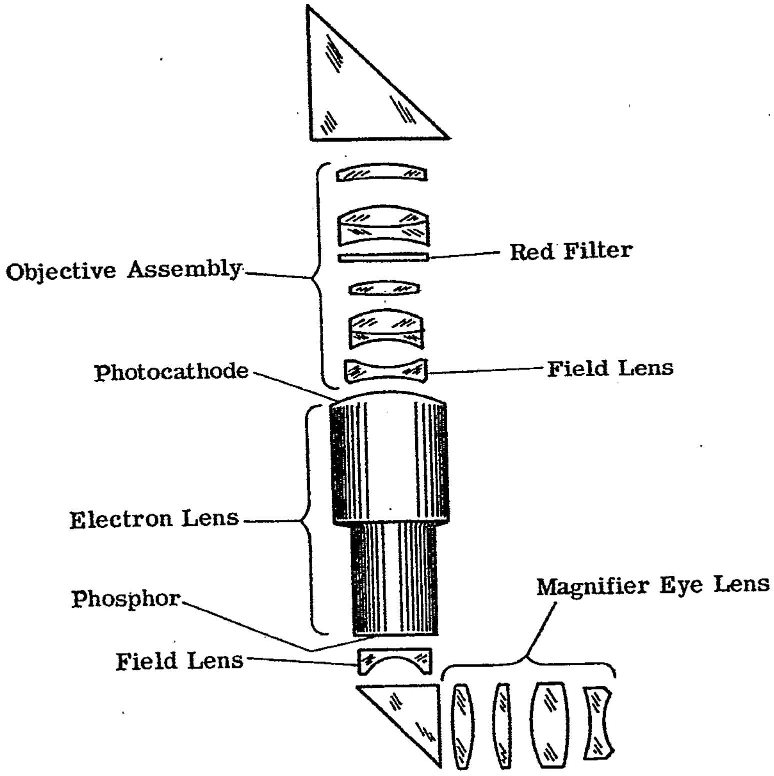


Figure 22.1- Optical schematic of an image converter system.

of the front system is 0.7X . 50mm = 35mm, and the magnifier must have a focal length of about 23mm.

22.5.3.3 The photocathode usually must be rather sharply convex toward the incident light, since the electronic, as well as the optical system, has field curvature. Thus, its curvature is of the opposite sign to that necessary to match the natural curvature of field of a refracting objective. For this reason, a strong negative field lens is employed in front of the cathode. The required power is such that it may be difficult to obtain the required correction for curvature with a single lens without getting total internal reflection of the light from the outer portions of the field; the work must then be divided between two elements. In addition to correcting the curvature, the field lens introduces coma, astigmatism and distortion, since it is not located exactly at the image. The coma and astigmatism must be balanced in the main part of the objective.

22.5.3.4 The electronic system of the image tube produces strong pincushion distortion in the image on the phosphor. This distortion, together with that of the objective and field lens, is dealt with, if at all, in the magnifier through which the phosphor is viewed.

22.5.3.5 The viewing lens system must be considered as a magnifier rather than an eyepiece. (See Section 13) Since the phosphor is a self-luminous surface, it emits light in all directions, and there is no natural exit pupil such as is present in an ordinary viewing telescope. This is an advantage in that there is more freedom to position the eye of the observer than would be the case with the presence of an exit pupil. However, it results in the necessity of correcting the magnifier for spherical aberration and coma, in order to avoid weird distortions and blurrings which can occur if the observer's eye does not happen to be exactly on axis. (This correction can be much cruder than necessary in an objective, since the pupil of the eye accepts only a portion of the bundle from a given object point at any one instant).

22.5.3.6 The conditions of use are such that it may be advantageous to use one or more aspheric surfaces in the magnifier. Aspherics of sufficient precision can be made by processes suitable for mass production. Such a magnifier is shown schematically in Figure 22.2. The magnifier consists of an eye lens and a field lens. One surface of the eye lens is aspherized to correct for spherical aberration. The bending of the lens is chosen to minimize coma. The field lens is aspherized to compensate for the pincushion distortion at the phosphor. (The aspheric may be given a slight power on axis to facilitate fabrication.)

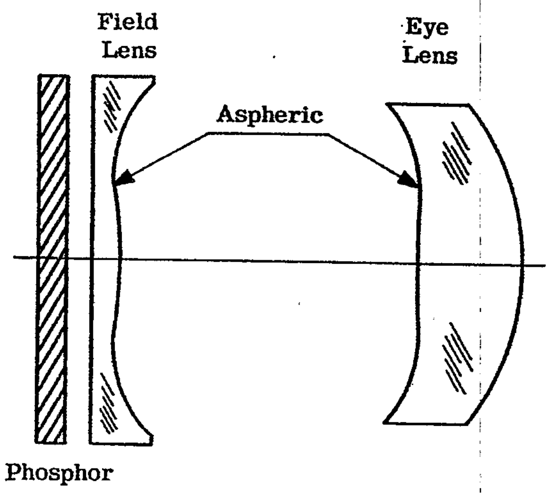


Figure 22.2- Aspheric Magnifier.

22.5.4 The triggered radiation type.

22.5.4.1 Instruments of the this type depend on the ability of some phosphors to store energy when irradiated by short wave radiation, and to emit it as visible light when triggered by irradiation with infrared. The short-wave radiation may be ultraviolet (or visible) light, or that from a bit of radioactive material. An objective forms the infrared image of the scene being viewed on the phosphor. The various parts of the phosphor emit visible light in proportion to the intensity of the infrared radiation.

22.5.4.2 In this system, it is necessary to provide optical means of inverting the image, since the system would otherwise perform like an inverting astronomical telescope. This may be done by using a lens or a prism erecting system, either preceding the phosphor (thus working in the infrared) or between the phosphor and the eye. (Other ingenious means have been used in special designs).

22.6 THE INTERMEDIATE AND FAR INFRARED REGION

22.6.1 Distinguishing characteristics.

22.6.1.1 Design in the region beyond the near infrared has two main distinguishing characteristics, in addition to the necessity of using materials other than optical glass. One is the limitation on image quality. The other is the limitation imposed by the combination of the performance requirements and the characteristics of the types of energy detectors which must be used in many applications.

22.6.1.2 Aside from these limitations, the design requirements are much as they are in other wavelength regions, and the designer must be prepared to deal with requirements quite similar to those found in other optical designs.

22.6.2 Limitation on image quality.

22.6.2.1 Since the diffraction limit of resolving power (see 16.28) depends on the wavelength of the light being used, the best image quality obtainable from a source with a given aperture is much poorer in the infrared than in the visible. Taking 0.56μ as typical of the visible region, the resolution at 3μ is five times as coarse, and at 10μ is eighteen times as coarse as in the visible. Since there is no gain to be achieved from improving the correction beyond the point at which the Rayleigh criterion is satisfied, the designer may stop his work with a residue of aberration which it might be well worth while to remove if he were working in the visible region. He may also have to warn the proposer of the system of the limited resolving power which can be obtained.

22.6.2.2 Currently the resolution requirements of most systems are even coarser than the limit which would be imposed by diffraction. However this is not always the case, even at present, and as the infrared art develops it is likely that there will be many more requirements for performance near the diffraction limit.

22.6.3 General functions of the optical system.

22.6.3.1 As an aid in discussing the relation of the energy detector to infrared optical design it is worth while to review the functions of the optical instrument in general terms.

22.6.3.2 Every optical instrument is designed to obtain information concerning the radiation characteristics of a portion of space. This portion of space is called the field of view of the instrument. It may be, for example, the crater of an arc, as in emission spectroscopy; a volume of space, as in absorption spectroscopy or in an infrared search system; or a surface, as in a slide projector. The radiation may or may not originate in the field; that is, the field may or may not be self-luminous. In emission spectroscopy and pyrometry the field is self-luminous. In absorption spectroscopy and with active viewing systems it is not self-luminous. In missile guidance systems a part of the field, the target, is self-luminous, while the background light for the most part originates outside the field of view. The importance of the distinction between self-luminosity and non-self-luminosity is only secondary. More basic is the question of whether it is possible to control the nature of the radiation, either in the design of the equipment or during its use.

22.6.3.3 One important function of the optical system is the rejection of radiation from outside the field of view. The field stop in many instruments is an embodiment of this function. In other cases, as in certain types of scanning systems, the limitation of the field is obtained by more elaborate means.

22.6.3.4 The type of information to be obtained from the field depends on the application. In spectroscopy, the object is to obtain a measure of intensity as a function of wavelength, without regard to the portion of the field of view from which the radiation comes. In spectroscopic systems for on-stream process control, the object is, in addition, to present this information, or a portion of it, as a function of time. In image-forming systems, the object is to have rather detailed information concerning radiation intensity as a function of position throughout the field. If the image-forming system is of the "color-translation" type, at least some information about the spectral distribution of intensity at each point of the field must also be provided. In infrared search systems, the object is to know, from moment to moment, the presence and location within the field of small areas, or targets, having radiation characteristics slightly different from those of the remaining background portions of the field.

22.6.4 Detector characteristics.

22.6.4.1 The manner in which the information is obtained from the incoming radiation, and in fact, subject to the over all requirements for the instrument, the precise nature of the information, depends greatly on the kind of energy detector which can be used.

22.6.4.2 To the optical designer, the energy detector is the last surface of his system, which must receive and absorb all the useful energy collected by the system. The nature of the detector and its associated equipment imposes limitations on his choice in bringing this about, and also on the minimum of light-gathering power which he must build into the optical system. Several characteristics of the detector are worthy of discussion.

22.6.4.3 For the purposes of this discussion, the detector may be considered as a figurative "black box" with an input, the radiation, and an output, usually an electrical signal in infrared devices. Ordinarily it is more useful to consider the input as flux, or flux per unit area of the detector, rather than as total energy. The flux may be expressed in watts, or in some other unit of power. As will be seen in the following paragraph, it is frequently necessary to consider the distribution of the flux as a function of wavelength. The output is measured in appropriate units, in megohms for example if it is the change in electrical resistance of the cell due to the incident radiant power. The responsivity of the cell is the output per unit input; in our example the number of megohms change in resistance per watt of input. (The term sensitivity is sometimes used for what is here called responsivity, but the word sensitivity has also been employed for a number of other concepts, so its use will be avoided altogether in this discussion.)

22.6.4.4 Two qualifications must be put on this concept. In the first place, the output per unit input may depend on the wavelength of the radiation. Thus to characterize a detector adequately it is necessary to give its responsivity as a function of wavelength; and to predict its response, the distribution of the incident power as a function of wavelength must be known. As a class, the detectors known as photoelectric detectors are highly wavelength dependent. The thermal detectors as a class have substantially constant responsivity regardless of wavelength, and the spectral distribution of the radiation can be ignored.

22.6.4.5 The second qualification arises from the fact that the output may not be strictly a linear function of the input, even allowing for spectral effects. Many detectors show saturation effects when strongly irradiated. Frequently the detector is operated under conditions such that the response is substantially linear. In other cases the concept of responsivity must be modified suitably.

22.6.4.6 Another important characteristic of the cell is called its detectivity. It is a measure of the smallest input, or smallest change in input, that can be reliably detected. All detectors have a random output, not related to the input, known as noise. As a rule of thumb, the increment of input necessary to produce an increment of output equal to the noise may be taken as the minimum detectable input. When expressed as power, this is known as "noise-equivalent power". The larger the noise-equivalent power, the poorer is the detector for small inputs. The reciprocal of the noise-equivalent power is called the detectivity. (The word sensitivity has sometimes been used to mean the detectivity.) Detectivity depends on the type of detector, on the way it is made, and on the environmental and electrical conditions under which it is used. Ordinarily the detectivity is improved by keeping the detector area small.

22.6.4.7 When the input is suddenly changed, the output does not change instantaneously, but takes a finite time to adjust to the new level. The time constant is a measure of the time required for such an adjustment. It is important in predicting the response of the detector to short bursts of radiation, and in determining its suitability for use in scanning systems and in other systems in which the input is made to vary at high frequency. As a class, thermal detectors have much larger time constants than photodetectors.

22.6.4.8 In choosing the size of the detector it is to be remembered that flux density, rather than total flux on the detector, is the criterion of the amount of output which will be obtained. For example, if two square lead sulfide cells of similar characteristics be operated under similar conditions, but the area of one is twice the area of the other, then the outputs of the two will be equal when the flux per unit area on the two is the same, although the total flux on the larger is then twice that on the smaller. Sometimes a large cell may be operated under conditions not practical with a smaller one so as to produce a higher output at a given flux density (for example, a photocell of large area can safely be operated at a higher bias voltage than a small one), but in many cases this is not practical (for example, the voltage available for biasing may be limited) and flux density is the criterion of the output obtainable. In general this is an advantageous situation, since it is usually possible to use a more compact optical system to produce a given flux density on a small area than on a large one. Within limits, small detectors have better detectivity than large ones of the same kind. As a consequence in critical situations where detectivity is an important characteristic small cells are used. In photoconductors, dimensions of a few tenths of a millimeter are common.

22.6.4.9 The responsivity of a detector may depend somewhat on the way the flux is distributed over its surface. There may be local "hot spots" which are more responsive than the rest of the surface. Since the photoconductive cells are used with a bias voltage, the response depends on the way the radiation falls with respect to the points at which the leads make contact with the detector. For this reason it is desirable to plan the optical system so that the non-uniformity will not be a disadvantage. This ordinarily means that the detector is not placed at, or immediately adjacent to, an image plane, but rather at an image of the entrance aperture of the system.

22.6.5 Target detection and location.

22.6.5.1 An important class of problems is exemplified by airborne infrared search systems. In applications of this sort it is necessary to have a large field of view, and to detect and locate small, weak targets which are at great distances from the optical system.

22.6.5.2 There may be large amounts of radiation in the field besides that coming from the targets which it is desired to detect. Such radiation is called "background". The problem thus is one of distinguishing the radiation of the target from that of the background. It is necessary to take all possible advantage of differences between the target and the background. One important difference lies in spectral distribution, the target radiation usually having a peak wavelength different from that of the background. (The spectral distribution must be evaluated at the optical system. The intervening atmosphere is in effect a part of the field of view, and its absorbing and scattering characteristics must be taken into account.) Contrast between target and background is further increased by the use of optical filters to absorb as much as practical of the radiation at wavelengths at which the background is stronger than the target. (For a discussion of infrared filters see Ballard et al, loc. cit.) Choice of the type of detector depends in part on its having adequate responsivity in the spectral region near the peak wavelength of the target.

22.6.5.3 The technique known as spatial filtering is frequently used to take advantage of the dimensional differences between the target and other sources of radiation likely to be in the background. For a discussion of spatial filtering see, for example, Aroyan. *

22.6.5.4 Detection and location of the target within the field of view is accomplished by dividing the field into elements by some means and observing either the difference in flux between an element and adjacent ones, or the change in flux in each element with time. Since the intensity difference between target and background is small, the detector must be chosen and used so as to have good detectivity, and the optical system must have a large aperture to insure that the difference between target and background can be recognized by the detector.

22.6.5.5 An attractively simple scheme for providing the necessary subdivision of the field uses an objective which forms an image of the field at its focal plane. In the plane of the image is placed a rotating opaque plate carrying a set of small apertures so arranged that at any instant a single aperture is transmitting light from some small portion of the image, and during a single rotation of the plate the whole image is scanned. A condenser system placed behind the image plane collects the radiation and brings it to the detector. (The condenser is usually designed to form an image of the aperture of the objective on the detector.) The rotation of the plate can be related electrically to the output of the detector so that the system as a whole recognizes the portion of the field from which radiation is being transmitted at any instant. Attractive though it is, this simple scheme is rarely adequate because of the simultaneous requirements of large field and wide aperture. The inadequacy is not due entirely to lack of ingenuity on the part of the optical designer, but results from a fundamental limitation on the light-receiving ability of a small surface.

22.6.6 Receiving ability of a surface as a limiting factor.

22.6.6.1 The method used in the following analysis of the light-receiving ability of a surface is old, though it does not seem to be so well-known as desirable. See for example Drude. ** The method applies to any surface through which all the useful energy must pass, and thus its conclusions apply to focal surfaces as well as detector surfaces.

22.6.6.2 Suppose the instrument to be confronted by a black body, the surface of which is at least large enough

*Aroyan, G. F. "The Technique of Spatial Filtering." Proc. I.R.E. 47; 1561-68; Sept. 1959.

**Drude, Paul, "Lehrbuch der Optik," Leipzig, 1900. English trans., "The Theory of Optics," N.Y. and Dover, 1959 to fill the whole field of view of the instrument. That is, it is large enough so that any ray which enters the optical system and passes into the surface whose light-receiving ability is being investigated can be considered to have originated in the surface of the black body. It is convenient though not necessary to assume that the black body is infinitely distant from the instrument. Assume that the black body is at some fixed, uniform temperature.

