Optics for physicists

M. Saffman

March 8, 2016
Contents

1 Optical waves .................................................. 2
  1.1 Maxwell equations ........................................ 2
  1.2 Wave equation in linear media .......................... 2
  1.3 Energy and momentum in optical fields ............... 3
    1.3.1 Broadband fields ................................ 5
  1.4 Refractive Index .......................................... 6
  1.5 Polarization .............................................. 7
    1.5.1 Angular momentum of light .......................... 10
    1.5.2 Jones matrices ..................................... 11
    1.5.3 Stokes parameters ................................. 11
  1.6 Fresnel coefficients .................................... 11
  1.7 The beamsplitter ......................................... 13
    1.7.1 Brewster’s angle .................................. 15
    1.7.2 Total internal reflection ......................... 15
    1.7.3 Reflection from metals ............................ 17
    1.7.4 Calcite beam displacer ........................... 18

2 Interference ................................................... 19
  2.1 Interference of two beams ............................... 19
    2.1.1 Plane wave interference ........................... 19
    2.1.2 Interference of plane and spherical waves ....... 20
    2.1.3 Interference of two spherical waves ............. 21
    2.1.4 Multibeam interference ............................ 21
  2.2 Fabry-Pérot Interferometer ............................. 21
    2.2.1 Transmitted intensity ............................. 23
    2.2.2 Reflected intensity ................................ 24
    2.2.3 Finesse ............................................. 25
    2.2.4 Phase shifts ....................................... 27
    2.2.5 Finite beam size effects ............................ 28
    2.2.6 Sensitivity to pressure ........................... 29
    2.2.7 Determination of mode matching coefficient ...... 29
  2.3 Etalon ..................................................... 31
    2.3.1 Tilted etalon ....................................... 32
  2.4 Fourier transform spectroscopy ......................... 33
  2.5 Multilayer coatings ...................................... 35

March 8, 2016 M. Saffman
3 Geometrical optics 41
  3.1 Reflection and refraction 41
    3.1.1 Paraxial Ray Matrices 41
    3.1.2 Ray matrix for a concave mirror 44
    3.1.3 Imaging with a thin lens 45
    3.1.4 Negative lens 46
    3.1.5 Compound lens 47
    3.1.6 Principal Planes 48
    3.1.7 Lens Maker’s formula 50
    3.1.8 Optical plates 51
    3.1.9 Sign conventions 52
    3.1.10 Brightness 53
  3.2 Aberrations 53
    3.2.1 Primary monochromatic aberrations 54
    3.2.2 Chromatic aberration 56
  3.3 Image forming instruments 57
    3.3.1 The eye 57
    3.3.2 Microscope 58
    3.3.3 Telescope 59
    3.3.4 Binoculars 59
  3.4 Solid Angle 60
  3.5 Eikonal equation 61
    3.5.1 Ray equation 63

4 Fourier optics 66
  4.1 Paraxial propagation and diffraction 67
    4.1.1 Fresnel Diffraction 69
    4.1.2 Fraunhofer Diffraction 71
    4.1.3 Diffraction from an aperture 72
    4.1.4 Beyond Fresnel diffraction 75
    4.1.5 Fourier transforming properties of lenses 76
    4.1.6 Fields with radial symmetry 78
    4.1.7 Airy disk 78
    4.1.8 Spot of Arago 80
    4.1.9 Fresnel zones 81
    4.1.10 Talbot effect 82
    4.1.11 Bottle beam lenses 83
    4.1.12 More general transformations 87
    4.1.13 Summary of diffraction formulae 89
  4.2 Fourier optical image processing 90
4.2.1 Repeated objects ........................................... 90
4.2.2 Periodic apertures ......................................... 91
4.2.3 Fourier filtering ........................................... 92
4.3 Diffraction gratings ........................................... 94

5 Gaussian beams .................................................. 95
5.1 Lowest order solution ....................................... 95
5.1.1 Higher order Gaussian beams ................................ 98
5.1.2 Laguerre-Gaussian beams .................................. 100
5.1.3 Gaussian beam transformation by a lens .................. 101
5.1.4 Gaussian beam focusing .................................... 104
5.1.5 Fiber collimation .......................................... 105
5.1.6 Refocusing with a mirror ................................ 107
5.1.7 Zoom lenses for Gaussian beams ......................... 107
5.1.8 Two-lens zoom ............................................. 108
5.1.9 Three-lens zoom ......................................... 109
5.2 Apertured Gaussian Beams ................................... 111
5.2.1 Apertured beam in intermediate plane .................. 114

6 Optical resonators ............................................ 117
6.1 Resonator eigenmodes ....................................... 117
6.2 Linear cavity .................................................. 118
6.3 Linear cavity with internal optical element ................. 120
6.4 Resonant frequencies ....................................... 121
6.5 Mode matching ............................................... 123
6.6 Ring cavity ................................................... 124
6.6.1 Ring cavity with angled crystal ......................... 126

7 Electro-optics ................................................. 129
7.1 Optical Modulators .......................................... 129
7.1.1 Acousto-optic deflector ................................... 129
7.1.2 Time bandwidth product .................................. 130
7.1.3 Electro-optic modulator ................................... 131
7.2 Pound Drever Hall locking .................................. 131
7.3 Tilt-Locking .................................................. 131
7.3.1 Littrow laser scanning ................................... 132
7.3.2 Frequency comb stability ................................. 134
7.4 Photodetection and signal to noise ratio .................... 135
7.4.1 Thermal noise ............................................. 135
7.4.2 Detector characteristics .................................. 136
7.4.3 Photodetection ........................................... 138
7.4.4 Heterodyne detection ..................................... 140
7.5 References ................................................... 142
List of Figures

1.1 Reflection of light from a mirror. The coordinate system for the incident wave is \((\epsilon_x, \epsilon_y, \epsilon_z)\) and for the reflected wave \((\epsilon_{x'}, \epsilon_{y'}, \epsilon_{z'})\). To make both coordinate systems right handed we set \(\epsilon_{x'} = \epsilon_x, \epsilon_{y'} = -\epsilon_y, \epsilon_{z'} = -\epsilon_z\). .............................. 9

1.2 The beamsplitter. .............................................................................................................. 13

1.3 Perfect transmission of p-polarized light occurs at Brewster’s angle. ......................... 16

1.4 TIR at 45 deg. incidence angle with \(n_1 > n_2\). .................................................. 16

1.5 Phase shift in TIR for \(n_1 = 1.50, 1.51, 1.52\). .................................................. 17

2.1 Fabry-Perot cavity. .............................................................................................................. 22

2.2 Transmission (solid line) and reflectivity (dashed line) of a Fabry-Perot cavity with \(R_{in} = .9, R_{out} = 0.8\), and \(R_{cav} = 0.99\). The inset zooms in near the resonance. .............................................................................................................. 24

2.3 Fabry-Perot cavity finesse(left) and build up factor (right) as a function of round trip internal loss for input coupler transmission of \(T_{in} = 0.05, 0.02, 0.01\) (solid, dashed, dotted lines) and \(R_{out} = 0.999\). .............................................................................................................. 26

2.4 Transmission (solid line) and reflection (dashed line) phase shifts for \(R_{in} = R_{out} = 0.95, R_{cav} = 0.99\). The finesse is \(F = 55.8\). .............................................................................................................. 28

2.5 Etalon at normal incidence (solid line) and for tilted rays (dashed lines). .......................... 32

2.6 Fourier transform spectroscopy with a Michelson interferometer. Moving one of the mirrors changes the differential path length \(p\) leading interference at the output port. .............................................................................................................. 33

2.7 Fourier transform spectroscopy signal for an input spectrum of two Gaussians centered at \(\omega_1\) and \(\omega_2\). The left two plots show the output with only one of the Gaussians present and the right hand plot shows the output when both input frequencies are present which leads to interference. ......................... 34

2.8 Reflection from a thin film of thickness \(d\). The external medium has index \(n_0\), the film has index \(n_1\) and the substrate has index \(n_2\). .................................................. 35

2.9 Reflectance of quarter wave film with index \(n_1\) in between air \((n_0 = 1)\) and glass \((n_2 = 1.5)\). .................................................. 37

2.10 Wave propagation through a multilayer film. .................................................. 38

2.11 High reflectivity coating made of alternating quarterwave layers of high and low index. .................................................. 39

3.1 Reflection and transmission at an interface. .................................................. 41

3.2 Fractional error in approximating \(\sin(\theta)\) by \(\theta\). .................................................. 42

3.3 Ray propagation through a thin lens. .................................................. 43

March 8, 2016 M. Saffman
3.4 Ray matrices in the tangential and sagittal planes. The relative refractive index is \( n_r = n_2/n_1 \). In the fourth and fifth lines \( R > 0 \) for the interface concave to the right. The matrices for transmission through the tilted interface can be found in G.A. Massey and A. E. Siegman, Appl. Opt. 8, 975 (1969). 44

3.5 Geometry for imaging with a concave mirror. ................................. 45

3.6 Imaging with a thin lens. The object and image distances are \( d_1 \) and \( d_2 \), respectively. The distances measured from the principal foci are \( z_1 \) and \( z_2 \). 46

3.7 Compound lens composed of two thin lenses. The principal planes are labeled PP. ................................................................. 47

3.8 Principal plane analysis. A composite optical system (top) is equivalent to an effective thin lens with the propagation distances measured to and from the principal planes (bottom). 49

3.9 Thick lens. ............................................................................ 50

3.10 A tilted slab of index \( n_2 \) and thickness \( t \). ................................. 51

3.11 A tilted mirror with radius of curvature \( R \), thickness \( t \), and substrate index \( n_2 \). 52

3.12 Seidel aberrations. ................................................................. 54

3.13 Spherical aberration of a plano-convex lens. .............................. 56

3.14 Optics of the human eye. ........................................................ 57

3.15 Microscope with objective lens and eyepiece. ............................. 58

3.16 Binocular optics. ..................................................................... 59

3.17 Calculation of solid angle of window or lens. .............................. 60

3.18 Fractional solid angle as a function of cone angle \( \theta \), \( f_\# \), and \( NA \). The dashed line is the cone angle as a function of the \( f_\# \) or \( NA \). 61

3.19 Ray propagation in a medium with continuously varying index \( n(r) \). 62

3.20 Periodic ray trajectories in a gradient index fiber. The numerical solutions are for \( n(x) = 1.2 \text{sech}(x) \), and \( \theta(0) = 0.1, 0.79 \). ............................ 64

4.1 Huygens’ construction for describing propagation of wavefronts. .... 66

4.2 Coordinate system for optical propagation. ................................. 68

4.3 Paraxial evolution can be solved by convolution with the impulse response in the spatial domain or multiplication by the transfer function in the Fourier domain. ................................................................. 70

4.4 Fresnel integrals \( C(s), S(s) \) and Cornu spiral. .............................. 73

4.5 Diffracted intensity using Fresnel diffraction (blue solid line) and Fraunhofer diffraction (green dashed line) from a slit of width \( L = 20 \mu m \), with \( \lambda = 1. \mu m \) at different distances \( z/z_{\text{Fresnel}} = 0.1, 0.5, 1, 5, 10 \). The intensity has been normalized to the incident intensity at the aperture plane. The aperture parameters give \( z_{\text{Fresnel}} = 54. \mu m \) and \( z_{\text{Fraunhofer}} = 400. \mu m \). ......................... 74

4.6 Diffracted intensity from Fresnel diffraction theory (blue solid line) and Rayleigh-Sommerfeld theory Eq. (4.17) (green dashed line) from a square aperture of width \( L = 20 \mu m \), with \( \lambda = 1. \mu m \) at different distances \( z/z_{\text{Fresnel}} = 0.1, 0.5, 1 \). The intensity is shown as a function of \( x_2 \) for \( y_2 = 0 \) and has been normalized to the incident intensity at the aperture plane. The aperture parameters give \( z_{\text{Fresnel}} = 54. \mu m \). ........................................ 76

4.7 Field transformation with a lens. .................................................. 77
4.8 Airy diffraction rings formed by uniform illumination of a lens. The circular lens parameters are diameter \( d = 100 \ \mu\text{m} \), focal length \( f = 1 \ \text{cm} \), \( \lambda = 1 \ \mu\text{m} \). The annulus has inner diameter \( 0.6d \) and outer diameter \( d \). The intensity is normalized to the incident intensity at the lens.  

4.9 Spot of Arago calculated for \( \lambda = 1 \ \mu\text{m} \), \( a = 10 \ \mu\text{m} \), giving \( z_{\text{Fresnel}} = 54 \ \mu\text{m} \). The diffracted intensity normalized to the incident intensity is shown for \( z/z_{\text{Fresnel}} = 1, 2, 10, 20 \).  

4.10 Fresnel zone plate for lensless focusing.  

4.11 Gaussian beam illuminating a lens with a \( \pi \) phase in the inner region of radius \( b \).  

4.12 BoB profiles for \( w_0 = 82.76 \ \mu\text{m}, w_3 = 3 \ \mu\text{m}, b = 69 \ \mu\text{m} \) and \( s = 1 \) (left). These parameters correspond to \( \lambda = 0.78 \ \mu\text{m} \) and \( f = 1 \ \text{mm} \). The plotted intensity is normalized to the peak of the input Gaussian with waist \( w_0 \). The inset shows that the on-axis intensity does not vanish for any \( b \). On the right the transverse profiles are shown at axial displacements up to \( 50 \ \mu\text{m} \) from the focal plane.  

4.13 Three dimensional intensity distribution about the focus. Same parameters as in Fig. 4.12.  

4.14 Fourier filtering setup. The filter function \( t(\rho_2) \) is positioned in the Fourier plane between the lenses.  

5.1 Gaussian beam of waist \( w_0 \).  

5.2 Intensity of Hermite-Gaussian beams in the focal plane \( z = 0 \) with \( \lambda = 0.5, w_0 = 1 \) in regions of size \( 6 \times 6 \).  

5.3 Intensity of Laguerre-Gaussian beams in the focal plane \( z = 0 \) with \( \lambda = 0.5, w_0 = 1 \) in regions of size \( 6 \times 6 \).  

5.4 Gaussian beam transformation by a lens of focal length \( f \). The incident waist \( w_{10} \) is positioned a distance \( z_1 \) in front of the lens. The output waist \( w_{20} \) occurs a distance \( z_2 \) after the lens.  

5.5 Output waist size (left) and output waist position (right) for transformation by a single lens.  

5.6 Numerical aperture for a given focusing ratio at aperture parameter \( p = 4 \). The dashed line shows a large \( r \) approximation.  

5.7 Output waist (left) and waist position (right) for collimation of the light emitted by a single mode fiber.  

5.8 Width on a screen at distance \( z_s \) vs. lens position \( z_1 \). The curves are labeled with \( z_s = 1, 2, 5, 10 \ \text{m} \).  

5.9 Beam width as \( z_1 \) is varied (left) and as \( z_2 \) is varied (right). The flat line in the left hand graph is an exact result for all \( z_1 \). The flat line in the right hand graph is an approximate result for \( z_2 \) similar to \( f \).  

5.10 Refocusing of light in a crystal with a curved mirror.  

5.11 Layout of a two lens zoom.
5.12 Layout of a three lens zoom. The three lens zoom is afocal so we include a
dispersion of a fixed $f_4$ to transform the output. The nominal distances are $z_{10}, z_{20}, z_{30}, z_{40}$.
The lenses move by amounts $\delta_1, \delta_2, \delta_3$ during zoom operation giving actual
distances $z_1 = z_{10} + \delta_1, z_2 = z_{20} + \delta_2 - \delta_1, z_3 = z_{30} + \delta_3 - \delta_2$, and $z_4 = z_{40} - \delta_3$. 110
5.13 Three lens zoom as in Fig. 5.12 with $f_1 = 5, f_2 = -2.5, f_3 = 5, f_4 = 9$. The
actual magnification is shown on the left and the lens displacements on the
right. ........................................ 111
5.14 Gaussian beam filtered with a circular aperture. ...................... 111
5.15 Intensity profiles in back focal plane for $w_1 = w_2 = 0.126$ mm, $\lambda = 0.5 \mu$m,
$f = 100$. mm and $D/w_1 = 1, 2, 3, 4, 5$ (left). The profile for $D/w_1 = 5$ is
indistinguishable from that with $D = \infty$ shown in purple. The right hand
plot shows the weak oscillating tails for $D/w_1 = 4, 4.25, 4.5$. .............. 112
5.16 Airy ring structure of the intensity profiles in the back focal plane for $D/w_1 = 
0.5, 1, 2$, and all other parameters the same as in Fig. 5.15. The red curves
are for a Gaussian beam with no aperture, the blue curves are the apertured
Gaussian and the green curves are from the Airy formula (4.22). ............ 113
5.17 Intensity profiles for $D/w_1 = 4$, and all other parameters the same as in Fig.
5.15 for $z = (0, 0.1, 0.5, 1, 2)L_R$ with $L_R = \pi w_2^2/\lambda = 200$. mm. ......... 114
6.1 Linear resonator with spherical mirrors. .............................. 118
6.2 Stability diagram for a linear resonator. The shaded regions are stable, and
everywhere else is unstable. The points marked with dots are symmetric plane-
plane ($g_1 = g_2 = 1, R_1 = R_2 = \infty$), confocal ($g_1 = g_2 = 0, L = R_1 = R_2$), and
concentric ($g_1 = g_2 = -1, L = 2R_1 = 2R_2$). The figure also shows asymmetric
hemispherical and concave-convex resonators. .......................... 119
6.3 Linear resonator containing an optical crystal of length $L_c$ and refractive index
$n_c$. ........................................... 121
6.4 Power overlap between a Gaussian beam and a Gaussian cavity mode. The
line plots show the variation with $s$ and $\zeta$. ............................... 121
6.5 Mode matching to a linear resonator. Without the input coupler (top) the
focusing lens produces a waist $w_0$ at position $z_0$. With the input coupler in
place (bottom) the waist is transformed to $w_2$ at position $z_2$. ............... 123
6.6 Ring bowtie cavity. The fields are drawn for a harmonic generation configu-
ration. .................................................. 125
6.7 Fundamental mode waist for a bowtie cavity with $R = 4$ cm, $L_c = 3$ cm,
$n_c = 2.3$, and $\lambda = 0.96 \mu$m. ........................................ 125
6.8 Ring cavity for an optically pumped laser with angled crystal. The geometrical
round trip distance $L_T$ is $L_T = L_2 + L_{d1} + L_{d2} + L_u$. The path length inside
the crystal is $L_c$. The detail in the upper right shows the geometry of the
Brewster cut crystal. The beam path in the lower right with positions 1 – 6
is used for generating a detailed optical layout. Position 0 is at the center of
the crystal. .................................................. 126
6.9 Calculated waists at the center of the Ti:Sa crystal (left) and between the flat
mirrors (right). Parameters: $\lambda = 0.685 \mu$m, $L_c = 1$ cm, $n_c = 1.77$, $R = 2.5$
cm, $\theta = 19$ deg., $L_T = 15$ cm, $h = 2.3$ cm. .............................. 127
<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>6.10</td>
<td>Layout of the Ti:Sa cavity for the parameters of Fig. 6.9.</td>
<td>128</td>
</tr>
<tr>
<td>7.1</td>
<td>Bragg diffraction in an acousto-optic deflector with acoustic wavelength $\Lambda$.</td>
<td>129</td>
</tr>
<tr>
<td>7.2</td>
<td>Error signal as a function of cavity tuning.</td>
<td>131</td>
</tr>
<tr>
<td>7.3</td>
<td>Geometrical layout of a Littrow laser cavity.</td>
<td>133</td>
</tr>
<tr>
<td>7.4</td>
<td>Detector electronics.</td>
<td>136</td>
</tr>
</tbody>
</table>
Chapter 1

Optical waves

1.1 Maxwell equations

In SI units the Maxwell equations are

\[
\begin{align*}
\nabla \times E &= -\frac{\partial B}{\partial t}, \\
\nabla \times H &= J + \frac{\partial D}{\partial t}, \\
\nabla \cdot D &= \rho, \\
\nabla \cdot B &= 0.
\end{align*}
\]

(1.1a) (1.1b) (1.1c) (1.1d)

These are supplemented by the constitutive relations

\[
D = \varepsilon_0 E + P, \quad H = \frac{1}{\mu_0} B - M.
\]

(1.2a) (1.2b)

In general the electric and magnetic polarization of the medium \( P \) and \( M \) are nonlinear and anisotropic functions of the fields. In the simplest case of linear and isotropic media the constitutive relations can be written

\[
D = \varepsilon E, \quad H = B/\mu.
\]

In vacuum there is no medium to be polarized so \( P = 0, M = 0 \) and \( \varepsilon = \varepsilon_0, \mu = \mu_0 \).

1.2 Wave equation in linear media

Assuming linear and isotropic media we have

\[
\nabla \times \nabla \times E = -\frac{\partial}{\partial t} \nabla \times B = -\frac{\partial}{\partial t} \nabla \times \mu H = -\frac{\partial}{\partial t} (\nabla \times \mu H + \mu \nabla \times H).
\]

The left hand side is

\[
\nabla \times \nabla \times E = \nabla (\nabla \cdot E) - \nabla^2 E = \nabla (\nabla \cdot D/\varepsilon) - \nabla^2 E = \nabla (\frac{1}{\varepsilon} \nabla \cdot D + \nabla (\frac{1}{\varepsilon}) \cdot D) - \nabla^2 E
\]
If $\epsilon$ and $\mu$ are uniform in space then $\nabla(1/\epsilon) = 0$ and $\nabla\mu = 0$. In addition if there are no free charges and no currents then $\nabla \cdot D = 0$ and $J = 0$ and we get

$$-\nabla^2 E = -\mu \frac{\partial}{\partial t} \nabla \times H = -\epsilon\mu \frac{\partial^2 E}{\partial t^2}$$

or

$$\frac{\partial^2 E}{\partial t^2} - \frac{1}{\epsilon\mu} \nabla^2 E = 0.$$

The wave speed is $v = 1/\sqrt{\epsilon\mu}$ and the wave equation is

$$\frac{\partial^2 E}{\partial t^2} - v^2 \nabla^2 E = 0.$$

In vacuum $\epsilon\mu = \epsilon_0\mu_0$ and the wave speed is $1/\sqrt{\epsilon_0\mu_0} \equiv c$ the speed of light. It is convenient to introduce the refractive index $n = \sqrt{\mu/\epsilon_0\mu_0}$ and write $v = c/n$. The refractive index has the value 1 in vacuum and can be greater than or less than 1 inside media. Remarkably it is also possible to have negative $n$ which leads to some unusual and very interesting effects.

1.3 Energy and momentum in optical fields

From the Maxwell equations (1.1a-1.1d) we have that

$$\mathbf{E} \cdot (\nabla \times \mathbf{H}) - \mathbf{H} \cdot (\nabla \times \mathbf{E}) = \mathbf{E} \cdot \frac{\partial \mathbf{D}}{\partial t} + \mathbf{H} \cdot \frac{\partial \mathbf{B}}{\partial t} + \mathbf{J} \cdot \mathbf{E},$$

which can be rewritten as

$$\mathbf{E} \cdot \frac{\partial \mathbf{D}}{\partial t} + \mathbf{H} \cdot \frac{\partial \mathbf{B}}{\partial t} + \nabla \cdot (\mathbf{E} \times \mathbf{H}) = -\mathbf{J} \cdot \mathbf{E}. \quad (1.3)$$

Assuming that the medium is linear in its electric and magnetic properties we can rewrite Eq. (1.3) in the form

$$\frac{\partial u}{\partial t} + \nabla \cdot \mathbf{S} = -\mathbf{J} \cdot \mathbf{E}, \quad (1.4)$$

where the energy density is

$$u = \frac{1}{2} (\mathbf{E} \cdot \mathbf{D} + \mathbf{H} \cdot \mathbf{B}), \quad (1.5)$$

and the energy flow is given by the Poynting vector

$$\mathbf{S} = \mathbf{E} \times \mathbf{H}. \quad (1.6)$$

The rate at which the electromagnetic fields do work on a current distribution $\mathbf{J}$ in a volume $V$ is given by

$$\int_V d^3x \; \mathbf{J} \cdot \mathbf{E}, \quad (1.7)$$

March 8, 2016 M. Saffman
from which we can identify \(-\mathbf{J} \cdot \mathbf{E}\) as the local rate at which charges do work on the fields. Equation (1.4) can therefore be interpreted as an energy continuity equation.

We now wish to find the intensity of a traveling electromagnetic wave. The Maxwell equations are written in terms of real fields, whereas it is convenient to use a complex notation for electromagnetic waves. Consider a component of the real electric field

\[ E = |E| \cos (\omega t - \theta) = \frac{|E|}{2} \left( e^{-i(\omega t - \theta)} + e^{i(\omega t - \theta)} \right) = \frac{|E|e^{i\theta}}{2} e^{-i\omega t} + \text{c.c.} = \frac{E}{2} e^{-i\omega t} + \text{c.c.} . \]  

(1.8)

Specializing to scalar, monochromatic traveling plane waves we can write

\[ E = \hat{x} \frac{\mathcal{E}}{2} e^{i(kz - \omega t)} + \text{c.c.} , \]  
\[ \mathbf{H} = \hat{y} \frac{\mathcal{H}}{2} e^{i(kz - \omega t)} + \text{c.c.} . \]  

(1.9a)

(1.9b)

The time averaged Poynting vector is then

\[ \langle \mathbf{S} \rangle = \hat{z} \frac{1}{4} (\mathcal{E} \mathcal{H}^* + \mathcal{E}^* \mathcal{H}) . \]  

(1.10)

Using the Maxwell equations, the expression for the speed of light \( c = (\epsilon_0 \mu_0)^{-1/2} \), and assuming the medium is non-magnetic so \( \mu = \mu_0 \), we find that \( \mathcal{H} = \epsilon_0 n c \mathcal{E} \), where \( n \) is the index of refraction \( (c/n = \omega/k) \). Hence

\[ \langle \mathbf{S} \rangle = \frac{\epsilon_0 n c}{2} |\mathcal{E}|^2 . \]  

(1.11)

The intensity is thus

\[ I = \frac{\epsilon_0 n c}{2} |\mathcal{E}|^2 \text{ [W/m}^2\text{]}, \]  

(1.12)

where the field \( \mathcal{E} \) has units of V/m. In cgs units \( [\mathcal{E}] = \text{statvolts/cm} \) and the corresponding intensity is

\[ I = \frac{n c}{8\pi} |\mathcal{E}|^2 \text{ [dyne/cm}^2\text{]} . \]  

(1.13)

Although the cgs expression is free of factors of \( \epsilon_0 \) the field units of statvolts/cm are not as convenient for practical calculations as are SI units where we have V/m. Although the choice of units is arbitrary, there has been a strong trend towards adoption of SI in the last decades and we will follow the trend.

As an example of an optical field strength consider a beam with power \( P = 1 \text{ mW} \) and a diameter of \( d = 1 \text{ mm} \). This corresponds to a low power laser pointer. The intensity is \( I = P/A = 4P/\pi d^2 = 1270. \text{ W/m}^2 \). Using \( n = 1 \), \( c = 299792458 \text{ m/s} \), and \( \epsilon_0 = 8.85 \times 10^{-12} \) we find a field strength \( \mathcal{E} = 980 \text{ V/m} = 0.98 \text{ V/mm} \). The (oscillating) potential difference across the beam is just under 1 V.
1.3 Energy and momentum in optical fields

1.3.1 Broadband fields

When the optical field consists of a continuum of frequencies with a spectral width that is not small compared to the mean frequency a description in terms of a single carrier frequency \( \omega \) as in Eqs. (1.9) is no longer convenient. Instead we allow for a continuum of frequencies with each frequency having a spectral amplitude \( \tilde{E}(\omega) \). The time domain field is then given by the Fourier representation

\[
E(z, t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} d\omega \tilde{E}(\omega)e^{i[k(\omega)z-\omega t]} = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} d\omega \tilde{E}(\omega)e^{i\tilde{\phi}(z-ct)}.
\]

The real electric field is

\[
E(z, t) = \frac{\mathcal{E}(z, t)}{2} + \frac{\mathcal{E}^*(z, t)}{2}
= \frac{1}{2\sqrt{2\pi}} \int_{-\infty}^{\infty} d\omega \tilde{E}(\omega)e^{i\tilde{\phi}(z-ct)} + \frac{1}{2\sqrt{2\pi}} \int_{-\infty}^{\infty} d\omega \tilde{E}^*(\omega)e^{-i\tilde{\phi}(z-ct)}.
\]

It is easily verified that the reality condition implies that

\[ \tilde{E}(\omega) = \tilde{E}^*(-\omega). \]

Using this relation we can write the real field as a single sided integral

\[ E(z, t) = \frac{1}{\sqrt{2\pi}} \int_{0}^{\infty} d\omega \left[ \tilde{E}(\omega)e^{i\tilde{\phi}(z-ct)} + \tilde{E}^*(\omega)e^{-i\tilde{\phi}(z-ct)} \right]. \]

This shows that it is sufficient to represent \( E(z, t) \) using only positive frequency spectral amplitudes.