22.6.6.3 Suppose that the surface whose light-receiving ability is being investigated is the surface of a black body which is at the same temperature as that of the external black body. It follows from the second law of thermodynamics that, regardless of the nature of the optical system, the internal surface cannot receive from the external one an amount of flux greater than that which the internal surface itself is radiating, for otherwise the system would be acting as a self-operating heat pump. In most systems the flux received by the internal surface from the external one will be considerably less than the maximum, due to the finite aperture of the system, absorptions within the system, etc.

22.6.6.4 Let I be the number of watts per unit solid angle radiated by either surface per unit area of the surface in the direction normal to the surface. This quantity is determined by the black body temperature, and for our purposes is to be considered constant. Let A be the area of the surface whose light-receiving ability is being investigated, and n the index of refraction of the medium on that side of this surface on which the light is incident. (The detector surface may be exposed to air, for example, or it may be in optical contact with glass, light reaching the surface through the glass.) Let F_s be the total flux radiated by the surface into the medium with which it is in contact. Then it can be shown that

$$F_s = \pi I n^2 A \qquad (1)$$

(Drude, loc. cit.) Then F_s also represents the upper limit of the flux we can hope the surface would be able to receive from the external black body.

22.6.6.5 As an example of the significance of this result, suppose that a system is contemplated which has an objective aperture 150mm. in diameter, at the front of the system. It is desired to have the objective focus the energy from a 20° field (i.e., from 0° to 10° off-axis) on an image plane, and then to condense the light on a photocell. It is desired to determine the minimum focal plane area and minimum cell area. It is first necessary to write down the expression for the flux received at the aperture from the 20° field. For purposes of later discussion this formulation will be made more elaborate than would otherwise be necessary. Let

$$\overline{A} = \text{the area of the aperture} = 75^2 \pi \text{ sq. mm.}$$

22.6.6.6 It can be shown that the element of flux dF received at the aperture from that elemental portion of the external blackbody which lies in a direction making an angle α with the axis of the system (i.e., with the normal to the aperture) and which subtends a solid angle dw as seen from the aperture is

$$dF = I \overline{A} \cos\alpha \; dw.$$

Here I has the same value as in (1) since both the internal and the external blackbody are at the same temperature. The quantity $\overline{A} \cos\alpha$ is the projection of the area of the aperture in the direction of the source. From the point in the external blackbody on the axis of the system the aperture appears circular. From points farther away from the axis the aperture appears foreshortened, so that the effective area is only $A \cos\alpha$. Let

$$B(\alpha) = \overline{A} \cos\alpha$$

and name

$$B(\alpha) \quad \underline{\text{the effective aperture function.}}$$

Then

$$dF = I \; B(\alpha) \; dw.$$

22.6.6.7 It is convenient to consider the elementary part of the external blackbody as being the annulus included between the two cones corresponding to field angles of α and $\alpha + d\alpha$ respectively. Then

$$dw = 2\pi \sin\alpha \; d\alpha$$

$$dF = 2\pi I B(\alpha) \sin\alpha \; d\alpha \qquad (2)$$

$$F = 2\pi I \int_0^{\alpha'} B(\alpha) \sin\alpha \; d\alpha . \qquad (3)$$

Here α' is the maximum value of the field angle from which radiation is to be collected.

22.6.6.8 In this example, since $B(\alpha)$ is known, (3) can of course be written

$$F = 2\pi I \bar{A} \int_0^{\alpha'} \cos\alpha \ \sin\alpha \ d\alpha .$$

$$= \pi I \bar{A} \sin^2 \alpha' , \tag{4}$$

or substituting for $\bar{A}$ and α',

$$F = \pi^2 I \ 75^2 \sin^2 10° . \tag{5}$$

The second law and equations (1) and (5) then require

$$F \leq F_s ,$$

whence

$$\pi^2 \ I \ 75^2 \ \sin^2 10° . \leq \pi I n^2 A$$

and

$$A \geq 75^2 \ \pi \ \sin^2 \ 10°/n^2 = 533/n^2 . \tag{6}$$

Thus if the focal surface must be in air, so that $n = 1$, we must expect its area to be greater than 533mm^2. Similarly, if the detector be optically immersed in a medium having an index of refraction $n = 2.2$, its area must be greater than 110mm^2. A detector having this area would have too poor a detectivity in many applications.

22.6.6.9 The treatment in the example can be generalized. In the example it was tacitly assumed that the system had rotational symmetry about the optic axis. Even if this is not the case it is still convenient to assume a set of spherical coordinates centered at some point in the optical system, using α for the polar angle and β for the azimuth angle. Consider an elementary portion of the external blackbody which lies in the direction (α, β) and subtends a solid angle dw as seen at the optical system. Consider a bundle of rays starting from the element of blackbody. When the rays reach the system, some will pass in and reach the focal plane and ultimately the detector. The others will be excluded by the various apertures of the system. The cross-sectional area of that portion of the bundle which does eventually reach the detector may be called the effective aperture of the system for the direction (α, β). We denote it by $B(\alpha, \beta)$, which may be called the effective aperture function of the system.

22.6.6.10 The flux collected by the system from the element of the blackbody is then

$$dF = I \, B(\alpha, \beta) \ dw \tag{7}$$

(we assume for the present that absorption and similar losses are negligible) and the total flux collected by the system is

$$F = I \int_\alpha \int_\beta B(\alpha, \beta) \ dw \tag{8}$$

the integral being taken over the whole field, i.e., over all directions (α, β) for which $B(\alpha, \beta) \neq 0$. It follows from the second law of thermodynamics, and from (1) and (8) that

$$\int_\alpha \int_\beta B(\alpha, \beta) \ dw \leq \pi n^2 A . \tag{9}$$

22.6.6.11 It is important to note that in (9) the radiation intensity, I, of the blackbody has cancelled out, and (9) is a condition on the characteristics of the system itself, regardless of the nature of the actual radiation field with which the system may be confronted. The expression on the right hand side may be taken as representing the maximum light-receiving ability of a surface of area A. The equation gives a fundamental limit to the combination of field coverage and aperture which can be achieved with a surface of area A.

22.6.6.12 When the system does have rotational symmetry we can proceed as in the example, considering $B(\alpha, \beta)$ to be a function of α only, and obtaining (3) which with (1) yields

$$2 \int_0^{\alpha'} B(\alpha) \ \sin\alpha \ d\alpha \leq n^2 A \tag{10}$$

22.6.7 Practical limits and techniques.

22.6.7.1 Of course practical difficulties prevent attaining the limiting light-receiving ability. To do so would require bringing in rays at all angles of incidence to the surface up to 90°. This is difficult to accomplish because of the precision required in constructing and focussing, and because of the resulting sizes and shapes required for the optical components. As a rule of thumb, something like 1/2 to 3/4 of the limiting light-receiving ability can be utilized. The former requires angles of incidence of 45° or greater at the surface; the latter, 60° or more. Consequently in practice one is limited to

$$\int_\alpha \int_\beta B(\alpha,\beta)\ dw \leqq k \pi n^2 A \qquad (9')$$

or

$$2 \int_0^{\alpha'} B(\alpha)\ \sin\alpha\ d\alpha \leqq k\, n^2 A \qquad (10')$$

wherein k may be 1/2 to 3/4. Losses by absorption and reflection in the system must of course also be taken into account.

22.6.7.2 As another example, suppose it is desired to monitor at 120° field, and that a detector area of 0.25mm^2 is chosen as having optimum detectivity under the expected operating conditions. Assume further that from knowledge of expected targets and backgrounds an aperture of at least 4400mm^2 is considered necessary to insure that the target will be recognized by the detector. (This aperture area is that of a circle 75mm in diameter). It is desired to use a simple scanning system of the type described in 22.6.5.5. The designer must decide whether he can do this.

22.6.7.3 It is evident at once that although in use the scanning plate permits only a small portion of the field to be viewed by the detector at any instant, nevertheless the system must be designed so that, if the plate were removed, the detector would view the whole field at once. That is, the light-receiving ability of the detector must be made to cover the whole field at the full aperture.

22.6.7.4 Without attempting to decide for the moment how it can be accomplished, assume that B (α) is to be made constant for the whole field of view, and equal to 4400mm^2. Then equation (10') becomes

$$8800 \int_0^{60^\circ} \sin\alpha\ d\alpha \leqq 0.25\, k\, n^2$$

or

$$8800\,(1 - \cos 60^\circ) \leq 0.25\, k\, n^2 .$$

Assuming k may be taken as 3/4, this becomes $4400 \leq 0.1875\ n^2$.

22.6.7.5 If the cell is to operate in air, or a vacuum, n = 1, and the inequality obviously is not satisfied. The system as proposed cannot possibly meet the requirements. The scheme would have to be abandoned and another adopted which permits the detector to be employed in a system with adequate aperture but with a limited field.

22.6.7.6 There are various ways of doing this. A simple one consists essentially of building a system of adequate aperture and suitably small field, and then pointing it rapidly first at one portion of the field and then another. Thus each part of the large field is observed intermittently, although not continuously. The pointing is usually done by rotating mirrors. Another scheme causes the detector to scan the image at the focal plane of an objective by moving it about in the focal plane. The difference between this scheme and the one originally suggested lies in the fact that here the whole light-receiving ability of the cell is operative on the small portion of the field being viewed at the moment, while in the former only a small part was used at any instant, the rest being blocked off by the opaque portions of the rotating plate. Still other schemes use an array of cells to scan the image, or a mosaic of cells fixed in position in the image plane, or a combination of these methods. See Figure 22.3. A complete discussion of choice of an optimum system is too lengthy to be included here.

22.6.7.7 In equations (9), (10), (9') and (10') the square of the index of refraction, n^2, appears on the right hand side. This is indicative of the fact that the light-receiving ability of a detector surface is increased if it is in optical contact with some medium other than air. From analogy with oil-immersion microscope objectives, it is customary to say that the surface is immersed in the medium. The increase in light-receiving ability is analogous to the increase in numerical aperture of a microscope objective obtained by using an immersion system. The surfaces of the glass or crystal on which the cell is formed must be so disposed as to permit light from the field of view to strike the cell at all incidences from zero up to very high angles. If the substrate is simply a plane parallel plate in air, for example, the critical angle of refraction in the glass will limit the light-receiving power to that which the cell would have in air. If the cell is placed at the center of a hemisphere of the glass, however, the full light-receiving power can be used.

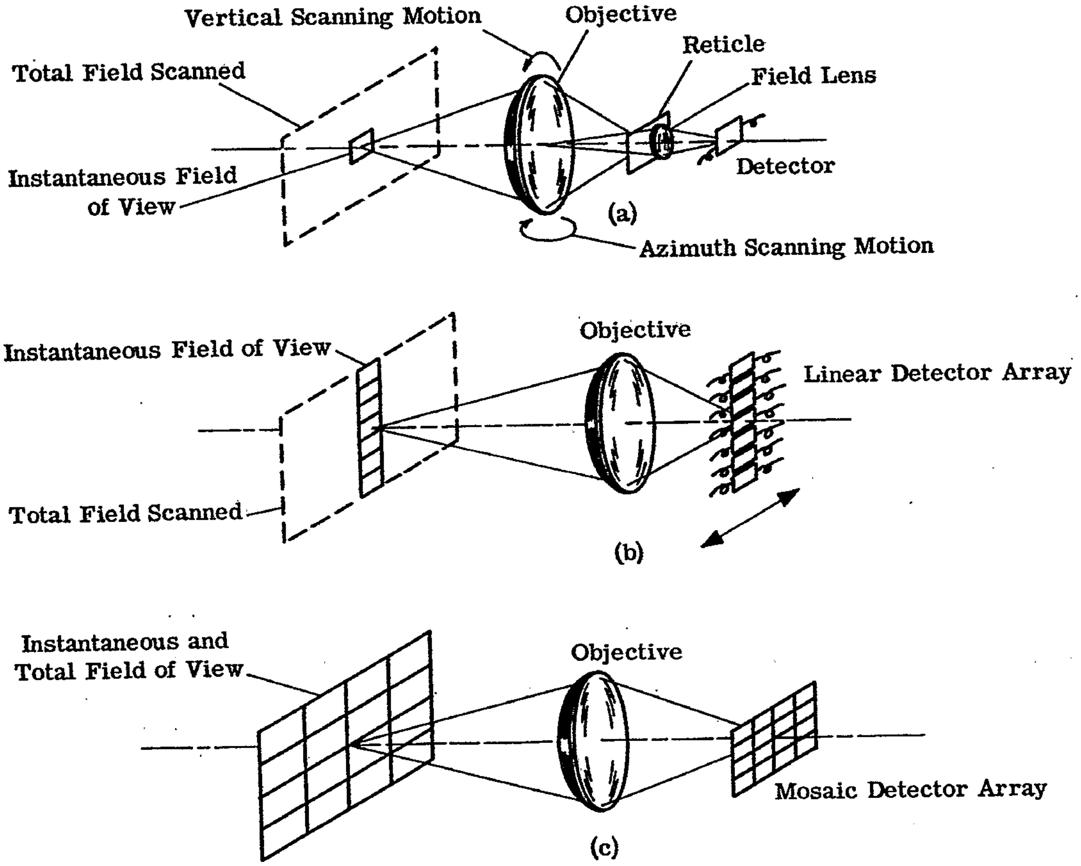


Figure 22.3 - Examples of basic scanning systems.

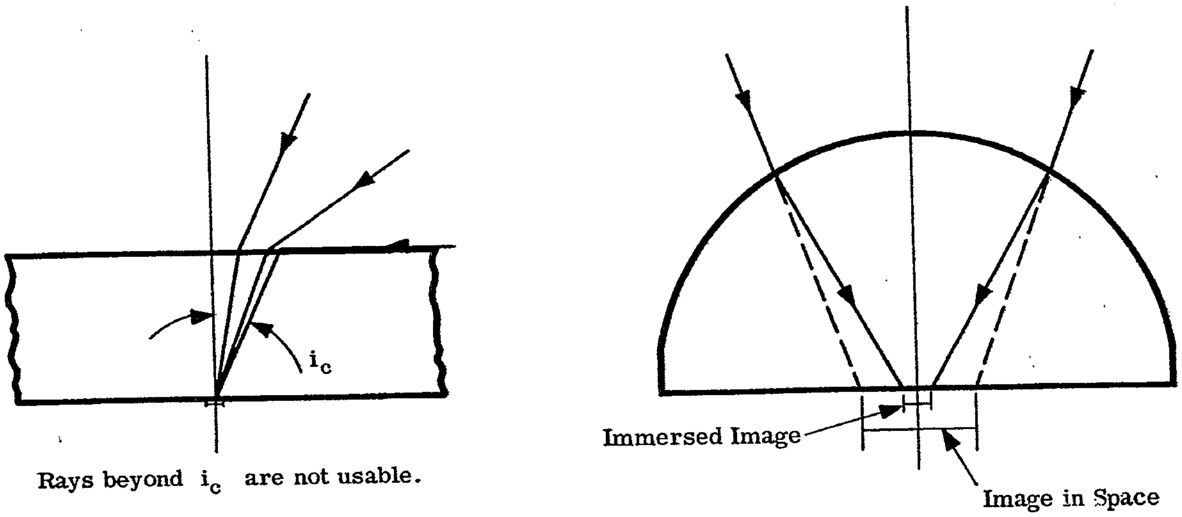


Figure 22.4- Illustration of advantages of an immersed detector.

22.6.7.8 The benefits of cell immersion are often overlooked, and the design of a system is thereby made harder. However, as the possibilities become more widely appreciated, cell makers are giving more attention to the problems of producing immersed cells. These problems are difficult, because in addition to being a suitable material for supporting the cell, the substrate must also transmit the desired radiation. The higher the index, the greater the increase in light-gathering ability. Strontium titanate, for example, has an index of 2.2, and increases the light-receiving ability of the cell by a factor of 4.8.

22.7 SUMMARY AND CONCLUSION.

22.7.1 Advantages and disadvantages. In infrared work the designer may meet problems as varied and complex as those encountered in the visible part of the spectrum. Since the laws of reflection and refraction are the same at all wavelengths, the same basic design principles are used in both regions. The most important differences to which he must become accustomed arise from the natures of the available optical materials on the one hand, and from the requirements of some of the currently important infrared applications in the other. In addition he must remember that the resolving power obtainable with a given aperture is poorer than in the visible, due to the longer wavelength of the light. Out to about 2.7μ he may use ordinary glasses, but must allow for the fact that the dispersion characteristics of the several glass types are more nearly alike than in the visible. Beyond 2.7μ he must use materials whose characteristics may vary widely from those of optical glass, sometimes favorably and sometimes not. He will want to use reflecting systems more frequently than in the visible. The use of many infrared devices for military applications, particularly airborne ones, adds requirements of ruggedness, resistance to adverse environmental conditions, compactness and lightness of weight to the optical ones. Most such devices are part of instruments which are complex combinations of optics, mechanics and electronics, and the choice of the basic optical characteristics is only part of the process of choosing the optimum design parameters to meet the performance goals for the whole instrument. The designer needs to know enough about the characteristics of the whole instrument and the interrelationships of its parts to be able to contribute intelligently to the decisions in the choice of parameters. Especially, he needs to know something about the energy detectors used in the infrared, and how their limitations of responsitivity and detectivity limit his design.

23 MICROSCOPE OPTICS

23.1 INTRODUCTION

23.1.1 Scope. The material in this section will be devoted primarily to a discussion of the compound microscope, its characteristics, components, and various special purpose adaptations. However, in any discussion relating to visual instruments, the designer must keep in mind that the eye of the observer is an integral part of the optical combination, and that the degree of optical perfection in the human eye is as influential on the final retinal image, as is the degree of image perfection formed by the instrument's optical elements. The reader is urged therefore, to refer to Section 4 for a discussion of visual optics.

23.1.2 Functional relationships of microscope components.

23.1.2.1 The primary function of the high-power, compound microscope is to obtain information regarding the structure and optical characteristics of small specimens. This information is obtained by visually interpretating the manner in which the light transmitted by, or reflected from, the specimen is affected.

23.1.2.2 Usually, the specimen must be illuminated by intense artificial light, and it is only in rare and special cases that the specimen can be self-illuminated. However, the action of the specimen on the illumination system used may consist of absorption, reflection, diffraction, scattering, birefringence, or localized changes in the phase of the illuminating light waves. The purpose of the microscope then, is to form an image, based on the action of the specimen on the illuminating light waves, which can be interpreted in terms of the particular information with respect to the specimen, that is desired.

23.1.2.3 Since the human eye is only sensitive to color and intensity contrasts, the information derived from the image by the observer must be interpreted from these two effects.

23.1.2.4 The primary source of light can be of any number of high intensity light sources, however the light used must be concentrated on the specimen by a condenser system. The specimen affects the light as stated in paragraph 23.1.2.2, and the objective system of the microscope must be capable of receiving the altered light so that the maximum effects of diffraction, absorption, scattering, etc., may be transmitted by the objective and appear in the image as interpretable spacial variations.

23.1.2.5 In order to interpret the spacial variations in illumination, the objective must be capable of accepting and transmitting a wide angular beam of light, since the effects of the specimen on the light, as previously mentioned, especially diffraction, fan out from the specimen over wide angles. In the important case of diffraction, the more of the spectral orders the objective can receive, the more exact is the correspondence between the specimen and the structural details of the image.

23.1.2.6 Another requirement of the microscope objective is that the points and lines in a specimen be imaged sharply, so that the details in the image have a point by point correspondence with those in the specimen. This requirement necessitates a high degree of correction for aberrations.

23.1.2.7 All the available information regarding the specimen, as a result of its action on transmitted or reflected light, is contained in the primary image formed by the objective. As long as this information is contained in the primary image it is useless, since it must be interpreted in the brain of the observer. The simplest method for increasing the interpretability of the image is by means of magnification. By means of the eyepiece, the smallest significant details of the image can be resolved by the human eye.

23.1.2.8 There are other intermediate means available for interpreting the information from a specimen. The primary image may be magnified by projection, and formed on a photographic plate. The image could also be viewed by a television tube and an enlarged image presented on a screen.

23.1.2.9 To summarize then, the compound microscope is an instrument which transforms the action of a small object specimen on light waves into interpretable visual impressions, and in a broad sense any apparatus which accomplishes this function may be designated as a microscope.

23.2 CHARACTERISTICS

23.2.1 General. The compound microscope is characterized by the following requirements: high magnification (without a sacrifice of definition over a restricted size of field), a comparatively small true angular field, an illumination system, and resolution limited only by the wavelength of light and the numerical aperture of the objective. Similarly, it is desirable that the oblique aberrations be as well corrected as is consistent with the requirement for axial definition of the highest possible order. These characteristics determine the basic design of the compound microscope.

23.2.2 High magnification. Since the major function of the compound microscope is to view extremely small specimens, it must be capable of magnifying to such a degree that the smallest resolvable detail can also be resolved by the human eye. The highest useful magnification, expressed in diameters, is approximately a thousand times the numerical aperture of the objective used in the microscope. It should be noted however, that the compound microscope is not always used to view extremely small specimens and, in some instances, magnifications as low as 25 diameters are advantageous for viewing larger specimens.

23.2.3 True angular field. The true angular field of the microscope is, in most instances, small due to the following factors. The diameter of the primary image cannot be larger than that of the eyepiece, and present day eyepieces have become standardized in order to afford interchangeability with those of different manufacturers. The optical tube length, i.e., the image distance of the objective, in practice has become standardized, within limits, so that it is not difficult to interchange objectives made by different manufacturers, without significantly changing the magnification and correction of the objective. Since the true angular field, α, can be expressed as

$$\tan \frac{\alpha}{2} = \frac{y}{d} ,$$

where y is the half-diameter of the primary image, and d is the optical tube length, it may be seen that the true angular field has a maximum value in practice. The exceptions to this characteristic are the special microscopes which have extra large diameter eyepieces, and hence larger true angular fields. Actually, the size of the primary image is limited by the field diaphragm in the eyepiece, and this diaphragm must always be slightly smaller than the outside diameter of the eyepiece itself. In so-called negative type eyepieces, it is the virtual image of the diaphragm formed by the field lens which limits the primary image formed by the objective, but since the magnification of the field diaphragm by the field lens is not large, the statements above regarding field size are still applicable. The true angular field of the microscope may be considered to have a maximum value of less than 7°. For example, the half-diameter of the primary image field may be taken as not exceeding 10mm., and the optical tube length as 170mm. Substituting these values in the equation previously mentioned, it will be seen that the true angular field is 6°8'.

23.2.4 Illumination. The intensity of illumination in a compound microscope is a major factor due to the usually small size of the specimen being viewed, and because of the high magnification required to resolve the details of the specimen. As a result of the intense illumination required, light from an artificial source must be condensed onto the specimen. It is noted that in some cases, sunlight or skylight are used for illumination. Most specimens are thin and transmit light. For such specimens, the illumination falls on the back of the specimen, and the light is transmitted through the specimen into the microscope. For opaque specimens, the illumination is condensed upon the upper surface and only the light which is reflected from the specimen enters the microscope. This method of illumination is designated as vertical illumination. For most specimens however, transmitted illumination is used.

23.2.5 High resolution. High resolution is a basic requirement of the compound microscope, for it is upon this characteristic that the ability of the microscope to distinguish the fine details thereof, is based.

23.2.5.1 Factors determining resolving power. In compound microscopes, the source of light is most often an incandescent lamp provided with a condenser (in photomicrography and micro-projection, arc lamps are often used). The lamp condenser concentrates the light into a second condensing system, which is a part of the microscope proper and is known as the substage condenser. When vertical illumination is required, the substage condenser is not used, and other various forms of condensers are employed to condense light onto the specimen from above. Therefore, the resolving power of the microscope depends on the following factors:

(a) the size of the angle of the illuminating cone of rays passing through the specimen.

(b) the ability of the optical system to accept that which has been transmitted by the specimen and to transmit a wide cone of rays.

(c) the refractive index of the material between the specimen and the first surface of the optical system comprising the microscope's objective.

23.2.5.2 Limit of resolution. The concept of numerical aperture (N.A.) is essential in expressing the limit of resolution of the microscope. As illustrated in Figure 23.1, n is the refractive index of the medium in which the specimen, S, is immersed; θ is the half angle of the cone of incident rays; and the numerical aperture (N.A.) of the cone of rays is n sin θ. In a compound microscope, a glass cover slip, and in the case of immersion objectives a layer of immersion fluid, intervenes between the specimen and the entrant surface of the optical system. In this case, the numerical aperture becomes the product of the lesser index of refraction and the sine of the angle θ in that medium. The limit of resolution of the compound microscope, i.e., the least distance between two objects that can be seen as separate, is equal to the wavelength of light (λ) divided by the sum of the numerical apertures of the substage condenser and the microscope's objective lens.

23.3 COMPONENTS OF A COMPOUND MICROSCOPE

23.3.1 General. In order to realize the requirements for microscopic observations, the simplest form of an optical system is shown schematically in Figure 23.2. Figure 23.3 illustrates how these optical components (except the light source A and the lamp condenser B of Figure 23.2) are incorporated into a modern compound microscope. In Figure 23.3, the mirror below the substage condenser and the reflecting prism in the body between the objective and eyepiece are for mechanical convenience and are of no optical significance. The initial optical element of many compound microscopes is a mirror, which reflects light from the source into the remainder of the optical system. The mirror usually presents no design problems. One side is flat and reflects the light into the substage condenser. However, when extremely low powered objectives are used, the substage condenser usually will not illuminate the entire field of view because of its increased size. In this instance, the substage condenser may be removed, and the second side of the mirror, which is concave, will condense the light onto the specimen.

23.3.2 Illumination systems.

23.3.2.1 Simple illuminator. The simplest form of illumination for a compound microscope is a broad diffusing source, such as a ground glass, placed in front of an incandescent bulb. This source is imaged directly onto the specimen by means of the substage condenser (Figures 23.2 and 23.3). However, any granules in the ground glass would be visible in the field of view unless the image of the ground glass was slightly defocussed. Therefore, such a source is satisfactory only for low power microscopy, because of its lack of illumination for high

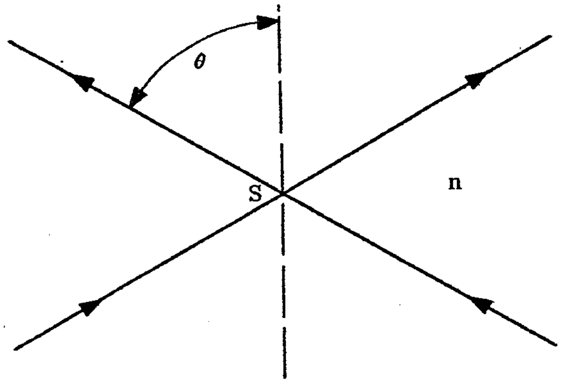


Figure 23.1- Determination of numerical aperture.

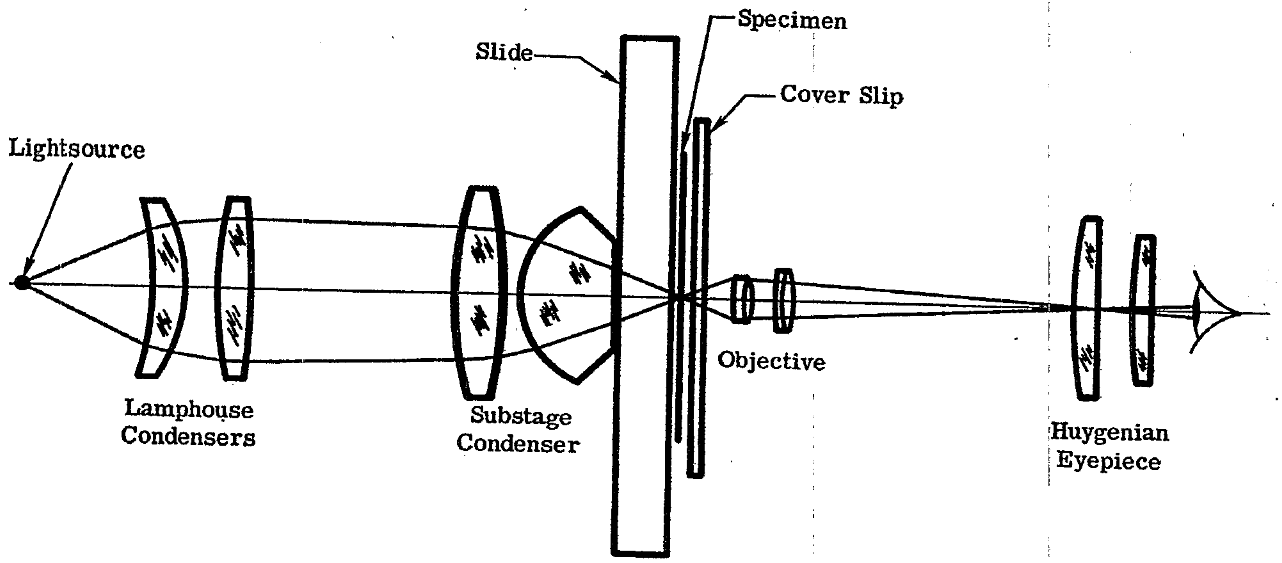


Figure 23. 2- Optical schematic of a compound microscope.

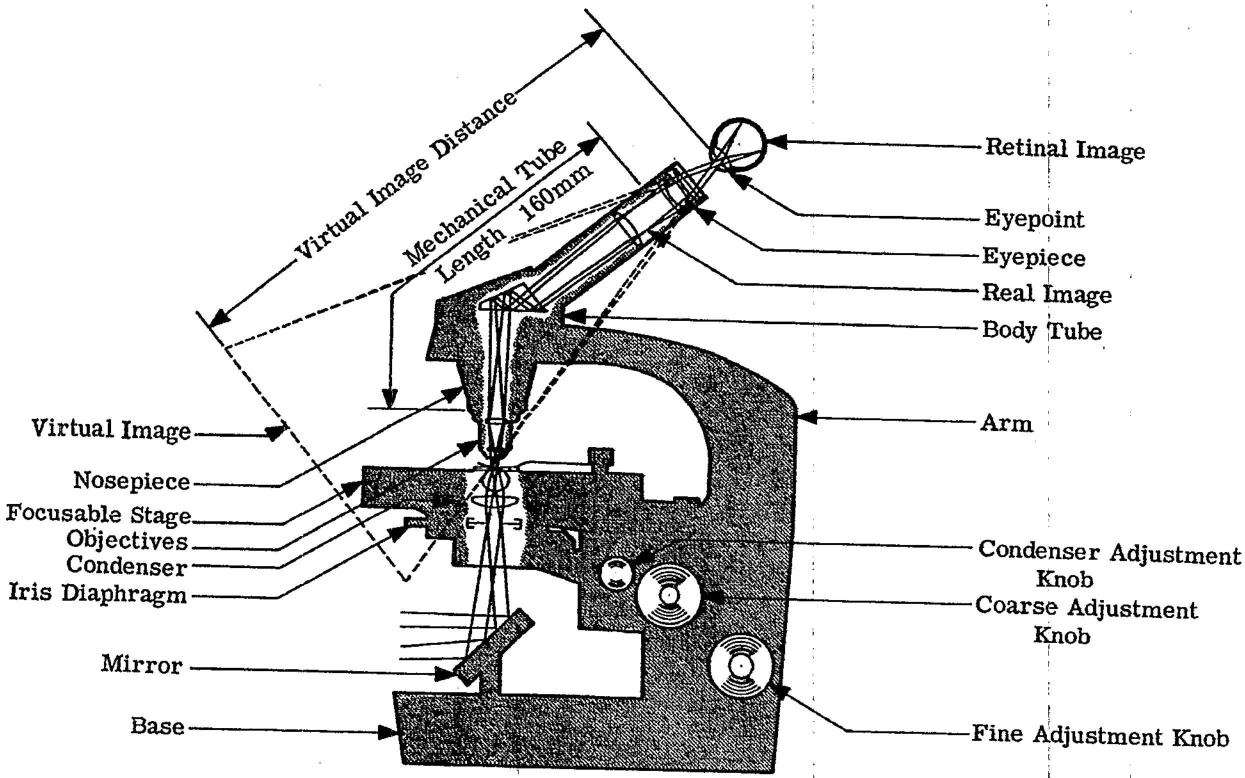


Figure 23. 3- Optical and mechanical features of the microscope.

power work. It should be noted that when the source of light is focussed directly onto the specimen, the illumination is designated as critical illumination. A more efficient microscope illuminator than that discussed here previously is shown in Figure 23.4 and consists of a monoplane filament lamp behind which is a spherical reflector and in front of which is a condenser--generally a two lens system. The filament of the lamp is near the center of curvature of the spherical reflector, and the reflected images of the strands of the filament are located between the strands themselves. This not only allows the reflected light to be utilized, but also forms a more nearly uniform primary source of light. The condenser in this system may be focussed so that the light source is imaged in the vicinity of a ground or opal glass, which in turn is focussed by the substage condenser onto the specimen (critical illumination).

23.3.2.2 From the preceeding paragraph it can be seen that critical illumination has the defect of not providing completely uniform illumination over the area of the specimen, and especially for photomicrography this is disadvantageous. A system known as Kohler illumination is used to overcome this difficulty. In this system, the lamp house condenser is used to focus the primary light source onto the substage iris diaphragm, placed at the front focal surface of the substage condenser, shown in Figure 23.4. Hence the light emerging from the substage condenser consists of parallel rays, which are re-imaged by the microscope's objective at its rear focal plane. The substage condenser now focuses the lamp house condenser, Figure 23.2, onto the specimen. Since the lamp house condenser is nearly uniformly illuminated by the light source, the field of the specimen is very uniformly illuminated. For cases in which the field must be uniformly illuminated, e.g., photomicrography, a form of Kohler illumination must be used unless the light source is very uniform as with a ribbon filament lamp.