The time averaged intensity is

\[
\langle I \rangle_T = \frac{1}{T} \frac{\varepsilon_0 c}{2} \int_{-T/2}^{T/2} d\tau \mathcal{E}(z, t+\tau)\mathcal{E}^*(z, t+\tau)
= \frac{1}{T} \frac{\varepsilon_0 c}{2} \int_{-T/2}^{T/2} d\tau \mathcal{E}(z, \tau)\mathcal{E}^*(z, \tau)
= \frac{1}{T} \frac{\varepsilon_0 c}{2} \int_{-T/2}^{T/2} d\tau \int_{0}^{\infty} d\omega \tilde{\mathcal{E}}(\omega)e^{i\tilde{\phi}(z-ct)} \int_{0}^{\infty} d\omega' \tilde{E}^*(\omega')e^{-i\tilde{\phi}(z-ct)}
= \frac{1}{T} \frac{\varepsilon_0 c}{2} \int_{0}^{\infty} d\omega \int_{0}^{\infty} d\omega' \tilde{\mathcal{E}}(\omega)\tilde{E}^*(\omega')e^{i(\omega-\omega')z/c} \int_{-T/2}^{T/2} d\tau e^{-i(\omega-\omega')\tau}.
\]

To proceed we assume the intensity is statistically stationary so that \( \langle I \rangle_T \) is independent of \( t \), and put \( t = 0 \). The integral over \( \tau \) is written as

\[
\frac{1}{2\pi} \int_{-T/2}^{T/2} d\tau e^{-i(\omega-\omega')\tau} = \delta_T(\omega - \omega').
\]

When the spectral support is band limited, i.e. \( \mathcal{E}(\omega) \approx 0 \) for \( |\omega| < \omega_{\text{max}} \) and \( T \gg 1/\omega_{\text{max}} \) we can make the approximation \( \delta_T(\omega - \omega') \approx \delta(\omega - \omega') \) to arrive at

\[
\langle I \rangle_T = \frac{1}{T} \frac{\varepsilon_0 c}{2} \int_{0}^{\infty} d\omega |\tilde{\mathcal{E}}(\omega)|^2.
\]
For sufficiently long averaging times we define an average intensity \( \bar{I} = \lim_{T \to \infty} (T \langle I \rangle) = \frac{\omega}{2} \int_0^\infty d\omega |\tilde{E}(\omega)|^2 \) and then introduce a normalized spectral intensity \( S(\omega) \) by
\[
S(\omega) = \frac{\epsilon_0 c^2 |\tilde{E}(\omega)|^2}{\bar{I}} = \frac{\tilde{I}(\omega)}{\bar{I}}
\]
such that
\[
\int_0^\infty d\omega S(\omega) = 1.
\]
With these definitions the frequency dependent field amplitude has units of \([\tilde{E}(\omega)] = V/Hz^{1/2}m\) and for the spectral intensity \([S(\omega)] = 1/Hz\).

### 1.4 Refractive Index

We introduced the refractive index \( n = \sqrt{\frac{\epsilon \mu}{\epsilon_0 \mu_0}} \) at the end of Sec. 1.2. Let’s take a closer look at the refractive index. In linear isotropic media with no free charges or currents the Maxwell equations are
\[
\begin{align*}
\nabla \times \mathbf{E} &= -\frac{\partial \mathbf{B}}{\partial t}, \\
\nabla \times \mathbf{H} &= \frac{\partial \mathbf{D}}{\partial t}, \\
\n\nabla \cdot \mathbf{D} &= 0, \\
\n\nabla \cdot \mathbf{B} &= 0,
\end{align*}
\]
and \( \mathbf{D} = \epsilon \mathbf{E}, \mathbf{H} = \frac{1}{\mu} \mathbf{B} \). Let’s write the fields in complex form as plane waves propagating in the direction \( \hat{k} \)
\[
\mathbf{E} = \frac{\mathcal{E}}{2} e^{i\theta} \hat{E} + c.c.
\]
with \( \theta = \mathbf{k} \cdot \mathbf{r} - \omega t, \mathbf{k} = k \hat{k} \), and similarly for \( \mathbf{D}, \mathbf{B}, \mathbf{H} \). Here \( \hat{E} \) is a unit vector in the direction of \( \mathbf{E} \).

The differential operators acting on this representation take on a simple form:
\[
\nabla \times \ldots \to i\mathbf{k} \times \ldots, \quad \frac{\partial}{\partial t} \ldots \to -i\omega \ldots
\]
With these rules the first two Maxwell equations can be written as
\[
\hat{k} \times \hat{E} = \frac{\omega}{k} \frac{\mathcal{B}}{\mathcal{E}} \hat{B}, \quad \hat{k} \times \hat{H} = -\frac{\omega}{k} \frac{\mathcal{D}}{\mathcal{H}} \hat{D}.
\]
In linear and isotropic media \( \mathcal{D} = \hat{E} \) and \( \mathcal{H} = \hat{B} \) so that
\[
\hat{k} \times \hat{E} = \frac{\omega}{k} \frac{\mathcal{B}}{\mathcal{E}} \hat{B}, \quad \hat{k} \times \hat{B} = -\epsilon \mu \frac{\omega}{k} \frac{\mathcal{E}}{\mathcal{B}} \hat{E}.
\]
Plane waves have the direction of \( \mathbf{E} \) and \( \mathbf{B} \) perpendicular to \( \hat{k} \) so
\[
\hat{k} \times (\hat{k} \times \hat{E}) = (\hat{k} \cdot \hat{E})\hat{k} - (\hat{k} \cdot \hat{k})\hat{E} = -\hat{E} = \frac{\omega}{k} \frac{\mathcal{B}}{\mathcal{E}} \hat{k} \times \hat{B} = -\epsilon \mu \frac{\omega^2}{k^2} \hat{E}.
\]
Thus we must have that
\[ \epsilon \mu \frac{\omega^2}{k^2} = 1. \]
Then use \( \omega/k = \nu \lambda = v = c/n \) so the refractive index is
\[ n^2 = c^2 \epsilon \mu = \frac{\epsilon \mu}{\epsilon_0 \mu_0} \]
or
\[ n = \sqrt{\frac{\epsilon \mu}{\epsilon_0 \mu_0}}. \]
This is the same result as we found earlier from the wave equation. However we can now deduce something else that is new. First, notice that the last equation can be written as
\[ \epsilon \mu \frac{n^2}{c^2}. \]
Next, since \( \hat{k} \perp \hat{E} \) it follows that
\[ \frac{\omega B}{k \hat{E}} = 1 \]
or
\[ |B| = \frac{|E|}{c/n}. \]
Therefore
\[ \left( \frac{\omega n}{k c} \right) \hat{E} \times \hat{B} = \hat{E} \times (k \times \hat{E}) = (\hat{E} \cdot \hat{k})\hat{k} - (\hat{E} \cdot \hat{E})\hat{k} = \hat{k} \]
or
\[ \hat{E} \times \hat{B} = \left( \frac{k c}{\omega n} \right) \hat{k}. \]
When all the quantities in the parentheses are positive we have \( \hat{E} \times \hat{B} = \hat{k} \), the Poynting vector \( S \) is parallel to \( \hat{k} \) and this is called a right handed medium. This is the typical situation.

However, it is also possible to have \( n < 0 \). In this case \( \hat{E} \times \hat{B} = -\hat{k} \), the Poynting vector \( S \) points opposite to \( \hat{k} \) and this is called a left handed medium. This is the case when both \( \epsilon \) and \( \mu \) are negative. This does not occur naturally but it is possible to engineer materials with \( n < 0 \). Such materials have some remarkable properties that we will learn more about later.

### 1.5 Polarization

The most general polarization of a plane electromagnetic wave traveling along \( \epsilon_z \) is\(^1\)
\[
E = E_x \cos(kz - \omega t + \phi_x)\epsilon_x + E_y \cos(kz - \omega t + \phi_y)\epsilon_y
\]
\[= |E| e^{i(kz-\omega t)} \frac{\epsilon_x E_x + \epsilon_y E_y}{|E|} + \text{c.c.}. \]

\(^1\)We will sometimes use \( \epsilon_x, \epsilon_y, \epsilon_z \) to denote the Cartesian unit vectors \( \hat{x}, \hat{y}, \hat{z} \).
The complex field amplitudes are $\mathcal{E}_x = E_x e^{i\phi_x}$, $\mathcal{E}_y = E_y e^{i\phi_y}$, and $|\mathcal{E}| = \sqrt{E_x^2 + E_y^2}$. The field $\mathbf{E}$ is a real quantity that we express in terms of complex amplitudes $\mathcal{E}_x, \mathcal{E}_y$ for mathematical convenience.

Defining a complex, unit polarization vector as

$$\mathbf{e} = \frac{\mathcal{E}_x \mathbf{e}_x + \mathcal{E}_y \mathbf{e}_y}{|\mathcal{E}|},$$

the vector field can be written as

$$\mathbf{E} = |\mathcal{E}| \left( \frac{e^{i(kz-\omega t)}}{2} \mathbf{e} + \frac{e^{-i(kz-\omega t)}}{2} \mathbf{e}^* \right).$$

If we pick a fixed reference plane, say $z = 0$, the time dependence of the field is

$$\mathbf{E}(z = 0) = |\mathcal{E}| \left( \frac{e^{-i\omega t}}{2} \mathbf{e} + \frac{e^{i\omega t}}{2} \mathbf{e}^* \right).$$

The most general complex unit vector in the $x-y$ plane can be written as

$$\mathbf{e} = \cos \theta e^{i\chi_x} \mathbf{e}_x + \sin \theta e^{i\chi_y} \mathbf{e}_y.$$ 

The field then takes the form

$$\mathbf{E} = |\mathcal{E}| \left[ \cos \theta (\cos \chi_x \cos \omega t - \sin \chi_x \sin \omega t) \mathbf{e}_x + \sin \theta (\cos \chi_y \cos \omega t - \sin \chi_y \sin \omega t) \mathbf{e}_y \right].$$

There are several basic states of polarization. When $\chi_y = \chi_x \pm \pi$ then

$$\mathbf{E} = |\mathcal{E}| \left[ \cos \theta (\cos \chi_x \cos \omega t - \sin \chi_x \sin \omega t) \mathbf{e}_x - \sin \theta (\cos \chi_x \cos \omega t - \sin \chi_x \sin \omega t) \mathbf{e}_y \right] = |\mathcal{E}| (\cos \theta \mathbf{e}_x - \sin \theta \mathbf{e}_y) \cos(\omega t + \chi_x).$$

and $E_y/E_x = -\tan \theta = \text{constant at all times}$ so the polarization is linear.

When $\chi_y = \chi_x \pm \pi/2$ we have

$$\mathbf{E} = |\mathcal{E}| \left[ \cos \theta (\cos \chi_x \cos \omega t - \sin \chi_x \sin \omega t) \mathbf{e}_x \pm \sin \theta (\sin \chi_x \cos \omega t \pm \cos \chi_x \sin \omega t) \mathbf{e}_y \right] = |\mathcal{E}| [\cos \theta \cos(\omega t + \chi_x) \mathbf{e}_x \pm \sin \theta \sin(\omega t + \chi_x) \mathbf{e}_y]\]$$

so

$$E_y/E_x = \pm \tan \theta \tan(\omega t + \chi_x).$$

The light is elliptically polarized since the field vector traces out an ellipse in the $x-y$ plane with period $2\pi/\omega$.

When $\theta = \pi/4$ so tan $\theta = 1$ the light is circularly polarized and choosing the time origin such that $\chi_x = 0$ the electric field is

$$\mathbf{E} = |\mathcal{E}| \frac{\cos(\omega t) \mathbf{e}_x \pm \sin(\omega t) \mathbf{e}_y}{\sqrt{2}}. \quad (1.16)$$

The plus(minus) signs correspond to an observer looking parallel with the direction of light propagation (towards positive $z$) seeing the electric field rotating about the $\mathbf{e}_z$ axis in a
clockwise (counterclockwise) direction. These two cases are referred to as left hand circular (lhc) and right hand circular (rhc) states of polarization. The electric field for circular polarization is thus

\[ E_{\text{rhc}} = |E| \frac{\cos(\omega t)\epsilon_x - \sin(\omega t)\epsilon_y}{\sqrt{2}}, \quad E_{\text{lhc}} = |E| \frac{\cos(\omega t)\epsilon_x + \sin(\omega t)\epsilon_y}{\sqrt{2}}, \]

and the complex basis vectors are

\[ \epsilon_{\text{rhc}} = \frac{\epsilon_x - i\epsilon_y}{\sqrt{2}}, \quad \epsilon_{\text{lhc}} = \frac{\epsilon_x + i\epsilon_y}{\sqrt{2}}. \]

When using complex basis vectors orthogonality is determined using a complex inner product. We define the overlap of two polarization states 1, 2 as

\[ f = \epsilon_{2}^{*} \cdot \epsilon_{1}. \]

With this definition we see that

\[ \epsilon_{\text{rhc}}^{*} \cdot \epsilon_{\text{rhc}} = \epsilon_{\text{lhc}}^{*} \cdot \epsilon_{\text{lhc}} = 1, \quad \epsilon_{\text{rhc}}^{*} \cdot \epsilon_{\text{lhc}} = \epsilon_{\text{lhc}}^{*} \cdot \epsilon_{\text{rhc}} = 0. \]

One may ask what happens to the handedness of circular polarization upon reflection from a mirror. The geometry is shown in Fig. 1.1. At normal reflection from a conducting surface the parallel component of the field must vanish so \( E_{x'} = -E_x, \ E_{y'} = -E_y \), where \( r \) refers to the reflected beam. Due to the inversion of the \( y \) axis to keep the coordinate system right handed we find

\[ E_{x'r} = E_{x'} = -E_x, \]

---

\(^2\)This convention is traditional, but is opposite to the usual right hand rule for vector cross products. Note that if we define the sense of field rotation as that seen by an observer looking towards negative \( z \) (opposite to the direction of light propagation) then a clockwise (counterclockwise) sense of rotation is right (left) hand circular polarization.

March 8, 2016 M. Saffman
but  
\[ E_{y' r} = -E_{y r} = E_y. \]

Thus  
\[ E_{y' r}/E_{x' r} = -E_y/E_x \]

which implies that the handedness of the light is inverted upon reflection.

### 1.5.1 Angular momentum of light

Light carries angular momentum due to the polarization (spin) state and the spatial structure of the mode. We will consider here only the polarization contribution for the simplest case of an electromagnetic field propagating in vacuum.

The classical expression for angular momentum is \( \mathbf{J} = \mathbf{r} \times \mathbf{P} \). The energy density can be written as  
\[ \frac{1}{c} \langle \mathbf{S} \rangle = \frac{1}{c^2} \langle \mathbf{E} \times \mathbf{H} \rangle \]

so the momentum density can be written as  
\[ \mathbf{p} = \frac{\hbar k}{\hbar \omega} \langle \mathbf{S} \rangle = \epsilon_0 \mu_0 \langle \mathbf{E} \times \mathbf{B} \rangle. \]

The angular momentum density is thus  
\[ \mathbf{j} = \epsilon_0 \mathbf{r} \times \langle \mathbf{E} \times \mathbf{B} \rangle. \]  
(1.17)

This expression accounts for the angular momentum due to the spatial structure of the field, which is called orbital angular momentum, as well as the angular momentum due to polarization or spin of the photons. When the field is a plane wave with planes of constant \( \mathbf{E} \) and \( \mathbf{B} \) there is no orbital angular momentum. In this case Eq. (1.17) describes the spin angular momentum.

The magnetic field can always be written as \( \mathbf{B} = \nabla \times \mathbf{A} \) with \( \mathbf{A} \) the vector potential. It can furthermore be shown that Eq. (1.17) is equivalent to  
\[ \mathbf{j} = \epsilon_0 \mathbf{E} \times \mathbf{A} \]

so the angular momentum of a beam is  
\[ \mathbf{J} = \epsilon_0 \int_{\text{beam}} d\mathbf{r} \mathbf{E} \times \mathbf{A} \]

where \( \mathbf{A} \) is the vector potential. The electric field is related to the vector potential by  
\[ \mathbf{E} = -\frac{\partial \mathbf{A}}{\partial t} \]

The vector potential for the field (1.16) is  
\[ \mathbf{A} = \frac{|\mathbf{E}|}{\omega} \sin(\omega t)\mathbf{e}_x \mp \cos(\omega t)\mathbf{e}_y \]  
(1.18)

giving  
\[ \mathbf{J} = \mp \epsilon_0 \frac{\mathbf{e}_z}{2\omega} \int_{\text{beam}} d\mathbf{r} |\mathbf{E}|^2 = \mp \frac{1}{c\omega} \epsilon_z \int_{\text{beam}} d\mathbf{r} I, \]

with \( I \) the optical intensity. We then write  
\[ \int_{\text{beam}} d^3\mathbf{r} I = L \int_{\text{area}} d^2\mathbf{r} I = LP \]

where \( P \) is the optical power and \( L = c\Delta t \) is the effective length of a photon mode. Combining the factors gives  
\[ \mathbf{J} = \mp \frac{P\Delta t}{\omega} \mathbf{e}_z. \]

The energy transported by the beam in a time \( \Delta t \) is  
\[ U = P\Delta t = N\hbar\omega \]

where \( N \) is the number of photons. Thus the angular momentum per photon is  
\[ \frac{\mathbf{J}}{N} = \mp \hbar \mathbf{e}_z. \]
We see that a photon with right (left) hand circular polarization carries angular momentum $-\hbar$ (\hbar) directed along the direction of propagation. This is the spin angular momentum of the photon which is a spin 1 particle. Note that when light is reflected normally from a mirror the handedness of circular polarization changes, but the direction of propagation also changes, so the angular momentum per photon does not change.