23.3.2.3 Optical requirements for an illuminator.

23.3.2.3.1 The spherical mirror offers no design problem other than that of its aperture being large enough so that the reflected light will pass through the optical elements that follow.

23.3.2.3.2 The lamp house condenser should be large enough so that its image (formed by the substage condenser on the specimen--Kohler illumination) will fill the field of view. In addition, the focal length of the lamp house condenser should be correctly determined, in order that the particular primary light source will be large enough to fill the substage iris diaphragm.

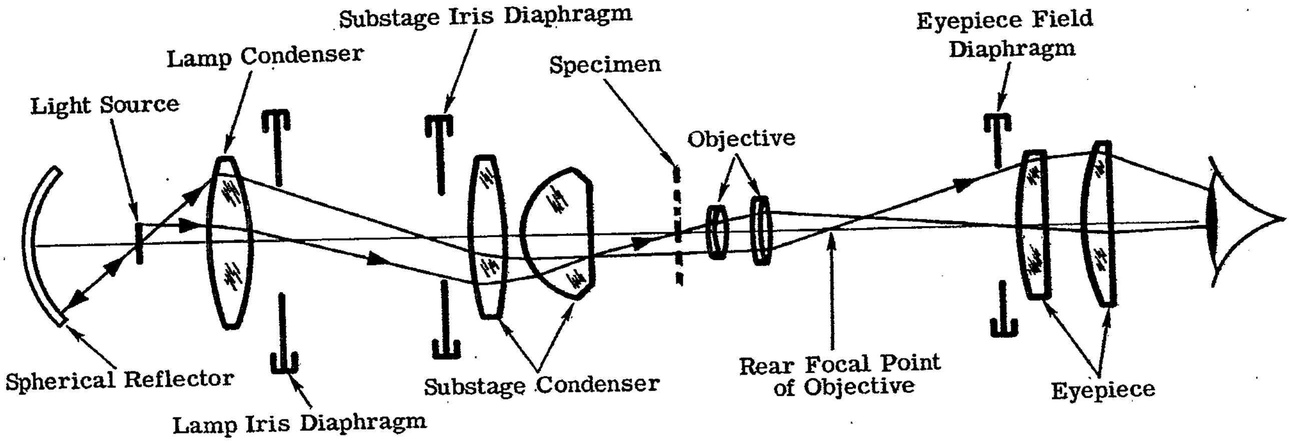


Figure 23.4- Kohler Illumination, schematic diagram.

23.3.2.3.3 An iris diaphragm is often located very close to the lamp condenser. With this design, the condenser (in Kohler illumination) and the iris diaphragm are imaged on the specimen, and if the iris diaphragm is adjustable in diameter, its image can be made to precisely fill the field. An adjustable iris diaphragm will prevent illumination of a greater area of the specimen than is necessary, and will also prevent scattered light, with its resultant loss of contrast, from entering the microscope.

23.3.2.3.4 The lamp condenser is usually a two lens, air-spaced system. It should be as well corrected for spherical aberration as is possible. The focal length must be correctly determined in order to image the filament of the lamp large enough to fill the iris diaphragm of the substage condenser (Kohler illumination) at a convenient distance (approximately 15 inches). The diameter of the condenser must be large enough so that its image, as formed by the substage condenser, covers the specimen field, when viewed with a 16mm focal length,(10x) microscope objective. It is readily apparent then, that the smaller the light source, the greater must be the speed of the condenser.

23.3.2.4 Vertical illuminators. For opaque specimens, vertical illumination is required for seeing surface details. Vertical illumination requires that the specimen field be uniformly and intensely illuminated, and that the illuminated field be limited to that portion of the specimen which is in the field of view. If the illuminated field is not limited as mentioned, an undesirable amount of light is scattered by the unviewed portion of the specimen, by the edges of the objective lenses, or by the walls of the objective. This scattering will reduce the contrast in the visual field.

23.3.2.4.1 Vertical illuminator, type A. In this type of vertical illuminator, as shown in Figure 23.5, the incident light is focussed on the specimen by being passed through the objective, in a reverse direction, and onto the specimen. The light source in this type is usually a low voltage, concentrated,filament bulb, located in a housing extending laterally to the axis of the microscope. The system consists of the light source, a condenser with an iris diaphragm mounted near the light source, a second condenser, and a semi-reflector at 45° to the microscope's axis for throwing light into the rear of the microscope's objective. The two condensers image the lamp filament at the exit pupil of the objective. The second condenser images the iris diaphragm (and the first condenser) at a virtual distance of about 160mm from the microscope's objective. Therefore, an image of the iris is formed by the microscope's objective on the specimen and it is uniformly illuminated. The field covered by the illuminated spot on the specimen can be regulated in size by adjusting the diameter of the field iris diaphragm, thereby preventing the scattering of light.

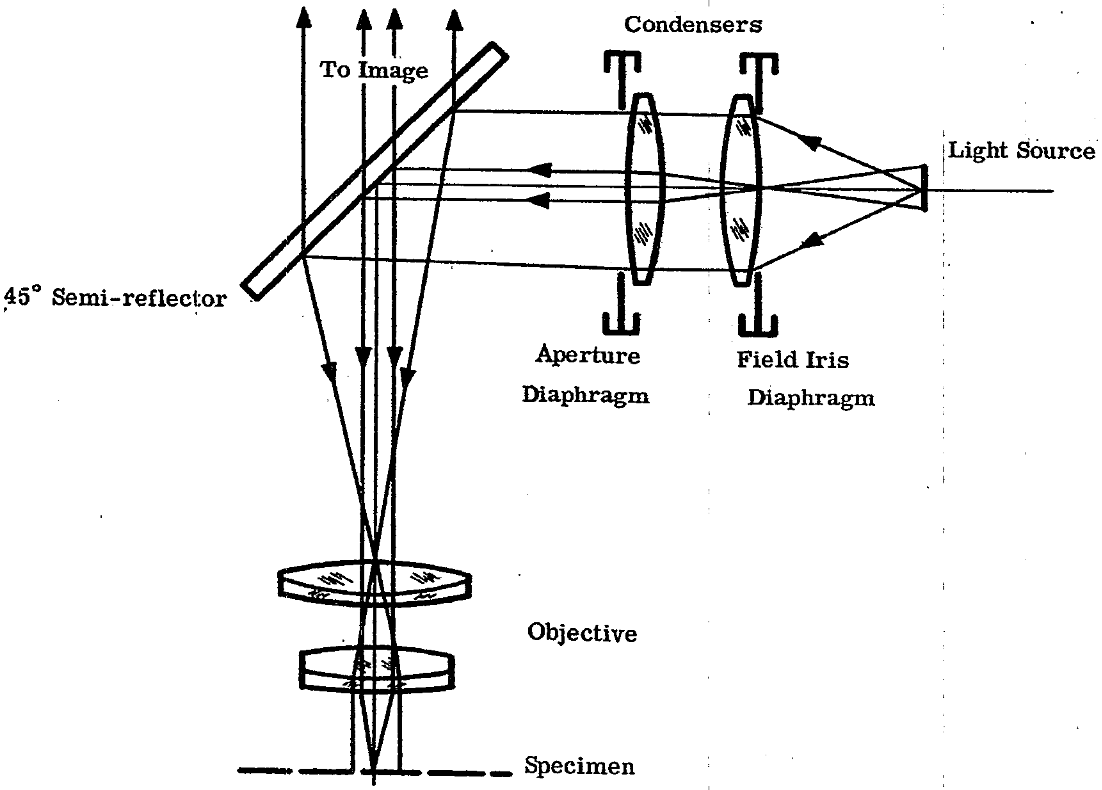


Figure 23.5- Vertical illumination.

23.3.2.4.2 Vertical illuminator, type B. In this type of illuminator, the light is not sent backwards through the objective, thereby preventing damaging internal reflections. This type of illuminator is best explained by picturing a condenser with a cylindrical hole through it axially, and the microscope's objective extending through this hole. In the actual design, the condensing elements surround the microscope's objective, and the illuminating cone of light is completely insulated from the objective's optics. Light is introduced into the condenser from the side through the use of an annular shaped full reflector. Such systems are called epi-condensers. Since the light must clear the lower rim of the microscope's objective and impinge on the specimen, it is obvious that the working distance from the objective to the specimen must be kept fairly large, therefore, this method of illumination is limited to the lower powers and numerical apertures of the objectives.

23.3.3 Substage condensers. The substage condenser concentrates light onto the specimen at a numerical aperture equal to that of the objective. There are two general classifications of condensers: the non-achromatic and the achromatic.

23.3.3.1 Non-achromatic. The non-achromatic substage condenser is used with achromatic microscope objectives which will be discussed in paragraph 23.3.4.3

23.3.3.1.1 The most common non-achromatic substage condenser is the Abbe condenser with a numerical aperture of 1.25. It is a two lens, air-spaced system, with a double convex lower lens and a plano-convex, hyperhemispherical upper element. When immersion fluids are used, the working distance between the upper plane surface and the image of the light source is equal to the optical thickness of the microscope's slide, plus a small space of a few tenths of a millimeter between the upper surface of the condenser and the slide. Reference to Figure 23.6 shows that no attempt is made to correct the chromatic aberration of the Abbe condenser, since both elements have positive power. The only variables available to the optical designer are the refractive indices of the elements, the shape of the first lens, and the separation of the two elements. The designer must adhere to a given working distance from the upper plano surface to the upper surface of the microscope slide, so that the light source (usually at a distance from infinity to a foot) can be focussed onto the specimen. In addition, the required numerical aperture is an additional consideration. A third requirement is that the focal length be such that the light source is imaged at a size sufficient to cover a specified area of the specimen, when the size and distance of the light source have been specified. (In practice, the light source is at least two inches in diameter at a minimum distance of one foot, and the field to be covered is that of a 16mm focus microscope objective). The requirements of numerical aperture and field coverage will determine the focal

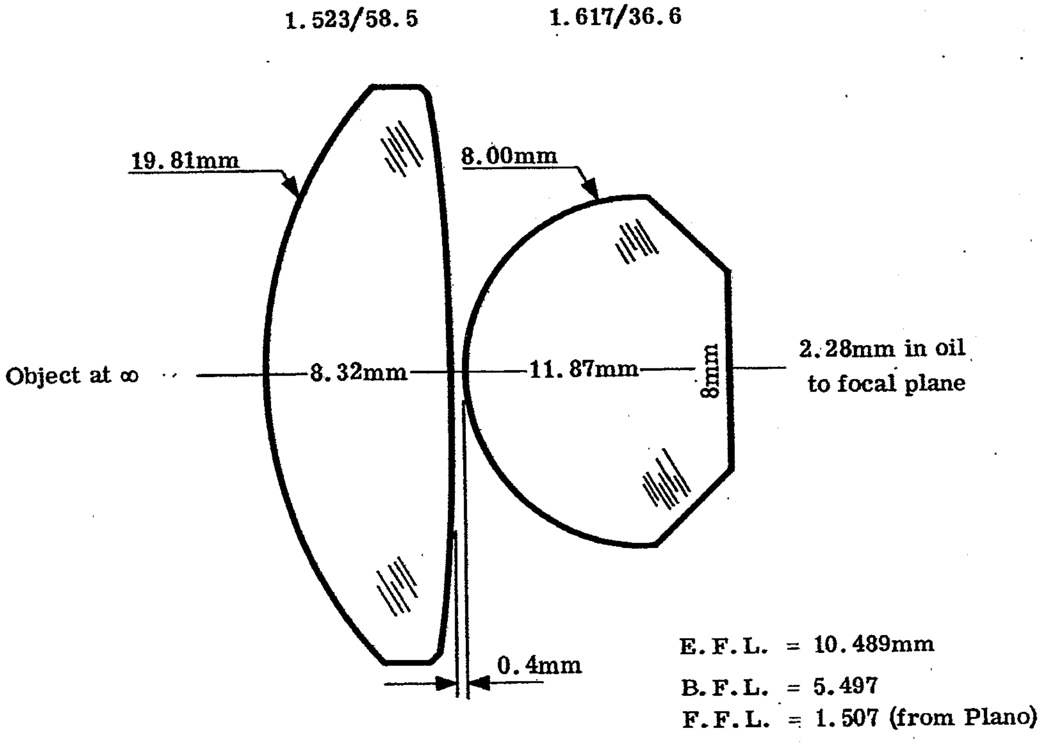


Figure 23.6- Optical layout of a 1.25 NA Abbe Condenser.

length of the front or first lens. The only aberration the designer may control in an Abbe condenser is spherical aberration. In the initial design of an Abbe condenser, graphical methods are useful and they are usually followed by mathematically triangulating a set of axial meridional rays through the system. In this way, the back lens may be bent to correct for spherical aberration. In conclusion, the two lens Abbe condenser is most commonly used with achromatic, rather than apochromatic objectives. For the latter, an achromatic condenser is used which more nearly approaches a microscope objective in form, construction, and correction. As the design principles of these condensers so closely approximates those of objectives, the reader is referred to paragraph 23.3.4.

23.3.3.2 Achromatic. Achromatic condensers are more complex than the Abbe condenser, and may consist of a triplet, doublet, meniscus, and front lens. This construction affords an opportunity for the correction of spherical aberration, coma, and chromatic aberration. This design will be seen to be essentially that of a microscope objective having the same numerical aperture, namely 1.30 and 1.40. Microscope objectives will be discussed in paragraph 23.3.4.3 through 23.3.4.5. Achromatic condensers find their most useful application when used in combination with apochromatic objectives.

23.3.4 Objectives.

23.3.4.1 Classification of objectives. Microscope objectives are classified as achromatic, semi-apochromatic, and apochromatic. If a microscope objective has been designed to correct for spherical aberration for one color of the spectrum, and for axial chromatic aberration for two colors, it is classified as an achromatic microscope objective. If the objective has been designed to correct for spherical aberration for two colors, and the axial chromatic aberration for three colors, it is classified as a apochromatic microscope objective. If the objective has been designed for correction between these two extremes, it is classified as a semi-apochromatic microscope objective.

23.3.4.2 Reasons for classification. With the magnification and resolving power (therefore the numerical aperture) corrected, the designer must take into account additional factors. Since the microscope is an instrument of almost fixed image distance, the magnification of the objective is almost proportional to its focal length. The image distance of the objective is not quite constant, since the corresponding fixed distance in the microscope is the mechanical distance from the mechanical shoulder of the objective, where it makes contact with the nosepiece, to the mechanical shoulder of the eyepiece, where it in turn makes contact with the upper end of the body tube. When the body tube of the microscope contains prisms or other optics for special purposes, the preceeding statement is no longer applicable. When the body tube contains prisms or other optics for special purposes, the optical tube length can be made the same as that of a microscope not having optical elements between the objective and the eyepiece, through the use of auxiliary compensating lenses. The distance between the front principal point of the objective and the specimen must be the same for a series of objectives, if these objectives are to be parfocal, i.e., no shift in focus should be required as a change of objectives is made. In order to meet this condition, the distance from the second principal point of the objective to the mechanical mounting shoulder, is certain to be different with the different powers of objectives, so that a really constant image distance cannot be obtained. In general, the numerical aperture of a microscope objective must be increased with magnification. However, since the difficulty of correcting aberrations increases rapidly with an increase in numerical aperture, the complexity of construction of the objective also increases. A cemented doublet is satisfactory for numerical apertures below 0.25, and focal lengths of 32 and 48mm. A 16mm focus having a numerical aperture of 0.25 requires two cemented doublets in the system.

23.3.4.3 Achromatic. The cost of microscope objectives depends on the complexity of their construction, and the cost of the optical materials used. These factors increase with the numerical aperture of the objective and the degree of correction required. For most routine examinations of biological or industrial materials, the moderate corrections and construction of the achromatic type objective are sufficient. This class of objective for higher powers are constructed of the following: a hemispherical or hyper-hemispherical lens known as the front; followed by a meniscus, the second front; a cemented component, the middle; and a cemented component, the back. Some of these components can be omitted from the lower numerical apertures. For example, a 16mm focus, 0.25 numerical aperture objective achromat has two cemented doublets; an 8mm focus, 0.50 numerical aperture objective achromat has a front, no second front, but a middle and back; a 4mm focus, 0.66 numerical aperture objective achromat has a front, second front, a middle, and a back. A 1.8mm focus, 1.25 numerical aperture achromat objective has a front, second front, cemented doublet middle, and a cemented doublet back.

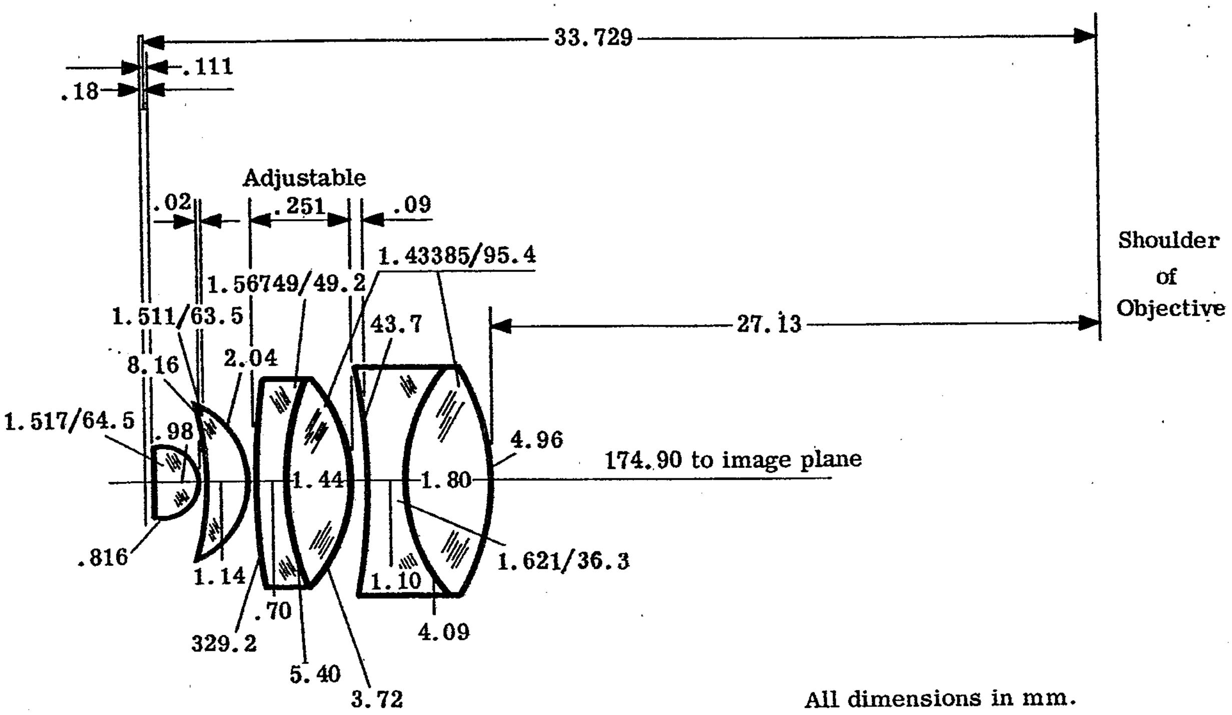
23.3.4.4 Semi-apochromatic. For more exacting routine microscopy, and for some kinds of research work, a higher degree of definition than that afforded by the achromatic objective is desirable. Semi-apochromats usually has flourite for one of its elements. Because flourite has a low refractive index, low dispersion, and a partial dispersion ratio different from glass, a better simultaneous correction for primary and secondary chromatic aberration and spherical aberration can be accomplished by its use as a positive element in a lens system. For example, if a flourite positive element is used with a flint glass negative element, a steep interface between the elements is attained, when the chromatic aberration is corrected. The over correction for spherical aberration, resulting from the steep interface and the large refractive difference at it, can be used to compensate for the under correction of other elements. By virtue of flourite's partial dispersion ratio being out of line with that of glass secondary chromatic aberration is favorably influenced. Constructional data for a semi-apochromat is shown in Figure 23.7

23.3.4.5 Apochromats. Apochromats are the most highly corrected of any of the microscope objectives. The optical design considerations involved are those described in 23.3.4.4, but they must be carried to the highest possible state of perfection. The correction is accomplished by the addition of optical components in the middle and back sections of the objective, and by the use of such crystals as alum and flourite to accomplish simultaneous correction for color, coma, and spherical aberration. Flint glass of the shortened spectrum type is also used in some constructions.

23.3.5 Eyepieces. There are three important considerations in the design of a microscope eyepiece. Since the object for the eyepiece is the image formed by the other elements of the optical system, the eyepiece can be designed to correct some of the residual defects in the other elements of the microscope's optical system. Also, the design of the eyepiece must be such, that a virtual image is formed anywhere between the point of most distinct vision (approximately 10 inches distant) and infinity. Finally, the eyepiece must be designed to correct lateral chromatic aberration.

23.3.5.1 Types of eyepieces. The main types of eyepieces used in compound microscopes are the Huygenian and Ramsden, and the type known as compensating eyepieces.

23.3.5.2 Huygenian eyepiece. For observational purposes, the Huygenian is often preferred to other eyepieces, since it can be completely freed of lateral color. The Huygenian eyepiece consists of two plano-convex lenses of the same type of glass (usually spectacle crown), with the field lens having a focal length approximately three times that of the eyelens, depending on the type of correction desired. The field lens and eyelens are separated in the body tube by a distance equal to twice the focal length of the eyelens. The combination is, therefore, free of lateral chromatism, and is most widely used with achromatic objectives (see paragraph 23.3.4). As is the case with all eyepieces, the limiting aperture for image forming bundles of rays is the exit pupil of the entire optical system of the microscope. The exit pupil is generally close to the second focal point of the eyepiece, and if a field stop or diaphragm is used, it should be positioned at the first focal point of the eyelens in order that its image will be formed at infinity. To some extent in microscopy, reticles are provided, and it follows that these should also lie in the plane of the first focal point of the eyelens. However, when such is the case the reticle is magnified by the eyelens alone, and even though the eyepiece combination as a whole



All dimensions in mm.

Figure 23.7- Optical layout of a 1.8mm flourite objective.

is free of lateral chromatic aberration, the corrections provided by the field lens are lacking and a large amount of aberration, particularly distortion and lateral color are introduced. To overcome this difficulty, the reticle used is kept so small that it is seen only at the center of the field. With respect to residual aberrations, the Huygenian eyepiece shows some spherical aberration, a large amount of longitudinal color, and marked pincushion distortion. An additional disadvantage occurs when this type of eyepiece has focal lengths less than one inch, since the eye relief is then usually too short for comfort. The reader is referenced to paragraph 6.11 where it is shown that a lens system such as this, can be designed to have constant equivalent focal lengths for all colors. An illustration and aberration graph of a Huygenian eyepiece is shown in Figure 23.8.

23.3.5.3 Ramsden eyepiece. A second type of eyepiece occasionally used with the microscope is the Ramsden eyepiece. The Ramsden eyepiece consists of two plano-convex lenses of the same type of glass (usually ordinary crown glass) and with equal focal lengths. The lenses are separated by a distance equal to two-thirds of the focal length of a single element. The focal point of this combination lies outside the system and so the eyepiece can be used to focus on an external reticle or cross hairs. With respect to aberration, the Ramsden eyepiece has more lateral color than the Huygenian, but the longitudinal color is only about half as great. The Ramsden eyepiece has about one-fifth the spherical aberration, and approximately half the distortion as found in the Huygenian eyepiece. The Ramsden eyepiece evidences no coma, and important advantage over the Huygenian is its 50 percent greater eye relief. An illustration of a Ramsden eyepiece is shown in Figure 23.9, and it is designated by usage as a positive eyepiece, in contradistinction to the negative Huygenian type.

23.3.5.4 Compensating eyepiece. The compensating eyepiece is used in conjunction with apochromatic objectives (paragraph 23.3.4.5) and as was previously stated, transverse chromatic aberration is a characteristic of these objectives. In order to correct for this aberration, an equal and opposite amount is introduced by the eyepiece. The eyepiece compensates the lateral color of the objective, and derives its name from this property. In addition to a definite amount of lateral color, the design of the eyepiece must correct for coma, spherical aberration and axial color, and its curvature of field and astigmatism must compensate those of the objective in so far as possible. In some cases, the observer wears spectacles, especially when the ocular defect is astigmatism (Myopia or hyperopia can be compensated by simply focussing the microscope), and it is therefore desirable to have the eyepoint of the microscope high enough so that there is sufficient space for the spectacle lenses between the back lens of the eyepiece and the eyepoint. This requires the back focal length of the eyepiece be sufficiently large in relation to the equivalent focal length, which determines the magnification. Such

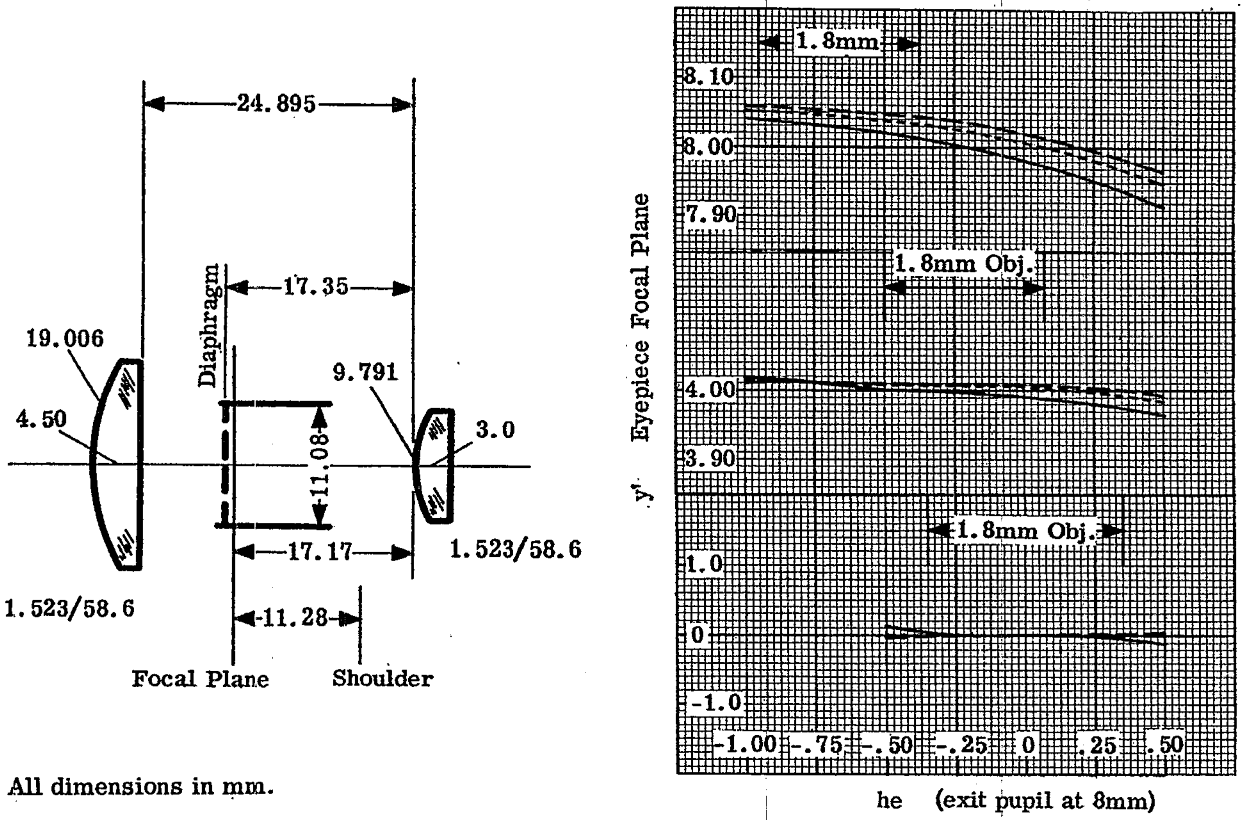


Figure 23.8- Optical layout of a 10X Huygenian eyepiece.

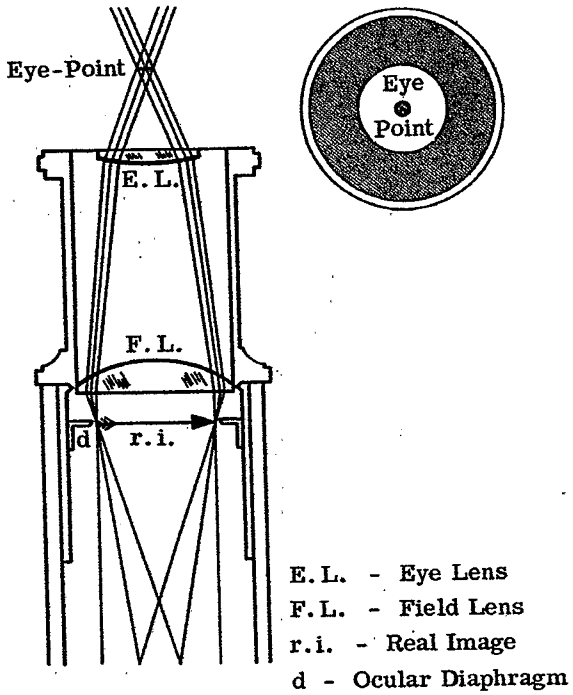


Figure 23.9- Optical schematic of a Ramsden eyepiece.

eyepieces are designated as "High-Eyepoint" and are illustrated in Figure 23.10. Compensating eyepieces can be of the positive or negative construction. Figure 23.11 shows the general construction of several powers of compensating eyepieces. The design of such eyepieces is dependent on the residual aberrations of the apochromatic objectives with which they are to be used. As shown in A and B of Figure 23.11, these eyepieces are of the negative type and are evolved from the Huygenian eyepiece by making the field lens and/or the eyelens cemented doublets, for purposes of correction. The high eyepoint compens is of the positive type and may be considered to be derived from the Ramsden eyepiece, by making the field lens a cemented triplet. The 30x compensator shown is essentially a ratioed form of the 10x High Eyepoint eyepiece. This construction prevents the eye distance from becoming too small, although the equivalent focal length necessarily is small in order to give the required relatively high eyepiece magnification.

23.4 DARKFIELD MICROSCOPY

23.4.1 General. In the ordinary microscope discussed previously, the illuminating bundles of rays enter the microscope objective and illuminate the entire field of view. The objects under examination are imaged as dark or colored details appearing against a bright background. Therefore, by this usual method of illumination, Brightfield Microscopy is accomplished. If the specimen is small, as for example with colloidal particles, or is practically transparent, the ordinary brightfield microscope does not offer sufficient contrast to render the objects visible. However, such particles have the property of scattering a portion of the incident radiation by means of diffraction, refraction, or reflection. In the field of darkfield microscopy, only the scattered light enters the microscope, while the direct illuminating beam entirely escapes the microscope's objective. Darkfield microscopy is accomplished by using the condenser to block the central portion of the light cone. The blocking of the entering light may be accomplished as detailed in 23.4.2 through 23.4.5. In both darkfield microscopy and ultra microscopy (paragraph 23.5) the objects appear to be self-luminous in a darkfield, and no light directly reaches the observer from an outside source. Light is only transmitted to the observer from the object being viewed.

23.4.2 Refracting darkfield condenser. A simple refracting darkfield condenser is an ordinary substage condenser provided with an opaque center stop which allows only rays traversing the outer zones of the condenser to be transmitted as shown in Figure 23.12. The effective numerical aperture of the microscope's objective must not be greater than the numerical aperture of the obscured central portion of the condenser, in order that the oblique hollow cone of rays transmitted by the condenser will not directly enter the objective. The oblique hollow cone of light will illuminate any object at its apex or focus, the object itself then deflects a part of this

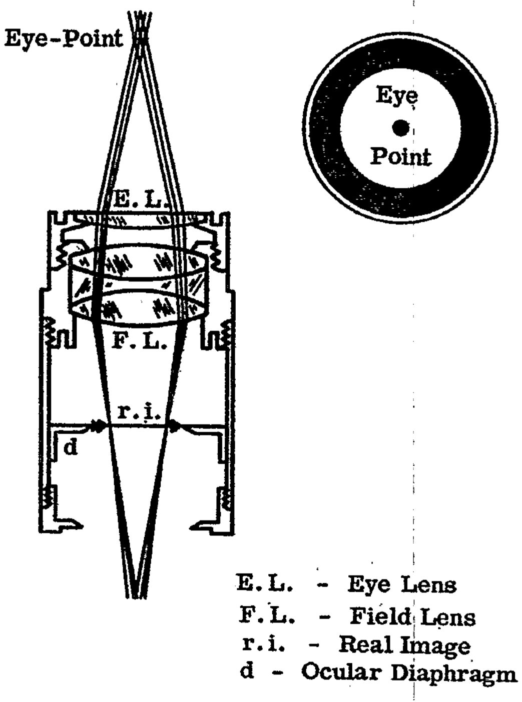


Figure 23.10- Optical schematic of a high eye-point compensating eyepiece.

light into the microscope, and the object will then seem to be self-luminous in a dark field. The smaller numerical aperture of the illuminating bundle is about 0.7, while the upper limiting numerical aperture of the condenser is about 1.2. While such an illuminator is suitable for non-critical work, the refracting condenser has too much spherical and chromatic aberration for exacting darkfield use. In order to obtain a sufficiently dark background, it is important to have a very thin section of the specimen receive the focussed light. This condition will preclude any significant amount of aberration being present in the condenser. Darkfield condensers of the reflecting type may be well corrected for those defects and are generally used for high power work.