### 1.5.2 Jones matrices

♣ to be added

### 1.5.3 Stokes parameters

An arbitrary polarization state depends on three real parameters: the angle $\theta$, and the polarization vector phases $\chi_x, \chi_y$. The Stokes parameters provide a convenient representation. See P & W Appendix 6.B for details. ♣

### 1.6 Fresnel coefficients

Wide optical beams (plane waves) propagating along the direction $k$ are polarized perpendicularly to $k$. When the beam is reflected from an interface we define a plane of incidence containing the incident beam with wavevector $k_i$ and the reflected beam with wavevector $k_r$. When the incident beam is polarized parallel to the plane of incidence we call the beam p polarized, and when the beam is polarized perpendicular to the plane of incidence we call the beam s polarized\(^3\).

The coefficients for transmission and reflection of optical waves at interfaces between different media are called the Fresnel coefficients. These coefficients can be found using the boundary conditions for electromagnetic fields at surfaces separating media with different electric and magnetic properties. The boundary conditions are continuity of the tangential components of $E$ and the normal components of $B$. The presence of surface charges or currents results in discontinuity of the normal component of $D$ or the tangential component of $H$.

For normal incidence the relations connecting the electric fields of the incident, transmitted and reflected waves, $E_i, E_t,$ and $E_r$ are

\[
E_t = \frac{2n_1}{n_1 + n_2} E_i \tag{1.19}
\]
\[
E_r = -\frac{n_2 - n_1}{n_2 + n_1} E_i. \tag{1.20}
\]

Here $n_1$ is the index of the medium containing the incident wave, and $n_2$ is the index of the medium containing the transmitted wave.

The Fresnel coefficients for light incident at angle $\theta_1$ in a medium with index $n_1$ and refracted at angle $\theta_2$ into a medium with index $n_2$ are (angles are measured from the surface

\(^3\)s stands for senkrecht, the German word for perpendicular.
normal)

\begin{align}
    r_s &= -\frac{\sin(\theta_1 - \theta_2)}{\sin(\theta_1 + \theta_2)} \\
    t_s &= \frac{2\cos(\theta_1)\sin(\theta_2)}{\sin(\theta_1 + \theta_2)} \\
    r_p &= -\frac{\tan(\theta_1 - \theta_2)}{\tan(\theta_1 + \theta_2)} \\
    t_p &= \frac{2\cos(\theta_1)\sin(\theta_2)}{\sin(\theta_1 + \theta_2)\cos(\theta_1 - \theta_2)}.
\end{align}

The subscripts \(s\) and \(p\) refer to polarization perpendicular and parallel to the plane of incidence respectively.

If we eliminate \(\theta_2\) using Snell’s law \(n_1 \sin \theta_1 = n_2 \sin \theta_2\) we get

\begin{align}
    r_s &= \frac{\cos(\theta_1) - \sqrt{n_2^2 - \sin^2(\theta_1)}}{\cos(\theta_1) + \sqrt{n_2^2 - \sin^2(\theta_1)}} \\
    t_s &= \frac{2\cos(\theta_1)}{\cos(\theta_1) + \sqrt{n_2^2 - \sin^2(\theta_1)}} \\
    r_p &= -\frac{n_2^2 \cos(\theta_1) - \sqrt{n_2^2 - \sin^2(\theta_1)}}{n_2^2 \cos(\theta_1) + \sqrt{n_2^2 - \sin^2(\theta_1)}} \\
    t_p &= \frac{2n \cos(\theta_1)}{n_2^2 \cos(\theta_1) + \sqrt{n_2^2 - \sin^2(\theta_1)}}.
\end{align}

where we have introduced the relative index \(n = n_2/n_1\).

The fraction of the incident power that is reflected from an interface is called the reflectance and is given by \(R_s = |r_s|^2\) or \(R_p = |r_p|^2\) for \(s\) or \(p\) polarizations. Energy conservation requires that the transmittance is given by \(T_s = 1 - R_s\) or \(T_p = 1 - R_p\). It is important to note that \(T_s \neq |t_s|^2\), \(T_p \neq |t_p|^2\). Instead we have the relations

\begin{align}
    |r_s|^2 + \frac{n_2 \cos(\theta_2)}{n_1 \cos(\theta_1)} |t_s|^2 &= 1, \\
    |r_p|^2 + \frac{n_2 \cos(\theta_2)}{n_1 \cos(\theta_1)} |t_p|^2 &= 1.
\end{align}

These expressions account for the dependence of intensity on refractive index and on the change of beam cross section at an interface proportional to \(1/\cos \theta\). It can be verified that Eqs. (1.21) satisfy these relations. Only when \(n_2 = n_1\) do we get the simpler relation \(|r|^2 + |t|^2 = 1\).

It is useful to define the Fresnel coefficients for light propagating backwards from medium
2 to medium 1. These are

\[ r'_s = \frac{\sin(\theta_1 - \theta_2)}{\sin(\theta_1 + \theta_2)} \]  
\[ r'_p = \frac{\tan(\theta_1 - \theta_2)}{\tan(\theta_1 + \theta_2)} \]  
\[ t'_s = \frac{2 \cos(\theta_2) \sin(\theta_1)}{\sin(\theta_1 + \theta_2)} \]  
\[ t'_p = \frac{2 \cos(\theta_2) \sin(\theta_1)}{\sin(\theta_1 + \theta_2) \cos(\theta_1 - \theta_2)} \]  

We see that \( r'_s = -r_s, r'_p = -r_p \) while \( t_s, t'_s \) and \( t_p, t'_p \) are related by angle dependent factors. It can be verified that the following mixed conservation relations are satisfied for arbitrary angles

\[ |r_s|^2 + t_s t'_s = 1, \]  
\[ |r_p|^2 + t_p t'_p = 1. \]

These energy conservation relations will be useful in analysis of Fabry-Perot resonators.

1.7 The beamsplitter

A generic beam splitter is shown in Fig. 1.2. There are two input ports (0 and 1) and two output ports (2 and 3). Without making any assumptions about the internal structure of the beamsplitter, apart from the requirement that it is lossless, the transmission and reflection coefficients must satisfy the following relations (Stokes, 1849)

\[ |r'| = |r|, \]  
\[ |t'| = |t|, \]  
\[ |r|^2 + |t|^2 = |r'|^2 + |t'|^2 = 1, \]  
\[ r t'^* + r'^* t = 0. \]  

Figure 1.2: The beamsplitter.
In this context the Fresnel coefficients $r, t, r', t'$ refer to the effective behavior of the beamsplitter for a given state of polarization. Relations (1.26) are thus separately valid for both s and p polarizations. It is important to keep in mind that Eqs. (1.26) describe a situation where the index of refraction is the same at all external ports of the beamsplitter. If this were not the case, then we would not find e.g. $|t'| = |t|$.

These reciprocity relations can be derived as follows. The incident intensity is

$$I_{\text{in}} = (\mathcal{E}_0, \mathcal{E}_1)(\mathcal{E}_0, \mathcal{E}_1)\dagger = |\mathcal{E}_0|^2 + |\mathcal{E}_1|^2 = I_0 + I_1.$$  

The output fields are

$$\begin{pmatrix} \mathcal{E}_2 \\ \mathcal{E}_3 \end{pmatrix} = \begin{pmatrix} t' & r \\ r' & t \end{pmatrix} \begin{pmatrix} \mathcal{E}_0 \\ \mathcal{E}_1 \end{pmatrix},$$

and the output intensity is

$$I_{\text{out}} = (\mathcal{E}_2, \mathcal{E}_3)(\mathcal{E}_2, \mathcal{E}_3)\dagger = (|r'|^2 + |t'|^2)I_0 + (|r|^2 + |t|^2)I_1 + (rt^* + r'^*t)\mathcal{E}_0^*\mathcal{E}_1 + (r^*t' + r't^*)\mathcal{E}_0\mathcal{E}_1^*.$$  

Conservation of intensity for arbitrary input fields gives the conditions

$$|r|^2 + |t|^2 = 1,$$  

$$|r'|^2 + |t'|^2 = 1,$$  

$$rt^* + r'^*t = 0.$$  

The additional conditions $|r| = |r'|$ and $|t| = |t'|$ follow from these relations.

We could also derive Eqs. (1.26) from the formal requirement that a lossless beamsplitter results in a unitary transformation of the fields. The scattering matrix of the beamsplitter is

$$S = \begin{pmatrix} t' & r \\ r' & t \end{pmatrix}$$

and unitarity means $SS\dagger = S^\dagger S = I$. These conditions result in

$$|r|^2 + |t|^2 = 1, \quad |r'|^2 + |t'|^2 = 1, \quad |r|^2 + |t|^2 = 1, \quad |r'|^2 + |t'|^2 = 1,$$  

$$rt^* + r'^*t = 0, \quad rt^* + r'^*t = 0.$$  

The relations (1.28a) are equivalent to (1.26a -1.26c). The second equality in (1.28b) looks different, but is equivalent to (1.26d). To see this write $r = |r|e^{i\phi_r}, r' = |r'|e^{i\phi_r}, t = |t|e^{i\phi_t}, t' = |t'|e^{i\phi_t'}$. Using the amplitude relations $|r| = |r'|, |t| = |t'|$ Eqs. (1.28b) both lead to

$$\phi_r + \phi'_r - \phi_t - \phi'_t = (2m + 1)\pi$$

with $m$ an integer. Different phase choices are possible. For example we can set $\phi_t = \phi'_t = 0$ so $t = t'$ and both quantities are positive. With this choice it follows that $\phi_r + \phi'_r = (2m + 1)\pi$, which gives $r = |r|e^{i\phi_r}$ and $r' = -|r|e^{-i\phi_r}$. A common choice for calculations is to take $\phi_r = 0$ so $r > 0, r' = -r$, and $t = t' > 0$.

The beamsplitter has some remarkable properties when used with weak optical fields containing single photons. Suppose a single photon is incident at port 0. For a 50/50
beamsplitter with $|r|^2 = |t|^2 = 1/2$ the photon will emerge at port 2 half the time and at port 3 half the time. Similarly if the photon is incident at port 1 it will emerge at port 2 half the time and at port 3 half the time. Based on these probabilities we might expect that if two photons were incident, one at port 0 and one at port 1 then half the time we would observe one photon at each output port, and a quarter of the time we would observe two photons at port 2 and a quarter of the time two photons at port 3.

However, what is observed is that two photons emerge at port 2 half the time and two photons emerge at port 3 half the time, but we never see one photon at each output port. This is known as the Hong-Ou-Mandel effect after the people who first demonstrated it experimentally. The reason is that for two identical photons we must add the amplitudes for the different reflection and transmission paths and these amplitudes cancel out for a lossless beamsplitter for the case when one photon exits each port.

### 1.7.1 Brewster’s angle

Light polarized in the plane of incidence (p-polarization) experiences zero reflectance and perfect transmission at Brewster’s angle. This can be seen from the geometrical construction in Fig. 1.3. The incident light interacts with the second medium to drive an oscillating dipole which radiates into the reflected and transmitted beams. The oscillating dipole is parallel to the electric field in the second medium, and there must be a non-zero component perpendicular to the $k$ vector of the reflected light in order to observe a non-zero reflection. When $\theta_i + \theta_t = \pi/2$ the oscillating dipole is parallel to $k_r$ and the reflection vanishes.

The zero reflection condition is thus $\theta_t = \pi/2 - \theta_i$ or, using Snell’s law,

$$n_i \sin(\theta_i) = n_t \sin(\theta_t) = n_t \sin(\pi/2 - \theta_i) = n_t \cos(\theta_i)$$

and

$$\theta_i = \tan^{-1}(n_t/n_i) = \theta_B.$$ 

For an air-glass interface $n_i = 1, n_t = 1.5$ we have $\theta_B = 56.3$ deg.

Light that is s-polarized has an electric field component that is perpendicular to $k_r$ for all incident angles, so there is no Brewster’s angle effect for this polarization. Repeated passage of a light beam through a stack of plates at Brewster’s angle will attenuate the s-polarized light while transmitting all of the p-polarized light. This can be used to create a highly polarized beam. Brewster’s angle is also widely used in gain cells in laser cavities as a way of obtaining a very low loss interface for p-polarized light.

### 1.7.2 Total internal reflection

Light is internally reflected on the hypotenuse of the prism shown in Fig. 1.4. The incidence angle is $\theta_1 = 45 \text{ deg}$. The refractive index of the prism is $n_1$ and outside the prism $n_2 = 1$. Let us find the condition on $n_1$ such that 100% of the light energy is reflected. This is referred to as total internal reflection (TIR). From Snell’s law $n_1 \sin \theta_1 = n_1/\sqrt{2} = n_2 \sin \theta_2$. Thus $\theta_2 = \pi/2$ results in $n_1 = \sqrt{2}n_2 = \sqrt{2}$ which marks the onset of TIR at this incidence angle.

The Fresnel coefficients at the onset of TIR are $r_s = 1, t_s = 2, r_p = -1, t_p = 2\sqrt{2}$. The $t_p$ and $t_s$ coefficients have the values found from Eqs. (1.22) but they do not imply that
energy is transmitted into the second region since $\theta_2 = \pi/2$. There is only an evanescent wave beyond the interface.

An interesting and useful feature of TIR is that the reflected s and p polarization components experience a differential phase shift which can be used to convert the polarization state of the light. We can calculate this phase shift as follows. The critical angle beyond which TIR occurs is $\theta_c = \sin^{-1}(n_2/n_1)$. The Fresnel reflection coefficient for s polarization is from (1.22a)

$$ r_s = \frac{\cos(\theta_1) - \sqrt{n_2^2 - \sin^2(\theta_1)}}{\cos(\theta_1) + \sqrt{n_2^2 - \sin^2(\theta_1)}} $$

with $n = n_2/n_1 < 1$. When $\theta_1 > \theta_c$, $\sin \theta_1 > n$ and we can write this as

$$ r_s = \frac{\cos(\theta_1) - i\sqrt{\sin^2(\theta_1) - n^2}}{\cos(\theta_1) + i\sqrt{\sin^2(\theta_1) - n^2}}. $$

This is an expression of the form $r_s = (a - ib)/(a + ib) = e^{-i2\tan^{-1}(b/a)}$ which has unit modulus. For p polarization we have

$$ r_p = \frac{-n^2 \cos(\theta_1) - \sqrt{n^2 - \sin^2(\theta_1)}}{n^2 \cos(\theta_1) + \sqrt{n^2 - \sin^2(\theta_1)}} = \frac{-n^2 \cos(\theta_1) - i\sqrt{\sin^2(\theta_1) - n^2}}{n^2 \cos(\theta_1) + i\sqrt{\sin^2(\theta_1) - n^2}}. $$

Figure 1.4: TIR at 45 deg. incidence angle with $n_1 > n_2$. 