23.4.3 Reflecting darkfield condensers. The advantage offered by a reflecting darkfield condenser, with respect to the refracting type, is its ability to form a good ring of light for darkfield work, and its ability to minimize spherical and chromatic aberration in the transmitted bundle. More light will be scattered by the specimen if the difference between the inner and outer numerical apertures of this hollow cone is large. On the other hand, the microscope's objective is functioning at a numerical aperture not greater than the lesser numerical aperture of the hollow cone. This factor determines the amount of scattered light which can be used for image formation, and also determines the resolving power of the microscope. It is common practice to have the numerical apertures of the hollow cone cover the image from 0.7 to approximately 1.25. When objectives having numerical apertures greater than 0.7 are used, it is necessary to equip them with a funnel stop. The funnel stop will reduce the numerical aperture so that no direct light passes through the objective. For high power darkfield microscopy, not all the light can pass from the condenser to the specimen unless the specimen and its slide are in oil contact with the condenser. Some reflecting darkfield condensers are made with spherical surfaces or aspheric surfaces. Aspheric reflecting darkfield condensers are more difficult to fabricate, but are theoretically better corrected than the spherical type.

23.4.4 Aspheric darkfield condensers.

23.4.4.1 Paraboloid. A paraboloidal darkfield condenser is shown in Figure 23.13 (a). This condenser is a plano-convex block of glass with the reflecting surfaces forming a true parabola, at whose focus the specimen is positioned. Since the microscope's slide is in oil contact with the upper surface of the condenser, no aberrations are introduced.

23.4.4.2 Cardioid. In the cardioid darkfield condenser, the light rays undergo two reflections; one from the inner surface which is spherical, and one from the outer surface, which is cardioidal as shown in Figure 23.13 (b). This condenser, as is the case with the paraboloidal type, is free from chromatic and spherical aberration and, since it obeys the sine condition, is termed aplanatic. It is possible to observe particles as small as

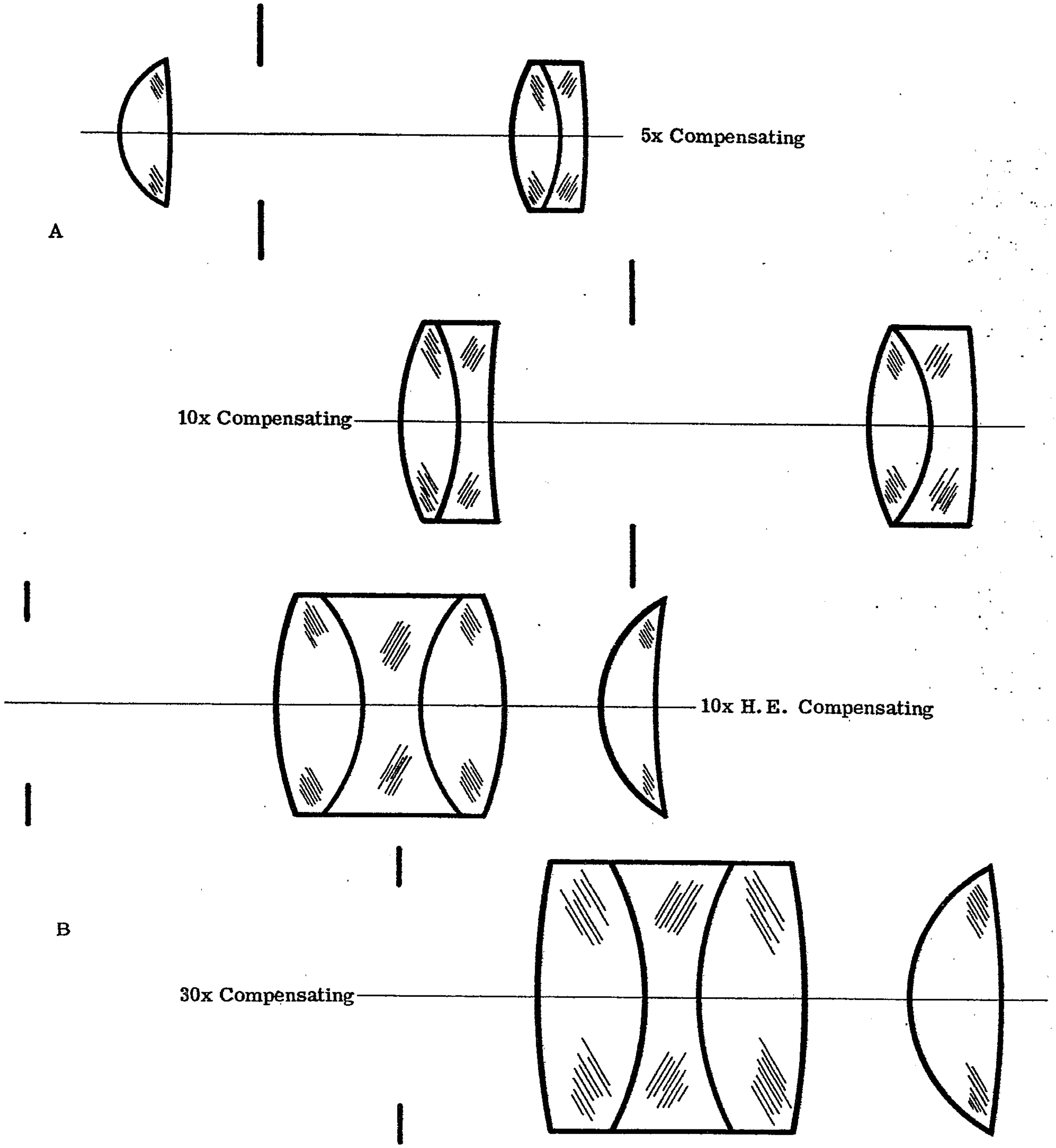


Figure 23.11- Typical compensating eyepieces.

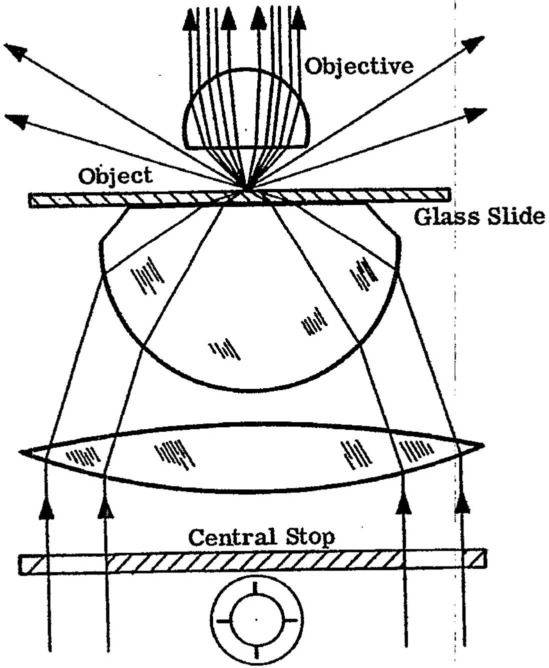


Figure 23.12- Refracting condenser with a central stop for dark-field illumination..

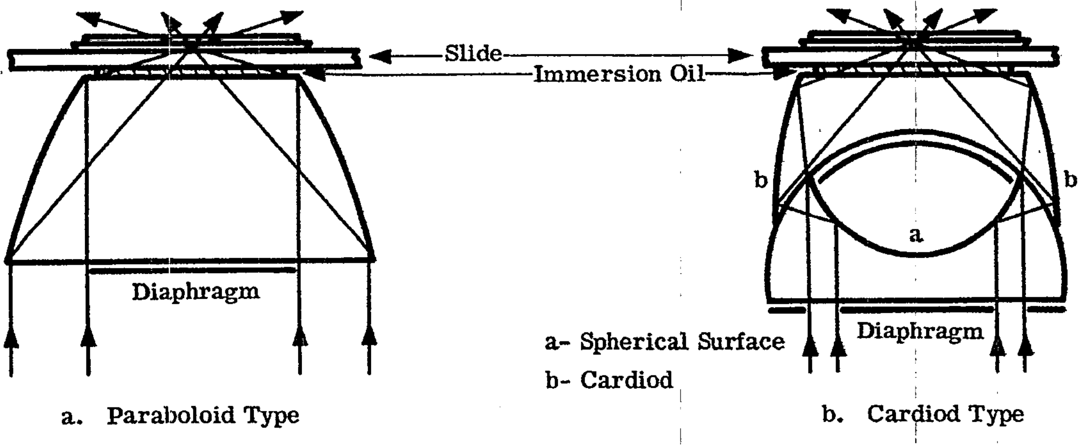


Figure 23.13- Aspheric darkfield condensers.

0.000004mm in diameter under favorable conditions with this type condenser. The disadvantage of this type of condenser is the difficulty encountered in grinding and polishing a precise cardioidal surface.

23.4.5 Spherical darkfield condensers.

23.4.5.1 Bispheric. The bispheric darkfield condenser as shown in Figure 23.14 is constructed with both surfaces spherical, thereby avoiding the difficulty of precise grinding and polishing (as is the case with the cardioid type). The highly precise spherical surfaces can then be used with only slight deviations from theoretical considerations.

23.5 ULTRAMICROSCOPY

23.5.1 General. As indicated in the conclusion of paragraph 23.4.1, darkfield microscopy and ultramicroscopy are similar in their approach to the problem of studying objects or specimens. The two approaches differ only in the size of the object to be observed. Darkfield microscopes deal with objects of approximately 0.2μ or more in diameter, that is, those which come within the resolving power of the microscope. Ultramicroscopy deals with objects so small that the details cannot be resolved, but the presence of the object is inferred by the presence of light which the object transmits in the instrument. Some of the details of the specimen viewed with the darkfield microscope can be resolved, but some details are so small that they show simply as points of light, usually in the form of so-called diffraction discs. The larger details in the specimen come within the province of darkfield microscopy, while the smaller details are the concern of ultramicroscopy.

23.5.2 Characteristics.

23.5.2.1 Ultramicroscopes pass a narrow beam of light through the specimen at right angles to the axis of the viewing microscope. With a strong light source, such as a carbon arc, ultramicroscopes are excellent for viewing and counting particles in colloidal suspension. Figure 23.15 illustrates the essential components of a slit ultramicroscope. The arc (a) is imaged by the lens (b) on the cross slits (c). The cross slits (c) are imaged by a long working distance,microscope objective (d) into cell (e), which is provided with two windows. The object to be viewed is introduced into the cell (e) and viewed by the microscope (f) through the upper window of the cell (e). In the cell, the Tyndall beam can be clearly seen in the microscope. By means of an eye-

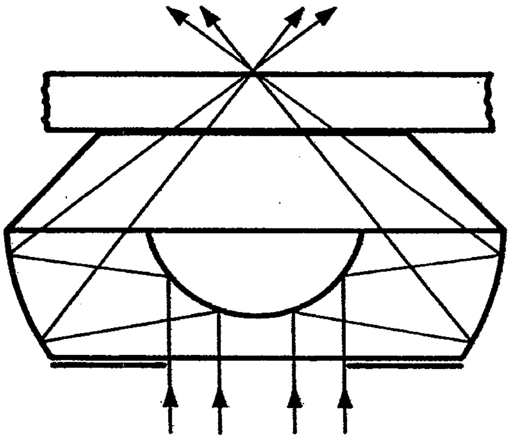


Figure 23.14- Bispheric (bicentric) darkfield condenser.

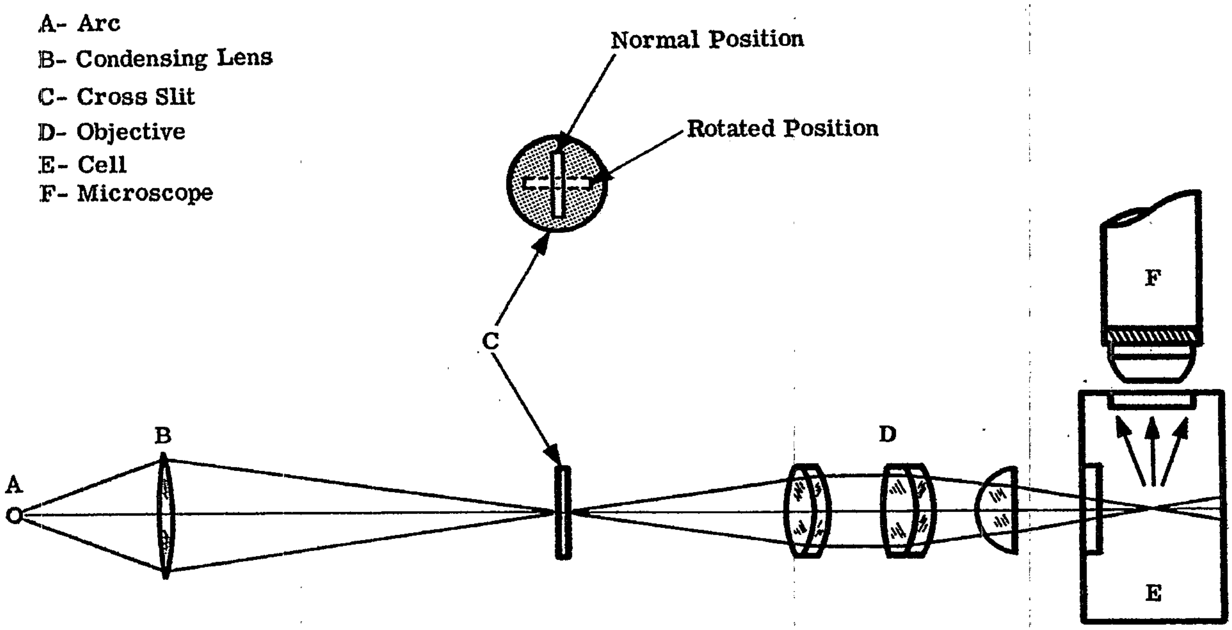


Figure 23.15- Elements of a slit ultramicroscope.

piece scale, the width and length of the beam can be measured. Therefore, if the cross slits (c) are rotated through 90°, the depth of the beam becomes the new width, and it can be measured. In this way, the volume of an illuminated portion of the contents of the cell (e) can be determined. In addition, the number of particles in the volume can be counted, and the number of colloidal particles per unit volume determined. Since the colloidal particles appear as diffraction discs, there is no need for high power or resolution in the viewing microscope.

23.6 PHASE MICROSCOPY

23.6.1 General. Nearly transparent materials having optical path (the product of the thickness and refractive index of the specimen) differences can be observed either with a phase or interference microscope. However, in contradistinction to the interference microscope (paragraph 23.7), the phase contrast is accomplished by the recombination in the image of direct light with the light deviated by the object after modification by a diffraction plate. It is interesting to note that a brightfield microscope may be converted to a phase microscope by the substitution of a phase condenser and phase objective. Similarly, the designer should keep in mind that a single contrast may be adequate for a given class of specimens, while other specimens may require several contrasts to reveal all of their structure.

23.6.2 Characteristics. The characteristics of a phase microscope can be seen from Figure 23.16. An annular diaphragm is placed in front of the condenser. When the annular diaphragm is uniformly illuminated, an image of it is formed in the objective near its focal plane, between the lens system. It can then be seen that all the light passes through this ring image or conjugate area when no specimen is present. However, when a specimen is being examined, some light is deviated through the rest of the area of the diffraction plate in the objective. The placing of a diffraction plate at this point differentially affects light deviated by the specimen, and the direct light from the background.