1 Optical waves
1.7 The beamsplitter

The relative phase shift of the p and s components is thus

$$\delta \phi = \phi_p - \phi_s = -2 \left[ \tan^{-1} \left( \frac{\sqrt{\sin^2(\theta_1) - n^2}}{\cos(\theta_1)} \right) - \tan^{-1} \left( \frac{\sqrt{\sin^2(\theta_1) - n^2}}{n^2 \cos(\theta_1)} \right) \right].$$

Let’s look at this for a glass prism in air so $n = 1/1.5 = 0.667$. Figure 1.5 shows the differential phase as a function of incidence angle. We see that for $\theta_1 \sim 54$ deg, the phase shift is very close to $\pi/4$ and changes by only a few degrees for changes in the refractive index which correspond to hundreds of nm worth of dispersion in glass. Thus two internal bounces will give a $\pi/2$ shift which will convert linearly polarized light to circular. This forms the basis for an achromatic polarization conversion device, known as the Fresnel rhomb, after Fresnel who invented it.

1.7.3 Reflection from metals

An ideal metallic conductor is a perfect reflector since the electric field vanishes inside the conductor. For real metals with finite conductivity the reflectivity is less than unity and is accompanied by a phase shift.

Let us write the refractive index as $\tilde{n} = n + i\kappa$. A plane wave propagates as

$$\mathcal{E}(z) \sim e^{ikz} = e^{ik_0\tilde{n}z} = e^{ik_0nz}e^{-k_0\kappa z}.$$  

We see that the attenuation depends on $\kappa$ leading to an exponential intensity loss $I(z) = I(0)e^{-\alpha z}$ with absorption coefficient $\alpha = 2k_0\kappa$.

Snell’s law and the expressions for Fresnel reflection coefficients remain valid with complex
Equations (1.22) take the form

\[ r_s = \frac{\cos(\theta_1) - \sqrt{\tilde{n}^2 - \sin^2(\theta_1)}}{\cos(\theta_1) + \sqrt{\tilde{n}^2 - \sin^2(\theta_1)}} \]  
(1.29a)

\[ t_s = \frac{2\cos(\theta_1)}{\cos(\theta_1) + \sqrt{\tilde{n}^2 - \sin^2(\theta_1)}} \]  
(1.29b)

\[ r_p = -\frac{\tilde{n}^2 \cos(\theta_1) - \sqrt{\tilde{n}^2 - \sin^2(\theta_1)}}{\tilde{n}^2 \cos(\theta_1) + \sqrt{\tilde{n}^2 - \sin^2(\theta_1)}} \]  
(1.29c)

\[ t_p = \frac{2\tilde{n}\cos(\theta_1)}{\tilde{n}^2 \cos(\theta_1) + \sqrt{\tilde{n}^2 - \sin^2(\theta_1)}} \]  
(1.29d)

There is still a Brewster’s angle with a minimum in the reflectivity for p-polarization, although the reflectivity does not go to zero.

Let’s evaluate the reflectivity for the simplest case of normal incidence, \( \theta_1 = 0 \). We have

\[ r_s = \frac{1 - \sqrt{\tilde{n}^2}}{1 + \sqrt{\tilde{n}^2}} = 1 - \frac{\tilde{n}}{1 + \tilde{n}} \]  
(1.30a)

\[ r_p = \frac{\tilde{n}^2 - \sqrt{\tilde{n}^2}}{\tilde{n}^2 + \sqrt{\tilde{n}^2}} = \frac{\tilde{n} - 1}{\tilde{n} + 1} = \frac{1 - \tilde{n}}{1 + \tilde{n}} \]  
(1.30b)

We see \( r_s = r_p \) as expected at normal incidence. We can separate the real and imaginary parts as

\[ r_s = \frac{(1 - \tilde{n})(1 + \tilde{n}^*)}{(1 + \tilde{n})(1 + \tilde{n}^*)} = \frac{1 - |\tilde{n}|^2 - 2i\text{Im}(\tilde{n})}{1 + |\tilde{n}|^2 + 2\text{Re}(\tilde{n})} = \frac{1 - n^2 - \kappa^2 - 2i\kappa}{1 + n^2 + \kappa^2 + 2n}. \]

Silver at visible wavelengths has \( n = 0.13 \) and \( \kappa = 4.05 \) which gives \( |r_s|^2 = 0.971 \) and \( \text{arg}(r_s) = -2.66 = -152 \deg \). The attenuation coefficient at \( \lambda = 0.5 \mu \) is

\[ \alpha = 2k_0\kappa = 1.0 \times 10^8 \text{ m}^{-1} = 100 \mu\text{m}^{-1}. \]

The intensity is attenuated by a factor of \( e^{-100} = 6. \times 10^{-45} \) at a distance of \( 1 \mu\text{m} \) below the surface.

### 1.7.4 Calcite beam displacer

Displacer length \( L \), and internal angle \( \alpha \). Transverse displacement is

\[ d = L \tan \alpha = L \frac{(n_e^2 - n_0^2) \tan \theta}{n_e^2 + n_0^2 \tan^2 \theta} \]

where \( \theta \) is the angle of the optical axis.
Chapter 2

Interference

The coherence properties of optical fields can be studied using interferometers which rely on interference between light waves. Let us start by considering several basic situations. When observing interference we usually use an intensity detector such as photographic film, or a CCD camera or our eyes. The intensity pattern shows the vector interference so if $E_1 = E_{1x} \epsilon_x + E_{1y} \epsilon_y$, $E_2 = E_{2x} \epsilon_x + E_{2y} \epsilon_y$, the total field is $E = E_1 + E_2$ and the intensity is

$$I \sim E \cdot E = (E_{1x} + E_{2x})^2 + (E_{1y} + E_{2y})^2.$$

In terms of the complex amplitudes introduced earlier $E_1 = (\frac{E_1}{\epsilon} + c.c.)$, $E_2 = (\frac{E_2}{\epsilon} + c.c.)$, $\epsilon_1, \epsilon_2$ are unit polarization vectors, and the intensity is

$$I = \frac{\epsilon_0 c}{2} \left( |\mathcal{E}_{1x} + \mathcal{E}_{2x}|^2 + |\mathcal{E}_{1y} + \mathcal{E}_{2y}|^2 \right)$$

$$= \frac{\epsilon_0 c}{2} \left[ |\mathcal{E}_{1x}|^2 + |\mathcal{E}_{2x}|^2 + |\mathcal{E}_{1y}|^2 + |\mathcal{E}_{2y}|^2 \right] + \frac{\epsilon_0 c}{2} \left[ (\mathcal{E}_{1x}^* \mathcal{E}_{2x}^* + \mathcal{E}_{1x}^* \mathcal{E}_{2x}) + (\mathcal{E}_{1y} \mathcal{E}_{2y}^* + \mathcal{E}_{1y} \mathcal{E}_{2y}) \right].$$

If the two fields are orthogonally polarized then the interference term, which depends on the relative phases of the fields, vanishes.

2.1 Interference of two beams

2.1.1 Plane wave interference

Consider two scalar fields propagating in the $x - z$ plane at angles $\pm \theta/2$ from the $\epsilon_z$ axis. Explicitly $k_{1,2} = k[\pm \sin(\theta/2) \epsilon_x + \cos(\theta/2) \epsilon_z]$ with $k = 2\pi/\lambda$. The total electric field is

$$\mathcal{E} \sim \mathcal{E}_1 e^{ik \cdot r} + \mathcal{E}_2 e^{ik \cdot r} \sim \mathcal{E}_1 e^{ik[\sin(\theta/2)x + \cos(\theta/2)z]} + \mathcal{E}_2 e^{i[k[-\sin(\theta/2)x + \cos(\theta/2)z]}.$$ 

At a fixed longitudinal position, say $z = 0$, the intensity is

$$I \sim |\mathcal{E}_1 e^{ik \sin(\theta/2)x} + \mathcal{E}_2 e^{-ik \sin(\theta/2)x}|^2$$

$$= |\mathcal{E}_1|^2 + |\mathcal{E}_2|^2 + \mathcal{E}_1 \mathcal{E}_2^* e^{i2k \sin(\theta/2)x} + \mathcal{E}_1^* \mathcal{E}_2 e^{-i2k \sin(\theta/2)x}.$$

Introduce a relative phase of the amplitudes by $\mathcal{E}_2^* = |\mathcal{E}_2| e^{i\phi}$ so

$$I \sim |\mathcal{E}_1|^2 + |\mathcal{E}_2|^2 + 2|\mathcal{E}_1||\mathcal{E}_2| \cos[2k \sin(\theta/2)x + \phi].$$

March 8, 2016 M. Saffman
Maxima in the intensity are separated by the period $\Lambda$ with

$$\Lambda = \frac{\lambda}{2 \sin(\theta/2)} \sim \frac{\lambda}{\theta},$$

where the last approximation holds for small $\theta$.

### 2.1.2 Interference of Plane and Spherical Waves

An outgoing spherical wave has electric field

$$E(\mathbf{r}, t) = E \frac{\cos(k \cdot \mathbf{r} - \omega t)}{f(r)} \epsilon(k).$$

The polarization $\epsilon(k)$ depends on the propagation direction $k$. Energy conservation for a wave expanding into $4\pi$ steradians requires that $f(r) = r = |\mathbf{r}|$. Using complex amplitudes the wave is

$$E(\mathbf{r}, t) = E e^{i(k \cdot \mathbf{r} - \omega t)} r \epsilon(k) + c.c. .$$

Note that the symbol $E$ has units of field times length. For an outgoing spherical wave we define $k$ and $r$ with respect to the same origin so $k \cdot r = kr$ and

$$E(\mathbf{r}, t) = E e^{i(kr - \omega t)} r \epsilon(k) + c.c. .$$

The interference of a plane wave $E_1$ and a spherical wave $E_2$ gives an intensity

$$I \sim |E_1|^2 + \frac{|E_2|^2}{r^2} + \left[ \frac{E_1 E_2^*}{r} e^{i(k_1 \cdot \mathbf{r} - kr)} \epsilon_1 \cdot \epsilon_2^* + c.c. \right].$$

Working in the $x-z$ plane we put $k_1 = k[\sin(\theta/2)\epsilon_x + \cos(\theta/2)\epsilon_z]$, $r = \sqrt{x^2 + z^2}$ and

$$I \sim |E_1|^2 + \frac{|E_2|^2}{r^2} + \left[ \frac{E_1 E_2^*}{r} e^{i[k_1(\sin(\theta/2)x + \cos(\theta/2)z - \sqrt{x^2 + z^2})]} \epsilon_1 \cdot \epsilon_2^* + c.c. \right].$$

At a fixed longitudinal position, say $z = L$, the intensity is

$$I \sim |E_1|^2 + \frac{|E_2|^2}{x^2 + L^2} + \left[ \frac{E_1 E_2^*}{\sqrt{x^2 + L^2}} e^{i[k(\sin(\theta/2)x + \cos(\theta/2)L - \sqrt{x^2 + L^2})]} \epsilon_1 \cdot \epsilon_2^* + c.c. \right].$$

We can simplify this by assuming a small angular range so that $\epsilon_1 \cdot \epsilon_2^*(k_2) \sim \text{constant}$, and set the constant equal to unity. A small angular range implies that $x \ll L$. Expanding to second order in $x$ we get

$$I \sim |E_1|^2 + \frac{|E_2|^2}{L^2} + \left[ \frac{E_1 E_2^*}{L} e^{i[k(\cos(\theta/2) - 1)L - \sqrt{x^2 + L^2}]} \epsilon_1 \cdot \epsilon_2^* + c.c. \right].$$

We have neglected the dependence of the amplitude, but not the phase, on $x^2$. Let us measure $x$ along the direction perpendicular to $k_1$ which corresponds to putting $\theta = 0$. The intensity simplifies to

$$I \sim |E_1|^2 + \frac{|E_2|^2}{L^2} + \left[ \frac{E_1 E_2^*}{L} e^{-i\frac{L^2}{2\pi}} + c.c. \right].$$
The interference term is then proportional to \( \cos(kx^2/2L) \). Neighboring intensity maxima are found for
\[
\frac{kx_2^2}{2L} - \frac{kx_1^2}{2L} = 2\pi
\]
which we solve for the period
\[
\Lambda_{21} = \frac{\lambda L}{\bar{x}}.
\]
Here \( \Lambda_{21} = x_2 - x_1 \) is the distance between neighboring maxima, and \( \bar{x} = (x_2 + x_1)/2 \) is the average \( x \) position. Away from the origin the fringe period decreases \( \sim 1/\bar{x} \), i.e. the interference is chirped. The first maximum is at \( \Lambda_{21} = x_{\text{max}}, \bar{x} = x_{\text{max}}/2 \) so \( x_{\text{max}} = \sqrt{2\lambda L} \) and the first intensity minimum is at \( x_{\text{min}} = \sqrt{\lambda L} \). When we get to a detailed discussion of diffraction phenomena we will encounter this type of scaling often, and will refer to a length scale \( \sim \sqrt{\lambda L} \) as diffractive scaling.

### 2.1.3 Interference of two spherical waves

Consider two spherical waves with one originating at \( r = 0 \) and the other one originating at \( R = d \epsilon_z \). The total field is
\[
E = \frac{\mathcal{E}_1}{2} e^{ikr_1} \epsilon_1 + \frac{\mathcal{E}_2}{2} e^{ikr_2} \epsilon_2 + \text{c.c.}.
\]
We will again make a small angle approximation and put \( \epsilon_1 \cdot \epsilon_2 = 1 \), \( r_1 \simeq L + \frac{x^2}{2\pi} \), \( r_2 \simeq (L-d) + \frac{x^2}{2(L-d)} \) to get
\[
I \sim \frac{|\mathcal{E}_1|^2}{L^2} + \frac{|\mathcal{E}_2|^2}{(L-d)^2} + \left[ \frac{\mathcal{E}_1 \mathcal{E}_2^*}{L(L-d)} e^{ikd} e^{ik\left[\frac{x^2}{2\pi} - \frac{x^2}{2(L-d)}\right]} + \text{c.c.} \right].
\]
The interference term at \( z = L \gg x \) is proportional to \( \cos \left[ kd + kx^2 \left( \frac{1}{2\pi} - \frac{1}{2(L-d)} \right) \right] \). We again have a chirped interference pattern with the fringe period decreasing like \( 1/x \). Taking \( kd = q2\pi \), with \( q \) an integer, the first minimum is at \( x_{\text{min}} = \sqrt{\frac{\lambda(L-d)}{d}} \) and the first maximum is at \( x_{\text{max}} = \sqrt{2x_{\text{min}}} \). The interference is azimuthally symmetric about the \( z \) axis so we get a “bull’s eye” pattern.

### 2.1.4 Multibeam interference

Sharper interference fringes can be obtained when more than two beams simultaneously interfere. ◆ details to be added

### 2.2 Fabry-Pérot Interferometer

An optical cavity can be used to enhance the intensity of an incident beam and has useful frequency dependent reflection and transmission properties. The presence of a gain medium inside a resonant cavity leads to laser operation. An interferometer is an instrument that

March 8, 2016 M. Saffman
uses interference from surfaces for analyzing optical properties of light. There are several different types of interferometer including Fabry-Pérot, Michelson, Mach-Zehnder, Fizeau, and Sagnac. We will start with a discussion of the Fabry-Pérot geometry.

Consider a general linear cavity with an input mirror, an output mirror, and internal losses. The amplitude of the internal field after transmission through the input coupler is $E_0 = t_{in}E_{in}$ where the intensity transmission of the input coupler is $T_{in} = |t_{in}|^2$. After each cavity round trip the internal field must be multiplied by a complex factor $g = |g|e^{i\phi(\nu)}$. The magnitude of $g$ accounts for coupling and internal losses and the phase shift depends on the optical frequency $\nu$ and the cavity parameters.

We will express the amplitude and phase as

$$|g| = \sqrt{R_{in}R_{out}R_{cav}} = \sqrt{R_{in}R_{out}(1 - L_{cav})},$$

$$e^{i\phi(\nu)} = e^{i\phi_0}e^{i\phi_1(\nu)} \frac{r_{in}'}{r_{out}'}.$$

Here $R_{cav}$ accounts for the round trip internal losses neglecting the input and output mirrors. For example if the medium inside the cavity has (intensity) absorption coefficient $\alpha$ then $R_{cav} = e^{-\alpha d}$, where $d$ is the length of the cavity. The primed coefficients refer to the case where the field is incident on a mirror from inside the cavity. The static phase accounts for
any interfaces inside the cavity, and the frequency dependent term is
\[ \phi_1(\nu) = k n(\nu) 2d = \frac{4\pi d}{c} \nu n(\nu) \]
where we have allowed for a possible dispersion of the index of the intracavity medium.

The total internal field due to the sum of the circulating fields is then
\[ E_{\text{cav}} = E_0 + E_1 + E_2 + ..., \]
where \( E_{j+1} = g E_j \). Thus
\[ E_{\text{cav}} = \sum_{j=0}^{\infty} E_0 g^j = E_0 \frac{1}{1 - g} = E_{\text{in}} \frac{t_{\text{in}}}{1 - g}. \] (2.1)

Referring to Fig. 2.1 \( E_{\text{cav}} \) is the field just to the right of the input mirror that is propagating to the right.

It is instructive to derive Eq. (2.1) in a different way that does not involve summing an infinite series of circulating waves. Looking at Fig. 2.1 we see that the internal field to the right of the input mirror is given by the sum of the transmitted input field plus the reflected cavity field. We can write \( E_{\text{cav}} = t_{\text{in}} E_{\text{in}} + E_{\text{cav}} \sqrt{R_{\text{in}} R_{\text{cav}} R_{\text{out}}} e^{i\phi} \) and solving for \( E_0 \) we find
\[ E_{\text{cav}} = E_{\text{in}} \frac{t_{\text{in}}}{1 - \sqrt{R_{\text{in}} R_{\text{cav}} R_{\text{out}}} e^{i\phi}} = E_{\text{in}} \frac{t_{\text{in}}}{1 - g} \]
which is the same as (2.1).

### 2.2.1 Transmitted intensity

With the cavity field known we can write the transmitted field as
\[ E_t = E_{\text{cav}} \sqrt{(1 + R_{\text{cav}})/2} t_{\text{out}}' = E_{\text{in}} t_{\text{in}} t_{\text{out}}' \sqrt{(1 + R_{\text{cav}})/2} \frac{1}{1 - g}, \] (2.2)
from which we find for the transmitted intensity
\[ \frac{I_t}{I_{\text{in}}} = T_{\text{in}} T_{\text{out}} ((1 + R_{\text{cav}})/2) \left| \frac{1}{1 - g} \right|^2, \]
\[ = \frac{T_{\text{in}} T_{\text{out}} ((1 + R_{\text{cav}})/2)}{1 + R} \left[ 1 - \frac{2 \sqrt{R}}{1 + R} \cos[\phi_0 - \theta_{\text{in}} - \theta_{\text{out}} + 4\pi \nu n(\nu)/c] \right]^{-1}, \] (2.3)

where \( R = |g|^2 = R_{\text{in}} R_{\text{out}} R_{\text{cav}} \). Assuming that the internal cavity phase and the phase shifts due to mirror reflections cancel we can write
\[ \frac{I_t}{I_{\text{in}}} = \frac{T_{\text{in}} T_{\text{out}} ((1 + R_{\text{cav}})/2)}{(1 - \sqrt{R})^2} \left[ 1 + \frac{4 \sqrt{R}}{(1 - \sqrt{R})^2} \sin^2(2\pi \nu n(\nu)/c) \right]^{-1}, \] (2.4)
which is the well known Airy formula showing a sequence of periodic maxima in the transmitted intensity as can be seen in Fig. 2.2. Note that when there are no internal losses, \( R_{\text{cav}} = 1 \), and the mirrors are matched, \( R_{\text{in}} = R_{\text{out}} \), the peak cavity transmission is unity.

March 8, 2016 M. Saffman
2.2.2 Reflected intensity

The reflected field is the sum of the field directly reflected from the input mirror and the cavity field that is transmitted back through the input mirror which gives

\[ E_r = E_{in}r_{in} + E_{cav}g t'_{in}/r'_{in} \]
\[ = E_{in}r_{in} \left( 1 + \frac{t'_{in}g}{r'_{in}} \right) \frac{1 - g}{1 - g} \]
\[ = E_{in}r_{in} \frac{R_{in} - g}{R_{in}}. \] (2.5)

In passing from the second to the third lines we have used \( r_{in}r'_{in} = -R_{in} \) and, from Eqs. (1.25), \( t'_{in} = 1 - R_{in} \). Note that the interference between the directly reflected and cavity leakage beams, and hence the magnitude of the field, is independent of the mirror phase \( \theta_r \).

A short calculation gives for the reflected intensity

\[ \frac{I_r}{I_{in}} = \frac{R_{in} + \mathcal{R}/R_{in} - 2\mathcal{R}^{1/2} \cos[4\pi d\nu\nu/\nu]/c}{1 + \mathcal{R} - 2\mathcal{R}^{1/2} \cos[4\pi d\nu\nu/\nu]/c}. \]

Using \( \cos 2x = 1 - 2\sin^2 x \) we can write this as

\[ \frac{I_r}{I_{in}} = \frac{R_{in} \left( 1 - \mathcal{R}^{1/2}/R_{in} \right)^2 + 4(\mathcal{R}^{1/2}/R_{in}) \sin^2[2\pi d\nu\nu/\nu]/c}{(1 - \mathcal{R}^{1/2})^2 + 4\mathcal{R}^{1/2} \sin^2[2\pi d\nu\nu/\nu]/c}. \] (2.6)

Note that \( |t_{in}| \neq |t'_{in}| \). This does not contradict the beamsplitter relations found in Sec. 1.7 where the index of refraction was the same on both sides of the beamsplitter, which is not generally true for the Fabry-Perot resonator.

Figure 2.2: Transmission (solid line) and reflectivity (dashed line) of a Fabry-Perot cavity with \( R_{in} = 0.9 \), \( R_{out} = 0.8 \), and \( R_{cav} = 0.99 \). The inset zooms in near the resonance.
We see that for $R_{\text{in}} = 1$ the reflected intensity is equal to the incident intensity, independent of frequency, as it must be. When $R_{\text{in}} = \mathcal{R}^{1/2}$ or, equivalently, $R_{\text{in}} = R_{\text{cav}} R_{\text{out}}$ the cavity is “impedance matched” and the reflectivity is exactly zero on resonance. Put another way, impedance matching occurs when the input mirror transmission $T_{\text{in}} = 1 - R_{\text{in}} = 1 - R_{\text{cav}} R_{\text{out}} = \mathcal{L}$, where $\mathcal{L}$ is the cavity round trip loss excluding the input coupler. Note that provided $R_{\text{cav}} = 1$ impedance matching occurs for $R_{\text{out}} = R_{\text{in}}$. For larger or smaller $R_{\text{in}}$ the cavity is over or under coupled and the reflectivity never vanishes.

An interesting case is the symmetric cavity with $R_{\text{in}} = R_{\text{out}}$ which implies $R_{\text{in}} = R_{\text{cav}} R_{\text{out}}$ and

$$\frac{I_r}{I_{\text{in}}} = \frac{R_{\text{in}} \left(1 - R_{\text{cav}}^{1/2}\right)^2 + 4R_{\text{cav}}^{1/2} \sin^2[2\pi d \nu n(\nu)/c]}{\left(1 - R_{\text{in}} R_{\text{cav}}^{1/2}\right)^2 + 4R_{\text{in}} R_{\text{cav}}^{1/2} \sin^2[2\pi d \nu n(\nu)/c]}.$$  \hspace{1cm} (2.7)

For a lossless cavity, $R_{\text{cav}} = 1$, and this last expression reduces to

$$\frac{I_r}{I_{\text{in}}} = \frac{\sin^2[2\pi d \nu n(\nu)/c]}{\left(1 - R_{\text{in}}^{1/2}\right)^2 + \sin^2[2\pi d \nu n(\nu)/c]}.$$  \hspace{1cm} (2.8)

The symmetric lossless cavity is impedance matched so the reflected intensity vanishes on resonance, in agreement with the discussion above.

### 2.2.3 Finesse

It is customary to write the transmitted intensity as

$$\frac{I_t}{I_{\text{in}}} = T_{\text{in}} T_{\text{out}} \left(\frac{(1 + R_{\text{cav}})/2}{\mathcal{R}^{1/2}}\right)^2 \frac{1}{1 + \left(\frac{2\mathcal{F}}{\pi}\right)^2 \sin^2(2\pi d \nu n(\nu)/c)},$$  \hspace{1cm} (2.9)

where the cavity finesse is given by

$$\mathcal{F} = \frac{\pi R_{\text{cav}}^{1/4}}{1 - \mathcal{R}^{1/2}}.$$  \hspace{1cm} (2.10)

The reason for this choice of definition will be apparent in a moment.\(^2\)

There are periodic maxima in the transmission at frequencies

$$\nu_N n(\nu_N) = \frac{c}{2d} N$$

with $N$ an integer. The free spectral range, neglecting dispersion in the refractive index, is $\nu_{\text{FSR}} = \nu_{N+1} - \nu_N = c/(2dn)$. Defining the full width at half maximum of the transmission peaks as $\nu_{\text{FWHM}}$ we find that

$$\nu_{\text{FWHM}} = \frac{c}{\pi d n} \sin^{-1}\left(\frac{\pi}{2\mathcal{F}}\right).$$

\(^2\)It is also possible to invert (2.10) for $\mathcal{R}(\mathcal{F}) = (2\mathcal{F}^4 + 4\pi^2 \mathcal{F}^2 + \pi^4 - \pi \sqrt{16\mathcal{F}^6 + 20\pi^2 \mathcal{F}^4 + 8\pi^4 \mathcal{F}^2 + \pi^6})/(2\mathcal{F}^4)$ and use this in (2.4) to express $I_t$ as a function of the finesse. This gives an unwieldy expression which is not convenient to work with.
so that when the finesse is large we have

\[ \nu_{FWMH} \approx \frac{\nu_{FSR}}{F}. \]  

(2.11)

We see that defining the finesse by Eq. (2.10) we get that the ratio between the longitudinal mode spacing and the full width of each mode is just the finesse.

Using the expressions for the finesse and the free spectral range resonance width we can write the transmitted and reflected intensity as

\[
\frac{I_t}{I_{in}} = \frac{T_{in} T_{out} ((1 + R_{cav})/2)}{(1 - R)^{1/2}} \frac{1}{1 + \left(\frac{2F}{\pi}\right)^2 \sin^2(\pi \nu/\nu_{FSR})},
\]

(2.12a)

\[
\frac{I_r}{I_{in}} = \frac{R_{in} \left(1 - R^{1/2}/R_{in}\right)}{(1 - R^{1/2})^2} + \left(\frac{2F}{\pi}\right)^2 \sin^2[\pi \nu/\nu_{FSR}] \frac{1}{1 + \left(\frac{2F}{\pi}\right)^2 \sin^2(\pi \nu/\nu_{FSR})}.
\]

(2.12b)

These equations are exact for any cavity reflectivities and losses with the definition \(R = R_{in} R_{out} R_{cav}\).

**Photon lifetime**

We can relate the FWHM of the cavity resonance to the photon lifetime inside the cavity as follows. Suppose in steady state the intensity inside the cavity is \(I_{cav}\) and the input field is suddenly turned off. The intensity will decay exponentially as \(I_{cav}(t) = I_{cav}(0)e^{-\gamma_{cav}t}\), where \(\gamma_{cav}\) is the cavity decay rate. The change in intensity in one round trip is

\[
\frac{\Delta I_{cav}}{\Delta t} = \frac{I_{cav} R - I_{cav}}{2dn/c} = -I_{cav} \nu_{FSR} \left[1 - (1 - T_{in})(1 - T_{out})(1 - L_{cav})\right].
\]

In the high finesse limit we have \(F \approx 2\pi/(T_{in} + T_{out} + L_{cav})\) so that

\[
\frac{dI_{cav}}{dt} \approx -I_{cav} \frac{2\pi \nu_{FSR}}{F}
\]
from which we see that

$$\gamma_{\text{cav}} \approx \frac{2 \pi \nu_{\text{FSR}}}{\mathcal{F}} \approx 2 \pi \nu_{\text{FWHM}} \equiv \omega_{\text{FWHM}}.$$  \hfill (2.13)

In other words the rate of energy decay from the cavity given by $\gamma_{\text{cav}}$ is equal to $\omega_{\text{FWHM}}$ which is the width of the cavity resonance expressed in angular units.

The cavity decay rate affords an interpretation of the finesse in terms of the number of photon round trips in one cavity lifetime $\tau_{\text{cav}} = 1/\gamma_{\text{cav}}$. Let us say that in a time $\tau_{\text{cav}}$ a photon makes $N$ cavity round trips. We have

$$N = \frac{\tau_{\text{cav}}}{(2dn/c)} = \tau_{\text{cav}} \nu_{\text{FSR}} = \tau_{\text{cav}} \mathcal{F} \nu_{\text{FWHM}} = \tau_{\text{cav}} \mathcal{F} \frac{\omega_{\text{FWHM}}}{2\pi} = \frac{\mathcal{F}}{2\pi}.$$  

We can therefore express the finesse as

$$\mathcal{F} = 2\pi N.$$  

Another useful quantity is the cavity build up factor which is the ratio of the on-resonance intensity inside the cavity to the input intensity. We have

$$I_{\text{cav, max}} = I_{\text{in}} \frac{T_{\text{in}}}{(1-\sqrt{R})^2} \approx I_{\text{in}} \frac{T_{\text{in}} \mathcal{F}^2}{\pi^2} \approx I_{\text{in}} \frac{4T_{\text{in}}}{(T_{\text{in}} + T_{\text{out}} + \mathcal{L}_{\text{cav}})^2} = I_{\text{in}} \frac{4}{T_{\text{in}} (1 + T_{\text{out}}/T_{\text{in}} + \mathcal{L}_{\text{cav}}/T_{\text{in}})^2},$$

where the last lines hold in the high finesse limit. Another useful relation is that when the cavity is impedance matched the buildup factor is $I_{\text{cav, max}}/I_{\text{in}} \sim \mathcal{F}/\pi$. Figure 2.3 shows the finesse and build up factor as a function of the internal losses for fixed input coupler transmission.