23.6.3 Principles.

23.6.3.1 The principles on which the phase microscope is based are shown in Figures 23.17 and 23.18. Figure 23.17shows a light wave A' passing through a transparent object C. You will notice that A' has slowed down with respect to light wave A, which did not pass through the transparent object, and accordingly the two light waves are out-of-phase. However, the human eye and light-sensitive photographic plates are insensitive to phase differences, and as a result the image can scarcely be seen or photographed. Light wave A" in Figure

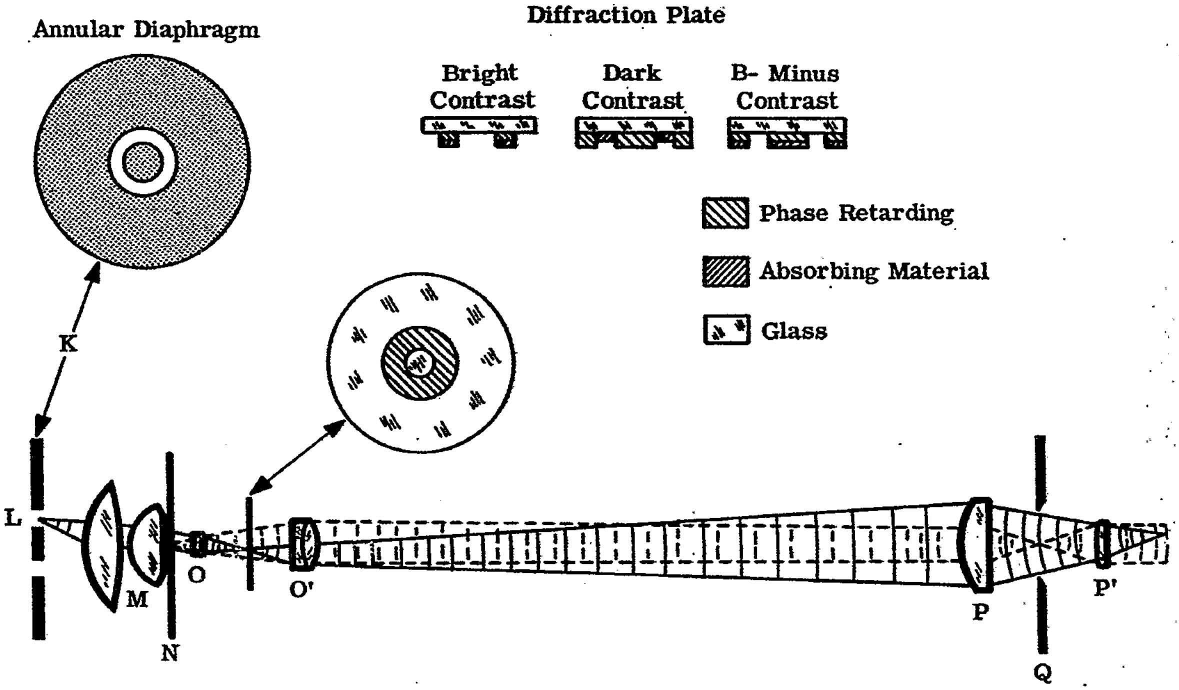


Figure 23.16- Elements of a phase microscope.

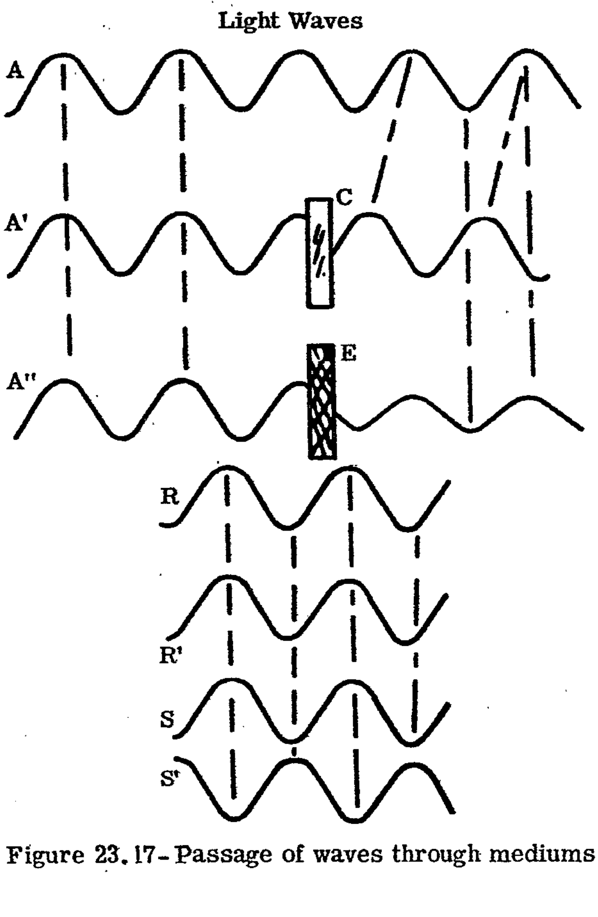


Figure 23.17- Passage of waves through mediums.

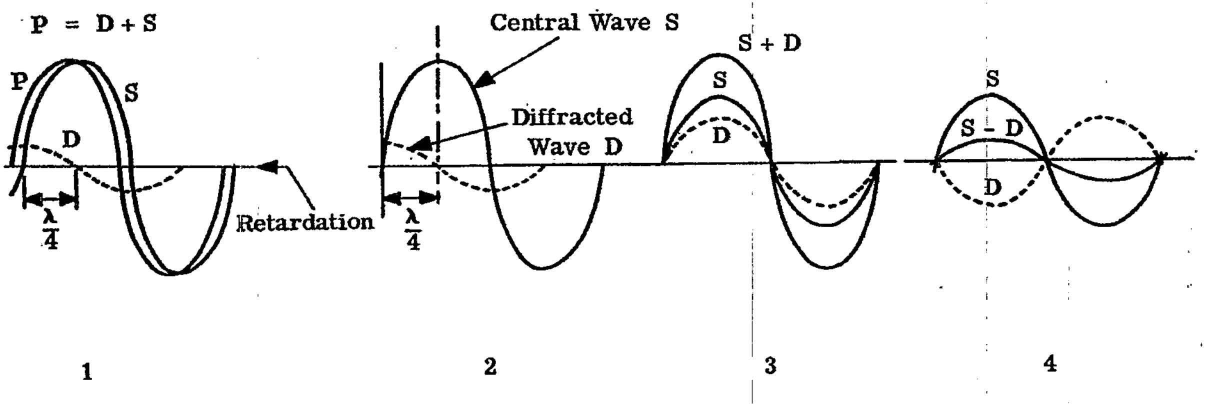


Figure 23.18 - Superimposed light waves.

23.17 passes through an absorbing medium E and is thereby reduced in amplitude as shown. However, in contradistinction to phase differences, amplitude differences are visible. When light waves of the same phase and amplitude are combined in the image as shown by R & R' in Figure 23.17, they add to produce brighter contrast. Similarly, dark contrast can be obtained by producing light waves which are out-of-phase or amplitude with each other as shown by S and S' in Figure 23.17, and combinations of amplitude and phase differences can be obtained which will produce lighter or darker greys.

23.6.3.2 Light waves may be superimposed as shown in Figure 23.18 Section 1 shows that the wave P, resulting from a slightly retarding particle, may be broken up into two waves S and D. The central wave S and the diffracted wave D, are shown again in Section 2. Section 3 demonstrates the result of using a bright contrast diffraction plate. The wave S has been partially absorbed, and wave D has been retarded, so that S and D are out-of-phase and produce a darker image.

23.6.3.3 The phase relationship of the light passing through a system of different optical paths has been altered, and the detail from the optical path differences, or slight absorptions of the specimen, will become visible within the microscope by the phase system elements (paragraph 23.6.2). By using an appropriate diffraction plate as previously discussed, it is then possible to increase or decrease the contrast of the image directly, or after reversing to change the contrast tone from bright to dark.

23.6.4 The diffraction plate. The diffraction plate consists of optical glass on which is evaporated,in vacuum, a very thin layer of metal, or a layer of a dielectric, or both. The layer of metal absorbs light, while the dielectric retards the light. The layer of metal or dielectric must be of sufficient size to cover either the image of the annular diaphragm formed in the objective or complimentary area of the remainder of the objective. These layers act upon the direct light from the background, and the deviated light from the specimen, so that recombination in the image will produce visible phase or absorption differences in the specimen.

23.6.4.1 The bright contrast diffraction plate absorbs, retards, or retards and absorbs the undeviated light, but has no effect on the deviated light. When this bright contrast diffraction plate is used, regions in the specimen of greater optical path will appear brighter than those of a lesser optical path.

23.6.4.2 The dark diffraction plate absorbs the undeviated light, and retards the light deviated by the specimen. The regions of greater optical path difference in the specimen will then appear darker. The effect of the dark diffraction plate is solely on the deviated light, and the degree of contrast is controlled by the width of the annulus and the thickness of the absorbing and retarding layer of the diffraction plate.

23.6.5 Disadvantages encountered with phase microscopy. As noted previously, phase contrast is accomplished in the phase microscope by the recombination of the direct and deviated light in the image, after diffraction. However, optical path differences and small absorption differences may be involved in this recombination and the resultant might be more appropriately termed "densiphase contrast" as suggested by Bennett, et. al. [1] Presently, phase microscopes have been modified to provide variable contrast, but not to measure the densiphase detail. Also, as the phase microscope redistributes the light in the image, haloes are often seen around the observed details, although the proper diffraction plate may lessen this condition. In addition, phase microscopes will make the optical path differences visible, but not their numerical magnitude.

23.7 INTERFERENCE MICROSCOPY

23.7.1 General. Interferometry, while well established in other fields, has only recently been applied to microscopy. Two methods of interference microscopy presently exist, the multiple beam method which is used extensively in the examination of surfaces of opaque materials having good reflection and in the examination of transparent materials, and the two beam method.

23.7.2 Characteristics.

23.7.2.1 Interference contrast. Interference contrast is accomplished by the recombination in the image of two beams of coherent light (from the same source), one of which is modified by passing through the specimen. In contradistinction to the phase microscope, the interference microscope will not produce haloes around the details. In addition, the interference microscope provides variable color contrast with white light illumination, and intensity variation in the color of the monochromatic light when monochromatic light is used. Similarly, with monochromatic light the interference microscope can provide measurement of the optical path differences in the specimen. It is interesting to note that when the thickness of the specimen is known, the refractive index can be measured, and in the case where the specimen is placed in a media of a different known refractive index, both the thickness and index can be measured. Also, interference microscopes have increased vertical resolution, but have the same lateral resolution as other light microscopes.

23.7.2.2 Multiple beam interference microscope. In the multiple beam method, the specimen to be examined is mounted between two flat,metalized,reflecting surfaces, and illuminated with parallel,monochromatic light. The recombinations resulting from repeated reflections of the light through the specimen produce fringes, which are used to measure the optical path differences (within reasonably transparent specimens).

23 7.2.3 Two beam interference microscope. With the two beam method, coherent illumination (from a single source) is so divided that part of the light passes in focus through the specimen, and the remainder passes to one side or is out-of-focus at the specimen. On recombining the light, the beams interfere to produce measureable patterns from which the optical path differences can be determined. This beam separation can be accomplished by reflection or polarization.

23.7.3 Principles.

23.7.3.1 The A O Baker interference microscope, Figure 23.19, illustrates the principles of interference microscopy. This microscope is fundamentally a polarizing microscope modified into a two-beam interferometer. The condenser has a birefringent plate which divides the light into two beams and the objective has a corresponding plate which recombines the beams after one of them has passed through the specimen. Above the objective is a quarter-wave compensator and an analyzer. Various eyepieces may be used to obtain different magnifications with the Shearing or Double Focus types of 10X, 40X and 100X objectives.

23.7.3.1.1 The polarizer below the condenser polarizes the light in a plane at 45° to the axis of the birefringent plate. The birefringent plate at the top of the condenser separates the polarized light into two beams which are plane-polarized at right angles to each other. One beam passes through the specimen, and the other passes to one side of the specimen in the Shearing system. In the Double Focus system one beam focuses at the specimen and the other spreads around the specimen to focus above it. The phase of the beam passing through the specimen is changed by the local variations in optical thickness in each portion of the specimen; while the changes in the reference beam depend on the average optical thickness of the specimen and the region around it in the Double Focus system; or the region to one side of the specimen in the Shearing system, as shown in Figure 23.20

[1] Bennett, A. H. Jupnik, H., Osterburg, H. and O. W. Richards, Phase Microscopy, pg. 11, John Wiley & Sons, New York, 1951.

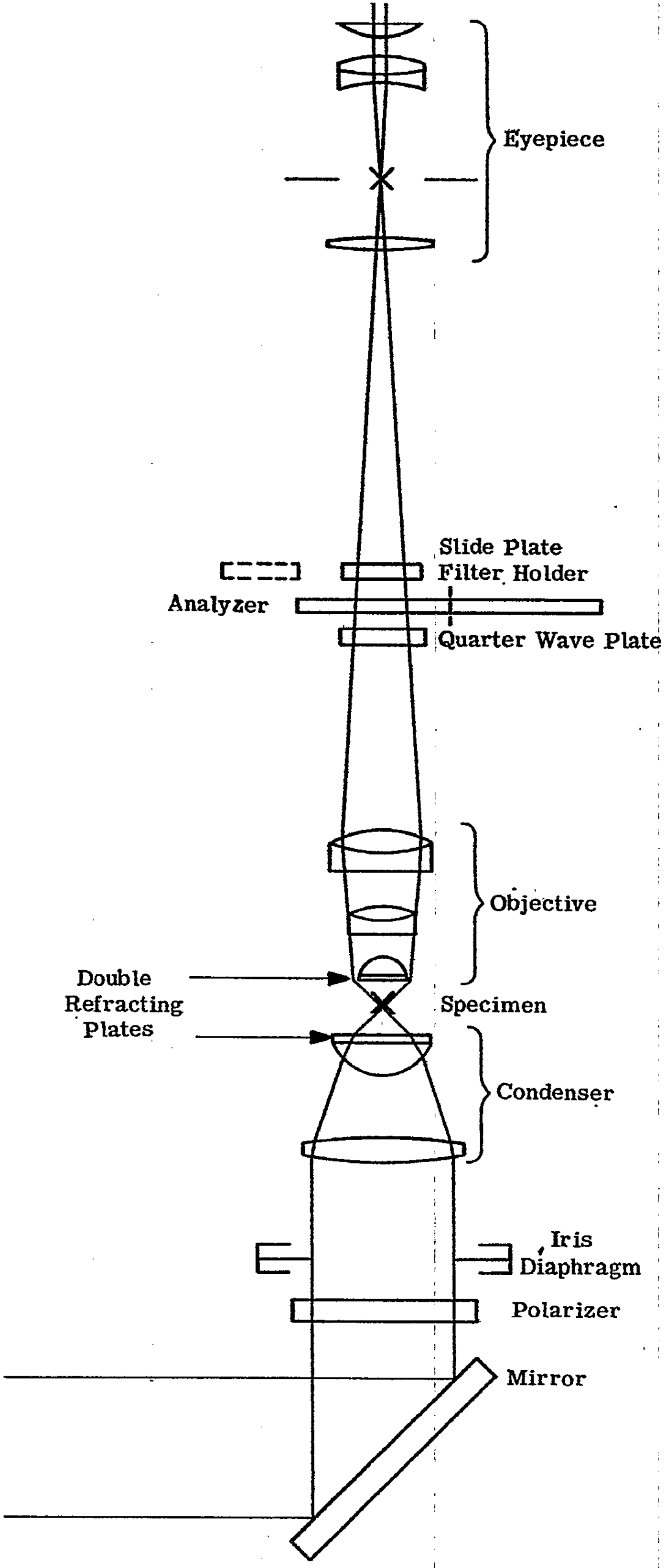
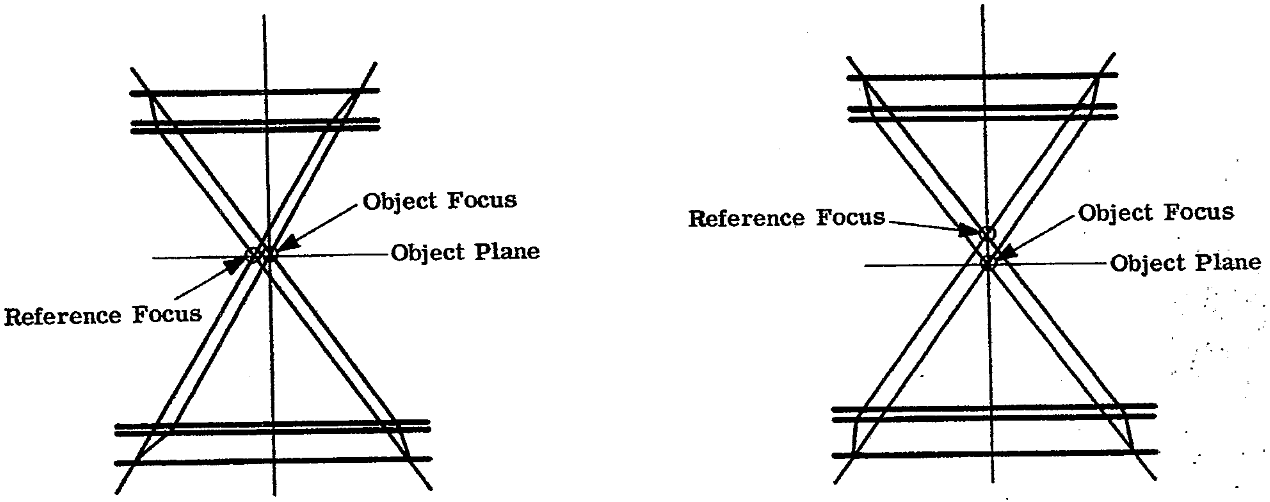


Figure 23.19- Optical schematic of AO Baker interference microscope.



Shearing System

Double-focus System

Figure 23.20- Path of light rays through a Shearing and Double-Focus system.