A quantity that is easily measured experimentally is the ratio of the maximum to minimum transmitted intensity as the cavity is scanned through resonance. This is given by

$$\frac{I_{t, \text{max}}}{I_{t, \text{min}}} = 1 + \left(\frac{2\mathcal{F}}{\pi}\right)^2.$$  \hfill (2.14)

Note that if the finesse is very high the ratio becomes large and difficult to determine. In such a case it is more convenient to measure the cavity bandwidth in order to determine the finesse.

### 2.2.4 Phase shifts

The phase shift of the transmitted beam with respect to the input beam is

$$\theta_t = \arg \left( \frac{t_{\text{in}} t'_{\text{out}}}{1 - g} \right).$$

March 8, 2016 M. Saffman
Figure 2.4: Transmission (solid line) and reflection (dashed line) phase shifts for $R_{\text{in}} = R_{\text{out}} = 0.95$, $R_{\text{cav}} = 0.99$. The finesse is $\mathcal{F} = 55.8$.

A little algebra gives the expression

$$\theta_t = \tan^{-1}\left( \frac{f \sin(2\pi\nu/\nu_{\text{FSR}})}{1 - f \cos(2\pi\nu/\nu_{\text{FSR}})} \right)$$

(2.15)

$$\approx \tan^{-1}\left( \frac{\sin(2\pi\nu/\nu_{\text{FSR}})}{1 - \cos(2\pi\nu/\nu_{\text{FSR}})} \right)$$

(2.16)

where $f = 1 + \frac{\pi^2}{2\mathcal{F}^2} - \frac{\pi}{\mathcal{F}} \sqrt{1 + \frac{\pi^2}{4\mathcal{F}^2}}$ and the second line holds in the high finesse limit.

The phase shift of the reflected beam is $\theta_r = \arg\left( r_{\text{in}} \frac{R_{\text{in}} - g}{1 - g} \right)$ which gives the expression

$$\theta_r = \tan^{-1}\left( \frac{(R_{\text{in}} - 1) f \sin(2\pi\nu/\nu_{\text{FSR}})}{R_{\text{in}} + f^2 - (1 + R_{\text{in}}) f \cos(2\pi\nu/\nu_{\text{FSR}})} \right)$$

(2.17)

$$\approx \tan^{-1}\left( \frac{(R_{\text{in}} - 1) \sin(2\pi\nu/\nu_{\text{FSR}})}{(1 + R_{\text{in}}) [1 - \cos(2\pi\nu/\nu_{\text{FSR}})]} \right)$$

(2.18)

where the second line is again applicable in the high finesse limit.

Figure 2.4 shows the phase shifts in transmission and reflection. The phase is an odd function of the detuning from cavity resonance and can be used to “lock” a laser beam to a cavity resonance. This is commonly done using the Pound-Drever-Hall method (1983). Although either transmission or reflection can be used the reflected phase has a larger slope, as can be seen from the inset in the figure, which provides for a more precise lock.

2.2.5 Finite beam size effects

If the beam has limited transverse size diffractive spreading will limit the attainable finesse irrespective of the mirror and cavity losses. This may be a limiting factor in the case of a planar mirror Fabry-Perot.
2.2 Fabry-Pérot Interferometer

We can estimate the necessary beam size to obtain a given finesse as follows. In the limit of low cavity loss and high finesse the effective number of cavity round trips is roughly the inverse of the single roundtrip loss, or $\mathcal{F}/(2\pi)$. The corresponding propagation distance is $L = 2d\mathcal{F}/(2\pi)$. If the input beam has waist radius $w_0$ then after propagation it will spread to $w = w_0\sqrt{1 + L^2/L_R^2}$, where $L_R = \pi w_0^2/\lambda$. Requiring that $w/w_0 \sim 1$ gives the condition $L^2 < L_R^2$ or

$$w_0 > \frac{\sqrt{\pi d\mathcal{F}}}{\pi}. \quad (2.19)$$

For example a 5 cm long cavity with 1% losses so $\mathcal{F} \sim 600$ operated at $\lambda = 1$ $\mu$m needs $w_0 > 1.7$ mm.

### 2.2.6 Sensitivity to pressure

A Fabry-Perot cavity with air inside is sensitive to pressure fluctuations due to the finite refractive index of air. In the visible part of the spectrum air at atmospheric pressure has refractive index $n \simeq 1 + 3 \times 10^{-4}$. In a cavity with length $d$ the air induced phase shift in one round trip is

$$\delta\phi = 2dk(n - 1).$$

The refractive index is proportional to density which is proportional to pressure so the pressure dependent shift is

$$\delta\phi = 2dk\frac{P}{P_{\text{atm}}}(n - 1)$$

where $P$ is the actual pressure and $P_{\text{atm}}$ is atmospheric pressure.

Using $\lambda = .5$ $\mu$m, $d = 0.1$ m, and a pressure of $10^{-6}$ mbar $\sim 10^{-9}$ of atmospheric, which corresponds to a good vacuum, but not UHV, gives $\delta\phi = 7.5 \times 10^{-7}$. A phase shift of $2\pi$ changes the cavity frequency by one FSR or $c/2d$. Thus the frequency shift at this pressure compared to the vacuum frequency is

$$\delta\nu = \frac{\delta\phi c}{2\pi 2d} = \frac{P}{P_{\text{atm}}}(n - 1)\nu = 180 \text{ Hz}.$$ 

Note that if the cavity were at atmospheric pressure we would be very sensitive to acoustic noise. In acoustics 0 dB, which is more or less the threshold of human perception, is defined as $2 \times 10^{-4}$ microbar. A soft whisper may be $+40$ dB or $2 \times 10^{-6}$ bar. Thus if the cavity were at atmospheric pressure and subjected to a 40 dB disturbance we would get a frequency shift of

$$\frac{P}{P_{\text{atm}}}(n - 1)\nu = (2 \times 10^{-6})(n - 1)\nu = 360 \text{ kHz}.$$ 

### 2.2.7 Determination of mode matching coefficient

The overlap of an incident beam with a resonant cavity mode can be measured as follows. Assume the intensity incident on the resonator is $I_{\text{in}} = I_0 + I_\perp$ with $I_0$ the mode matched intensity and $I_\perp$ the intensity in orthogonal modes. The field reflected from the input mirror is approximately $E_r = r_{\text{in}}E_0 + E_\perp + r_{\text{in}}E_{\perp}$. Here $E_\perp$ is the modematched intracavity field
transmitted through the input mirror. We neglect the non modematched intracavity field since it is much weaker than the mode matched part in a high finesse cavity.

At a cavity resonance the sum $r \text{in} E_0 + E_t$ leads to a reflected intensity for the mode matched light given by Eq. (2.6) as

$$I_0 \times \frac{1}{R_{\text{in}}} \frac{(R_{\text{in}} - \sqrt{R})^2}{(1 - \sqrt{R})^2} = m I_{\text{in}} \times \frac{1}{R_{\text{in}}} \frac{(R_{\text{in}} - \sqrt{R})^2}{(1 - \sqrt{R})^2}.$$

The directly reflected non modematched light gives a reflected intensity

$$I_{\perp} \times R_{\text{in}} = (1 - m) I_{\text{in}} R_{\text{in}}.$$

The coefficient $m = I_0/I_{\text{in}}$ is the fraction of the intensity that is modematched to the cavity. Thus the observed reflected intensity at resonance will be given by

$$\frac{I_r}{I_{\text{in}}} = m \frac{1}{R_{\text{in}}} \frac{(R_{\text{in}} - \sqrt{R})^2}{(1 - \sqrt{R})^2} + (1 - m) R_{\text{in}}. \quad (2.20)$$

For an impedance matched cavity $R_{\text{in}} = \sqrt{R}$, i.e. $R_{\text{in}} = R_{\text{cav}} R_{\text{out}}$ so the input transmission equals the cavity losses ($\mathcal{L} = 1 - R_{\text{cav}} R_{\text{out}}$) excluding the input coupler. In this case the reflected intensity vanishes on resonance when $m = 1$ but is finite for imperfect modematching.

To measure $m$ we need three measurements. First determine $R_{\text{in}}$ by measuring the input coupler reflectivity with the intracavity field blocked. Then measure the cavity finesse which determines $\mathcal{R}$, and finally measure the reflection dip $I_r/I_{\text{in}}$ and use (2.20) to determine $m$. 
2.3 Etalon

A short, fixed separation Fabry-Perot is referred to as an etalon. The etalon is useful for spectral filtering of a light beam. A solid etalon is inherently stable consisting of a piece of transparent optical material with highly polished and parallel faces. If the faces are separated by a distance $d$, the material has an index $n$ and the surfaces are coated with reflectivity $R$ then the free spectral range is

$$\nu_{FSR} = \frac{c}{2nd}$$

and the transmission function from Eq. (2.12) is

$$\frac{I_t}{I_{in}} = \frac{T_{in}T_{out}((1 + R_{cav})/2)}{(1 - R^{1/2})^2} \frac{1}{1 + \left(\frac{2\pi}{\nu}\right)^2 \sin^2(\pi\nu/\nu_{FSR})}$$

with the definition $R = R_{in}R_{out}R_{cav}$.

The contrast between the maximum and minimum transmitted intensity is

$$q = \left(\frac{2\mathcal{F}}{\pi}\right)^2.$$ 

A moderate finesse of 100 gives a contrast of 4050 or 36 dB. Say we use fused silica of index $n = 1.45$ and thickness $d = 1.\text{mm}$ then $\nu_{FSR} = 103$ (GHz) and the transmission bandwidth is $\nu_{FWHM} = \nu_{FSR}/\mathcal{F} = 1.03$ GHz. Neglecting loss in the fused silica ($R_{cav} = 1$) and $R_{in} = R_{out} = R$ we have $\mathcal{F} = \pi R^{1/2}/(1 - R)$ so $\mathcal{F} = 100$ implies $R = 0.97$. The peak transmission is then

$$I_{t,\max} = \frac{(1 - R)^2}{(1 - R)^2} = 1.$$ 

The detuning $\delta\nu$ at which the transmission has fallen to 0.99 of the peak is found from

$$\frac{1}{1 + \left(\frac{2\pi}{\nu}\right)^2 \sin^2(\pi\delta\nu/\nu_{FSR})} = 0.99$$

which has the approximate solution $\delta\nu = 0.00049\nu_{FSR} = 50.4$ MHz.

An interesting question is whether or not a solid etalon is sufficiently stable to maintain the resonance to 50 MHz so that the intensity varies by less than 1%. Let the temperature change by an amount $\delta T$ giving a change $\delta\nu_{FSR}$ then the detuning relative to an etalon peak is

$$\delta\nu = -\nu\frac{\delta\nu_{FSR}}{\nu_{FSR}} = -\frac{c}{\lambda}\frac{\delta\nu_{FSR}}{\nu_{FSR}}.$$ 

The change in the free spectral range is

$$\delta\nu_{FSR} = -\nu_{FSR}\left(\frac{1}{n}\frac{dn}{dT} + \frac{1}{d}\frac{dd}{dT}\right)\delta T.$$ 

At room temperature and 587 nm wavelength fused silica has$^3$

$$\frac{dn}{dT} = 0.87 \times 10^{-5}.$$ 

The coefficient of thermal expansion is

$$\frac{1}{d} \frac{dd}{dT} = 0.55 \times 10^{-6}$$

per degree C. Thus for a 1°C change in temperature which is probably typical for a quiet laboratory we find

$$\delta \nu_{\text{FSR}} = -\nu_{\text{FSR}} \times 6.6 \times 10^{-6}$$

so for a 0.5 µm wavelength

$$\delta \nu = \frac{c}{\lambda} 6.6 \times 10^{-6} = 4.0 \text{ GHz}.$$

If the etalon is stabilized to 1 mK we get a drift of 4 MHz.

We can reduce the requirement on temperature stability by using a thinner etalon. Say we take $d = 0.25$ mm which is commercially available as a standard component. Then $\nu_{\text{FSR}} = 414 \text{ GHz}$, $\nu_{\text{FWHM}} = 4.1 \text{ GHz}$, and the frequency error giving 99% transmission is 202 MHz which corresponds to a temperature drift of 50 mK.

### 2.3.1 Tilted etalon

Consider propagation through an etalon plate of index $n$ and thickness $d$ as in Fig. 2.5. At normal incidence the phase shift between the beam at the entrance point $(x, z) = (0, 0)$ and the exit point $(0, d)$ is $\phi = knd$, $k = 2\pi/\lambda$. There will be a transmission resonance when $2\phi = 2\pi m$ with $m$ an integer. This gives a wavelength

$$\lambda_0 = \frac{2nd}{m}.$$

When the beam is incident at angle $\theta$ the phase shift between the same entrance and exit points is $\phi = nk' \cdot r = knd \cos \theta'$, with $\theta'$ the angle from the normal inside the slab. The wavelength at a transmission resonance is

$$\lambda = \frac{2nd \cos \theta'}{m}. \quad (2.22)$$
It may seem odd that the phase shift crossing the tilted etalon gets smaller, not larger, since the thickness apparently increases as we tilt the etalon. However, following a ray across the tilted etalon also gives a transverse displacement, whereas we should compare the phase at the same entrance and exit points as in the case of normal incidence. Doing so gives Eq. (2.22).

Writing (2.22) in terms of the input angle $\theta$ gives

$$\lambda = \frac{2d}{m} \left[ n^2 - \sin^2(\theta) \right]^{1/2}$$

$$= \lambda_0 \left[ n^2 - \sin^2(\theta) \right]^{1/2} \frac{1}{n}$$

This equation can be used to predict the decrease of the resonant transmission wavelength $\lambda_0$ through an interference filter as $\theta$ is increased.

### 2.4 Fourier transform spectroscopy

The Michelson interferometer can be used to measure the spectral content of an optical beam. When the input light has a single spectral component $I(\lambda)$ and the two arms have a path length difference $p$, as shown in Fig. 2.6 the output of the interferometer is

$$I(\lambda, p) = \frac{I(\lambda)}{2} [1 + \cos(kp)].$$

Here $p = 2(d_2 - d_1)$ is the path length difference between the arms. We could equivalently use the time delay variable $\tau = p/c$ and the angular frequency $\omega = 2\pi c/\lambda$ to write

$$I(\omega, \tau) = \frac{I(\omega)}{2} [1 + \cos(\omega \tau)].$$

![Image of Fourier transform spectroscopy with a Michelson interferometer. Moving one of the mirrors changes the differential path length $p$ leading interference at the output port.](image-url)

March 8, 2016 M. Saffman
Following the discussion in Sec. 1.3.1 we will work with a single sided spectral intensity distribution $\tilde{I}(\omega)$ with $\omega \geq 0$. The total intensity at fixed $\tau$ is then

$$I(\tau) = \int_0^\infty d\omega \tilde{I}(\omega, \tau) = \frac{1}{2} \int_0^\infty d\omega \tilde{I}(\omega) + \frac{1}{2} \int_0^\infty d\omega \tilde{I}(\omega) \cos(\omega \tau).$$

At zero delay

$$I(0) = \int_0^\infty d\omega \tilde{I}(\omega)$$

and

$$I(\tau) - I(0) = \sqrt{\frac{2}{\pi}} \int_0^\infty d\omega \frac{\pi^{1/2} \tilde{I}(\omega)}{2^{3/2}} \cos(\omega \tau).$$

Taking the inverse cosine transform we find the spectral intensity from

$$\tilde{I}(\omega) = \frac{4}{\pi} \int_0^\infty d\tau \left[ I(\tau) - \frac{I(0)}{2} \right] \cos(\omega \tau).$$

If we record $I(\tau) = I(p/c)$ for different path length differences $p$ we can calculate the spectral distribution $\tilde{I}(\omega)$ by taking the inverse cosine transform. This is the basis of Fourier transform spectroscopy (FTS). FTS is often more efficient than using a scanning spectrometer since a larger intensity is incident on the detector which reduces the influence of photon shot noise and detector noise. Figure 2.7 shows an example of the output signal for a two frequency input.