23.7.3.1.2 The birefringent plate on the front of the objective unites the two beams and the quarter-wave plate changes the two oppositely polarized beams into left and right-hand circularly polarized light. The resultant of two circularly polarized beams is plane polarized light, the direction of the plane depending on the phase difference between the circularly polarized beams. Thus the phase differences in the specimen can be determined by turning the analyzer to the position of minimum luminance, or extinction, in the image.

23.7.3.1.3 A vector theory (2,3) and an integration theory (4) have been proposed for the mathematical analysis of this type of microscope.

23.7.4 Interference colors.

23.7.4.1 Light passes through an optically denser medium more slowly than in a less dense medium and is retarded,with respect to light,through the less dense medium. The amount of retardation (phase difference) is proportional to the difference in refractive index for the particular wavelength considered.

23.7.4.2 For example, should the denser regions of a specimen illuminated with light from a mercury arc retard the blue light exactly one-half wavelength and the analyzer be set to extinguish the blue light, the specimen then would be seen only in the remaining yellow-green light.

23.7.4.3 With tungsten light,the blue is not limited to such a single wavelength as the mercury arc, but is a band of light (±440 - 490μ). A single particle cannot retard exactly to a half wavelength all of these blue wavelengths, therefore some blue will be lost and some transmitted and the particle will appear, more or less yellow,depending on the amount of blue lost.

23.7.4.4 Phase changes affect other colors in a similar manner and the actual interference colors depend on the composition of the light from the illuminator and on how the optical paths in the specimen retard or advance each wavelength (color). The relative amount of each wavelength passing through the analyzer determines the color of the particle.

[2] J. Roy, A Vector Theory of Phase Contrast and Interference Contrast. Micr. Soc. 75:23-37, 1955.

[3] G. Oster, A. W. Pollister, Physical Techniques in Biological Research 29-90 Vol. III, Academic Press, New York.

[4] ibid (3), 310-437, Vol. I.

23.7.4.5 Light is radiation to which the eye is sensitive (380-740Mμ) and can be seen. Interference microscopy is possible with invisible, infrared and ultraviolet radiation with receptors sensitive to the radiation used.

23.7.5 Illumination and filters.

23.7.5.1 Measurement of the optical path requires the use of monochromatic light to avoid the color interferences mentioned in paragraph 23.7.4. Monochromatic light is light of a single wavelength and is usually obtained from a single line of a spectral source. As the mercury arc has most of its radiation concentrated in a few lines it is the usual source for such illumination. The mercury arc has the further advantage in that the green line of 548mμ wavelength is quite close to the maximum sensitivity of the human eye (555mμ). Sodium light is suitable, although not as comfortable for visual use.

23.7.5.2 To isolate the light from a single line in the mercury spectrum a filter is used that transmits the light from the desired line and absorbs the light from the other bright lines in the spectrum of this source.

23.7.5.3 Some filters for use with mercury arcs are listed in Table 23.1. The least expensive is the Wratten 62 or 74, but these transmit only about 10% of the green light and do not exclude the light from the yellow line. The Corning CS4-120 transmits more of the green and no yellow. The Wratten 77 and 77A also transmit more green light than the 62 or 74, but for monochromatic light need to be combined with the 58 filter which reduces the light correspondingly. Filters for isolating the blue and yellow mercury lines are included in the table.

23.7.5.4 For some measurement, where the highest precision is not required, approximately monochromatic light is adequate and may be obtained with tungsten lamps and "narrow band" filters or with "interference" filters. The latter often have the disadvantage of low transmission.

23.7.5.5 The H85-C3 or H100-A4 (formerly AH3 and AH4) mercury arcs are satisfactory for many visual applications, but require long exposure when photomicrographs are to be made. More intense mercury arc sources such as the B-T-H 250 and the Osram HB0200 give more light, especially when monochromatic light is used, and are desirable for photography.

23.7.5.6 Light from the mercury arc without a color filter can be used for variable color contrast microscopy

Mercury line	Blue, 0.436μ		Green, 0.546μ		Yellow, 0.577μ	
Relative energy	80%		100%		88%	
Eye Relative luminosity (100 at 0.555μ)	1.8		98.4		89.8	
Filter	% Trans. #	Rel. Vis.*	% Trans. #	Rel. Vis.*	% Trans. #	Rel. Vis.*
Corning CS4-120	0	-	44	43	0	-
Corning CS-584	22	3	0	--	0	-
Ilford 625	0	-	35	34	8	6
Wratten 50	6.4	0.9	0	0	0	0
Wratten 62	0	-	10	9.8	0.5	0.4
Wratten 74	0	-	10	9.8	0.2	0.2
Wratten 77A	0	-	68	67	0	0
Wratten 77A 58	0	-	29	28	0	0
Wratten 77	0	-	74	73	0.5	0.4
Wratten 77 58	0	-	31	31	0.06	-
Wratten 58	0	-	42	31	11	9
Wratten 22	0	-	0	-	71	62

*Relative luminosity - Relative energy - filter transmission - relative luminosity of the ICI Standard Observer.

#Trans. = transmittance.

Table 23.1- Table of visual efficiency of isolating filters for the H100-A4 Mercury Arc (based on nominal values from manufacturer's literature).

although it will be seen that the mercury light has very little orange and red as compared to tungsten or daylight. When used with a filter the path differences in the specimen are seen only in the color of the filter, but with variable intensity.

23.8 POLARIZING MICROSCOPE

23.8.1 General.

23.8.1.1 The polarizing microscope is a brightfield microscope modified for examining the specimen in polarized light, with auxiliary equipment for measuring the effect of the specimen on the polarized light. A laboratory microscope can be used with polarized light by placing discs of Polaroid under the condenser and in or on the ocular. Such simple polarization will reveal the colors in birefringent materials and show strain.

23.8.1.2 For measurement, a specialized microscope is necessary. The polarizing microscope has a polarizing prism of the Nicol or Ahrens type under the condenser. The chemical type has a cap analyzer over the eyepiece and the petrographic type has the analyzer in a slide so that it can be pushed into or out of the optical axis in the body tube of the microscope. The upper lenses of the condenser are arranged so that they may be moved into or away from the optical axis of the microscope.

23.8.2 Characteristics.

23.8.2.1 Strain-free optics are necessary in the design of a polarizing microscope, and the objectives are usually mounted in centering rings of a quick change type of nosepiece.

23.8.2.2 When a Bertrand lens is pushed into the optical axis, it forms, with the ocular, a telescope for viewing the back aperture of the objective. A slot is provided for moving a quartz wedge or other compensators into the optical axis.

23.8.2.3 The ocular contains cross hairs and is positioned in the ocular tube to prevent rotation. The polarizer is rotatable to position it at 180° to the polarization angle of the analyzer and a centering rotatable stage with graduated scale and vernier are used to measure the orientation of the specimen.

23.8.2.4 The improved Polaroid is satisfactory and is used to replace the expensive crystal polarizers in some modern instruments and many special compensators, multiaxis stages and other auxiliary equipment are available and useful. The birefringence of biological materials is small, and more elaborate polarizing microscopes have been built to meet this need. One marked improvement is the rectifier for compensating depolarization from the curved objective lens that makes possible the use of the nearly full aperture of the oil immer sion objective.

23.9 FLUORESCENCE MICROSCOPES

23.9.1 General.

23.9.1.1 Fluorescence microscopy can be accomplished with a brightfield microscope when the specimen is irradiated with ultraviolet radiation. A source of filtered radiation (usually a high pressure mercury arc) is necessary and an ultraviolet absorbing filter is placed on, or in, the ocular to prevent ultraviolet radiation not absorbed by the specimen from reaching the eye. A front-surface,aluminized mirror is more efficient than a silvered mirror.

23.9.2 Characteristics.

23.9.2.1 For short wavelength ultraviolet, necessary in the study of some minerals, the condenser and slide must be of quartz or other UV transmitting materials, or a catoptric condenser be used. Long wavelengths (>330mμ) UV pass through the ordinary microscope optics and they are satisfactory.

23.9.2.2 The most efficient system uses a crossed filter technic with a brightfield condenser. The lamp filter passes the radiation absorbed by the specimen and the ocular protective filter is chosen to absorb the ultraviolet, but to pass the light emitted by the specimen. When an efficient cross-filter system is not possible, a darkfield condenser is used with a thinner UV isolating filter. A colorless UV filter is usually required for the ocular as some of the UV may be scattered by the specimen into the objective.

23.9.2.3 Achromatic objectives and Abbe condensers are preferable as the chromatically corrected ones often contain fluorescent materials which introduce glare and reduces visibility. The Abbe NA 1.40 will concentrate more energy on the specimen than the usual NA 1.25 condenser.

23.10 THE STEREOSCOPIC MICROSCOPE

23.10.1 General.

23.10.1.1 The bi-objective, binocular microscope, reinvented by Greenough is made by combining two microscopes, Figure 23.21, so that the right eye sees with the right hand side,and the left eye with the left hand side. As each eye receives a separate disparate view, true stereopsis occurs. When the angles of the objective and ocular convergence are the same,true or orthostereopsis is provided. By changing these angles increased or decreased depth can be provided.

23.10.2 Characteristics.

23.10.2.1 Prisms are included to erect the image and such instruments are useful for dissection and for the examination of small parts.

23.10.2.2 Since two objectives are required, the mechanical limitations of placement limits the numerical aperture to about 0.12 and there is no advantage in using magnifications over about 120X.

23.10.2.3 A recent modification places the paired objectives in a rotatable turret with a single,large, corrected lens between them and the specimen. By turning the turret, magnification can be readily varied within the limitations of the cycloptic microscope. Another improvement is to build the paired objectives into a zoom system so that the magnification can be varied continuously throughout its range.

23.11 PETROGRAPHIC MICROSCOPE

23.11.1 General.

23.11.1.1 The optical system of the petrographic microscope has been so adapted that the methods of petrographic measurements can be made.

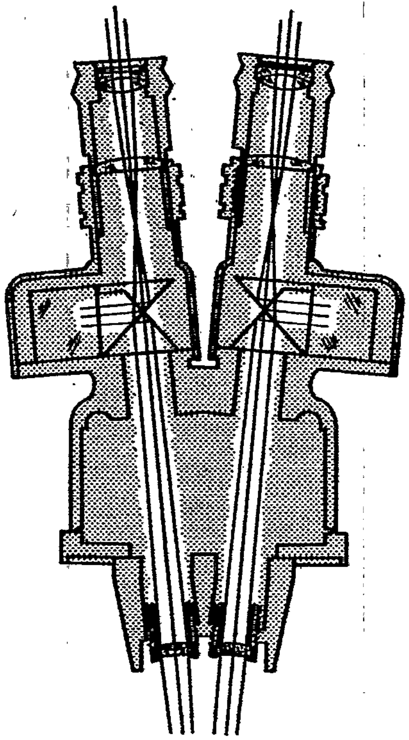


Figure 23.21- Optical schematic of a stereoscopic microscope.

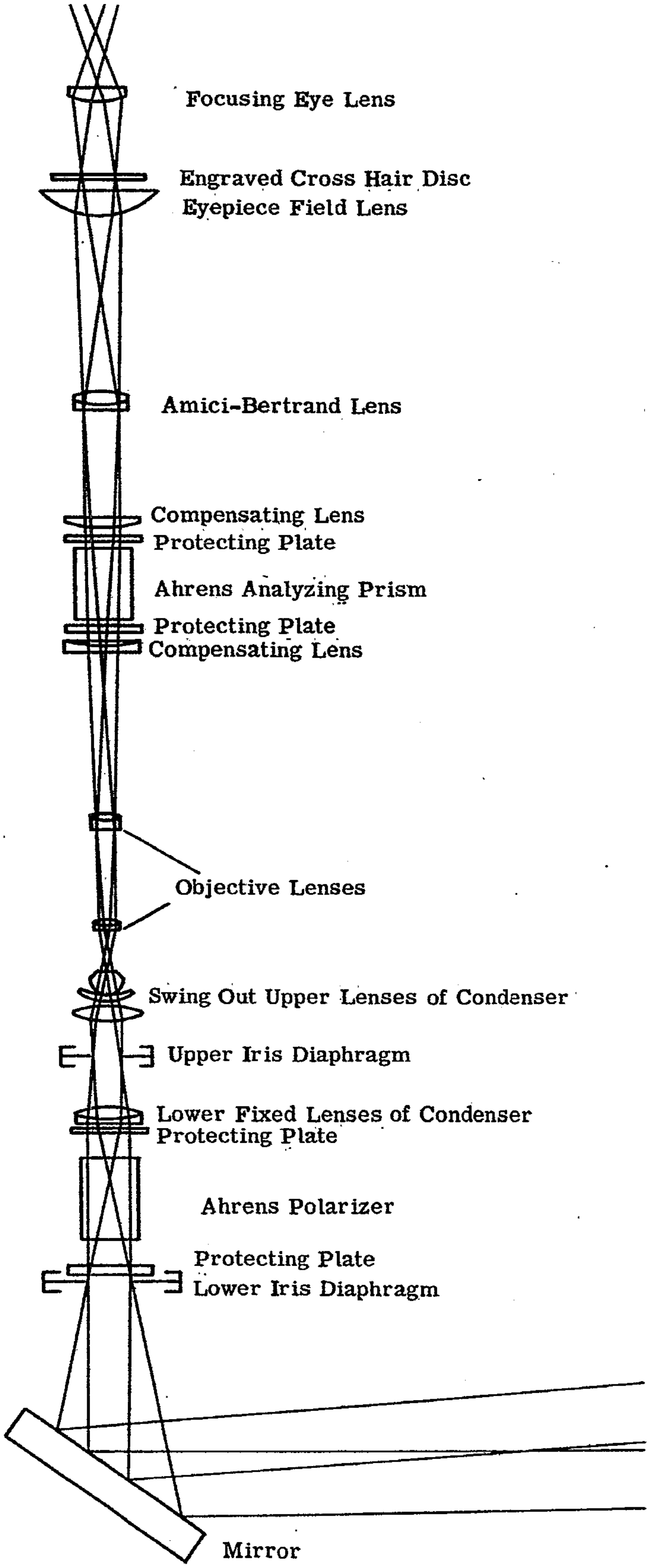


Figure 23. 22- Optical system of a petrographic microscope.

23.11.2 Characteristics.

23.11.2.1 Substage condenser. The substage condenser is made up of two parts one of which is designated as the lower fixed lenses of the condenser and the other as the swing out upper lenses of condenser. When these two systems are working together,the NA of the combination may be as great as 1.40. When only the lower part is used,the NA may be as low as 0.25. At the lower (front) focal plane of each of the above condensers,is located an iris diaphragm designated in Figure 23.22 as the lower iris diaphragm and the upper iris diaphragm. Beneath the lower condenser and between it and the lower iris diaphragm is the polarizer which may consist of a Nicol or Ahrens polarizing prism or a sheet of Polaroid. The polarizer is generally rotatable and provided with an angular scale. A detent stop may indicate the zero setting of the polarizer.

23.11.2.2 Objectives. The objective lenses used in the petrographic microscope are identical in design with the achromatic series of microscope objectives already mentioned. However these objectives must be free from strain otherwise their birefringence will interfere with measurements made upon mineral specimens.

23.11.2.3 Analyzing system. The analyzer may be a Polaroid plate or a polarizing prism of the Nicol or Ahrens type. The light passing from the objective to the eyepiece is convergent. Since a polarizing prism will produce astigmatism under such circumstances, it is necessary to parallelize the light traversing the polarizing prism. For this purpose a negative lens is used below the analyzer to cause the convergent light to become parallel. Above the analyzer is placed a convergent or positive lens of such focal length that the rays are focussed on the cross hairs of the eyepiece. These lenses need not be achromatic as the image forming bundles of rays are of such a small aperture. The introduction of these compensating lenses necessarily change the initial magnification and to avoid having different magnifications when the analyzer is inserted or withdrawn from its position on the axis of the instrument, the compensating lenses are fixed inside the body tube. The analyzing prism itself is protected against dust, fumes, and moisture by two windows labelled in Figure 23.22 as protecting plates. These plates should not be plane and parallel as there would be detrimental reflections between the surfaces of the plates. These windows should be in the form of menisci of zero power.

23.11.2.4 Amici-Bertrand Lens. This lens is located in such a position, and is of the correct focal length to image the back focal plane of the objective onto the cross hairs of the eyepiece. It will be seen that in this case the entire microscope becomes a telescope focussed for infinity. Of course,when the Amici-Bertrand lens is slid out of the instrument the system is a microscope. The Amici-Bertrand lens may be focusable and equipped with an iris diaphragm.

23.11.2.5 Eyepieces. The eyepieces in the illustration are of the Huygenian type with the eyelens focusable upon the cross lines of the reticle. The entire eyepiece is prevented from rotating by means of a tongue,or screw,in the eyepiece engaging a slot in the upper end of the body tube.