An important practical question is how large a scan range is needed to resolve a frequency interval $\delta \omega$. From the usual Fourier uncertainty relations we know that $\delta \omega \sim 1/\delta \tau$. Thus
2.5 Multilayer coatings

The utility of the Fabry-Perot resonator, and many other optical instruments, is dependent on the availability of mirrors with very high reflectivity. The technology for making high reflectivity mirrors has advanced to the point where more than five nines of reflectivity are commercially available. Such mirrors are fabricated using multilayer dielectric coatings. As is the case with lens systems for high performance imaging the design of multilayer coatings is based on computerized numerical calculations. We can nevertheless understand some of the basic ideas from analytical considerations.

2.5.1 Quarterwave anti-reflection coating

We will first consider the task of producing a low reflectivity or “anti-reflection” coating on a high index substrate. Such coatings are needed for lenses used in imaging systems to avoid ghosting from undesired multiple reflections. The geometry is shown in Fig. 2.8. We know from our analysis of the Fabry-Perot interferometer that the reflected field should be found by adding up an infinite number of partial reflections. We can nonetheless gain a qualitative understanding of the behavior by considering the first two reflected waves. There is the external reflection from the incident beam, $E_{r1} = E_{in} r_{01}$ and the first reflection from the substrate, $E_{r2} = E_{in} t_{01} r_{12} t_{10} e^{i 2d k n_1}$. The notation for the Fresnel coefficients uses $r_{ij}$ for the reflection inside medium $i$ with the second medium labeled $j$ and $t_{ij}$ the transmission from
medium $i$ to medium $j$. This notation will be useful below in the treatment of multilayer films.

Qualitatively we expect the reflection to be minimized if $E_{r1}$ and $E_{r2}$ interfere destructively. Assume a phase convention where $t_{01}, t_{10}$ are both positive. We then require that $\frac{r_{12}}{r_{01}} e^{2dn_1}$ is negative. At normal incidence the Fresnel reflection amplitudes are negative when the second medium has a higher index. If we set $n_2 > n_1 > n_0$ then we will get destructive interference when $2dn_1 = \pi$ or $d = \lambda/(4n_1)$. In order to match the magnitude of $E_{r1}$ and $E_{r2}$ we need to choose the film index $n_1$ judiciously.

At normal incidence $r_{ij} = \frac{n_i - n_j}{n_i + n_j}$. If $n_i, n_j$ are similar in magnitude then $|r_{ij}|$ will be small and $|t_{ij}|$ will be close to unity. This suggests the condition $r_{01} = r_{12}$ or

$$n_1 = \sqrt{n_0 n_2}.$$ 

We see that a quarter wave film ($d = \lambda/4n_1$) with $n_1$ the geometric mean of the external and substrate indices can be expected to minimize the reflection coefficient. To use this idea for air the external medium and a glass substrate we need $n_1 \sim \sqrt{1.5} = 1.22$. This condition is approximately met by MgF$_2$ which has an index of 1.38.

A quantitative evaluation of the reflected field can be found using Eq.(2.5) which becomes in the notation of this section

$$E_r = E_{in} \frac{r_{01} R_{01} - g}{R_{01}}$$ 

with

$$g = \sqrt{R_{01} R_{12}} e^{i\phi_0} e^{i\phi_1(\lambda)} \frac{r_{12}}{|r_{10}| |r_{12}|}$$

and $\phi_1 = 2dn_1/\lambda$. Setting $\phi_0 = 0, \phi_1 = \pi$ we get $g = \sqrt{R_{01} R_{12}}$ and

$$E_r = E_{in} \frac{n_0 n_2 - n_1^2}{n_0 n_2 + n_1^2} \tag{2.24}$$

Perhaps surprisingly the condition for canceling the reflection, $n_1 = \sqrt{n_0 n_2}$, which we found from a simplified consideration of only the first two reflected waves, is borne out by a full analysis. The reason why this condition works perfectly is that it corresponds to an impedance matched cavity. Recall the Fabry-Perot impedance matching condition when there are no internal losses is simply $R_{in} = R_{out}$ which coincides with $n_1 = \sqrt{n_0 n_2}$.

Using a MgF$_2$ layer $n_1 = 1.38$ on a glass substrate $n_2 = 1.52$ we find an intensity reflectivity of

$$\frac{I_r}{I_{in}} = 0.014$$

which is considerably less than that of uncoated glass which has $\frac{I_r}{I_{in}} = (\frac{1.5 - 1}{1.5 + 1})^2 = 0.04$.

The attentive reader may notice one peculiar point. In our analysis of the Fabry-Perot resonator we found that transmission resonances occur for $2dn = 2\pi N$ with $N$ an integer. Thus the first resonance is at $d = \lambda/2$. Here the first resonance is at $d = \lambda/4$. Why do we have an apparent discrepancy?
2.5 Multilayer coatings

Figure 2.9: Reflectance of quarter wave film with index $n_1$ in between air ($n_0 = 1$) and glass ($n_2 = 1.5$).

### 2.5.2 Quarterwave high-reflection coating

We now wish to enhance the reflectivity of the substrate. From our study of the Fabry-Perot resonator we know that the transmitted intensity never goes exactly to zero. We therefore do not expect to be able to make a perfect reflector. Nevertheless extremely high performance can be achieved with a multilayer coating.

The most basic idea uses the quarter-wave antireflection coating but with the low index film replaced by a high index film. With $n_1 > n_2 > n_0$ the reflection coefficient $r_{12}$ becomes positive and instead of destructive interference we get constructive interference and enhanced reflectivity. We now have $g = -\sqrt{R_{01}R_{12}}$ but Eq. (2.24) remains unchanged giving

$$E_r = E_{in} \frac{n_0 n_2 - n_1^2}{n_0 n_2 + n_1^2}.$$  

For example using TiO$_2$ which has $n = 2.30$ we get $I_r/I_{in} = 0.31$. As can be seen in Fig. 2.9 very high reflectance requires large $n_1$. Unfortunately there are no convenient transparent materials available which motivates the use of multilayer films as will be discussed in the next section.

### 2.5.3 Multilayer coatings

High performance antireflection or highreflection films rely on multilayer coatings. We can analyze these cases using a matrix formalism. Consider the geometry in Fig. 2.10. Each layer has index $n_j$ and thickness $d_j$. The fields propagating to the right and the left at the left hand side of layer $j$ are labeled $E_{jr}, E_{jl}$. The fields at the right hand side of the same layer are $\tilde{E}_{jr}, \tilde{E}_{jl}$ and they are related by

$$
\begin{pmatrix}
E_{jr} \\
E_{jl}
\end{pmatrix}
= 
\begin{pmatrix}
e^{-i\phi_j} & 0 \\
0 & e^{i\phi_j}
\end{pmatrix}
\begin{pmatrix}
\tilde{E}_{jr} \\
\tilde{E}_{jl}
\end{pmatrix},
$$
with $\phi_j = kn_jd_j$ and $k = 2\pi/\lambda_{\text{vac}}$. We can also write this in matrix notation as $\mathbf{E}_j = \Phi_j \tilde{\mathbf{E}}_j$ where $\mathbf{E}_j = \begin{pmatrix} E_{jr} \\ E_{jl} \end{pmatrix}$. Concentrating on the interface between layers $j$ and $k$ we have

$$E_{kr} = t_{jk} \tilde{E}_{jr} + r_{kj} E_{kl},$$

$$\tilde{E}_{jl} = r_{jk} \tilde{E}_{jr} + t_{kj} E_{kl}.$$  

Solving for the $\tilde{E}$ we find

$$\tilde{E}_{jr} = \frac{1}{t_{jk}} E_{kr} - \frac{r_{kj}}{t_{jk}} E_{kl},$$

$$\tilde{E}_{jl} = \frac{r_{jk}}{t_{jk}} E_{kr} + \frac{t_{kj}t_{jk} - r_{kj}r_{jk}}{t_{jk}} E_{kl}.$$  

Using the Stokes relation $t_{kj}t_{jk} - r_{kj}r_{jk} = 1$ we can write the solution as $\tilde{\mathbf{E}}_j = \mathbf{G}_{jk} \mathbf{E}_k$ where

$$\mathbf{G}_{jk} = \frac{1}{t_{jk}} \begin{pmatrix} 1 & r_{jk} \\ r_{jk} & 1 \end{pmatrix}.$$  

Combining with the $\Phi_j$ matrix we have for the transformation between two layers

$$\mathbf{E}_j = \Phi_j \mathbf{G}_{jk} \mathbf{E}_k = \mathbf{M}_{jk} \mathbf{E}_k$$  \hspace{1cm} (2.25)  

with

$$\mathbf{M}_{jk} = \frac{1}{t_{jk}} \begin{pmatrix} e^{-i\phi_j} & r_{jk}e^{-i\phi_j} \\ r_{jk}e^{i\phi_j} & e^{i\phi_j} \end{pmatrix}.$$  \hspace{1cm} (2.26)  

This can be straightforwardly extended to an arbitrary number of layers simply by multiplying together the matrices for each layer.

Let’s consider the most general situation where there is an external beam incident from the left ($E_{0r} \neq 0$), and also a beam incident from the right at the last layer labeled $N$ ($E_{Nl} \neq 0$). If the total transformation matrix is $\mathbf{M}_{0N} = \begin{pmatrix} M_{11} & M_{12} \\ M_{21} & M_{22} \end{pmatrix}$ then we can solve for
2.5 Multilayer coatings

the output fields to find

\[ E_{ol} = \frac{M_{21}}{M_{11}} E_{0r} + \frac{M_{11} M_{22} - M_{12} M_{21}}{M_{11}} E_{Nl}, \]  
\[ E_{Nr} = \frac{1}{M_{11}} E_{0r} - \frac{M_{12}}{M_{11}} E_{Nl}. \]  

Equations (2.27) provide a compact solution to any multilayer interference problem.

As an example let’s check the previous results for a single \( \lambda/4 \) film with incident medium 0, quarter wave layer 1, and substrate 2. The transformation between layers 0 and 2 is

\[ E_0 = M E_2 \]

with

\[ M = M_{01} M_{12} = \frac{1}{t_0 t_1} \left( \frac{e^{-i(\phi_0 + \phi_1)}}{e^{i(\phi_0 - \phi_1)}} \left( 1 + e^{i2\phi_1} r_{01} r_{12} \right) \right) \left( 1 + e^{-i2\phi_1} r_{01} r_{12} \right). \]

We take \( \phi_0 = 0 \) since we define the incident fields as those at the surface of the film, \( \phi_1 = \pi/2, \)
\( r_{01} = \frac{n_0 - n_1}{n_0 + n_1}, \)
\( r_{12} = \frac{n_1 - n_2}{n_1 + n_2}, \)
\( t_0 = \frac{2n_0}{n_0 + n_1}, \)
\( t_1 = \frac{2n_1}{n_1 + n_2}. \)

Plugging in we get

\[ M = \frac{i}{2n_0 n_1} \begin{pmatrix} -n_0 n_2 - n_1^2 & n_0 n_2 - n_1^2 \\ -n_0 n_2 + n_1^2 & n_0 n_2 + n_1^2 \end{pmatrix}. \]

With the boundary condition on the right hand side that \( E_{2l} = 0 \) we get the solutions

\[ r \equiv \frac{E_{ol}}{E_{0r}} = \frac{C}{A} = \frac{n_0 n_2 - n_1^2}{n_0 n_2 + n_1^2} \]
\[ t \equiv \frac{E_{2r}}{E_{0r}} = \frac{1}{A} = i \frac{2n_0 n_1}{n_0 n_2 + n_1^2}. \]

The expression for \( r \) agrees with (2.24) and \( |r|^2 + \frac{n_0^2}{n_0^2}|t|^2 = 1 \) agrees with (1.23).

2.5.4 Multilayer high reflectivity coating

As an example of the use of Eqs. (2.27) let’s calculate the reflectivity for a stack of alternating high and low index \( \lambda/4 \) layers as shown in Fig. 2.11. A pair of \( \lambda/4 \) layers with indices \( n_L, n_H \) give

\[ M_L = \frac{i}{2n_L} \begin{pmatrix} -n_H - n_L & n_H - n_L \\ n_L - n_H & n_H + n_L \end{pmatrix}, \quad M_H = \frac{i}{2n_H} \begin{pmatrix} -n_H - n_L & n_L - n_H \\ n_H - n_L & n_L + n_H \end{pmatrix}. \]
and the two-layer segment $M_L M_H$ gives

$$M_s = M_L M_H = \frac{-1}{2n_H n_L} \begin{pmatrix} n_H^2 + n_L^2 & n_H n_L^2 - n_H^2 \\ n_H n_L^2 - n_H^2 & n_H^2 + n_L^2 \end{pmatrix}. $$

Note that $\det(M_L M_H) = 1$. The matrix for $N$ such segments is $M_N = (M_L M_H)^N$. To calculate this we can use a result known as Sylvester’s theorem. This says that for matrices with $\det(M) = 1$ the matrix $M_N = M^N$ is given by

$$M_N = \frac{1}{\sin(\theta)} \begin{pmatrix} M_{11} \sin(N\theta) - \sin[(N-1)\theta] & M_{12} \sin(N\theta) \\ M_{21} \sin(N\theta) & M_{22} \sin(N\theta) - \sin[(N-1)\theta] \end{pmatrix}$$

with $\cos(\theta) = (M_{11} + M_{22})/2$.

For a large number of layers we can neglect edge effects due to air on the left and a glass substrate on the right. In this limit the intensity reflectance is

$$R_N = \left( \frac{M_{N,21}}{M_{N,11}} \right)^2 = \frac{M_{21}^2 \sin^2(N\theta)}{(M_{11} \sin(N\theta) - \sin[(N-1)\theta])^2}.$$

Let us use MgF$_2$ with $n_L = 1.38$ and TiO$_2$ with $n_H = 2.30$. For $N = 2, 4, 6, 8, 10$ corresponding to 4, 8, 12, 16, 20 layers we find $R = 0.5937, 0.9350, 0.9913, 0.9989, 0.9999$.

We can also derive a closed form expression for $R_N$ in terms of $n_H, n_L$ as follows. The segment matrix $M_s = M_L M_H$ is a real symmetric matrix. It can be diagonalized by an orthogonal matrix $P$ constructed from the eigenvectors of $M_s$. To do this we calculate the eigenvectors of $M_s$ which are $\frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ -1 \end{pmatrix}$ and use them to construct

$$P = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 1 \\ -1 & 1 \end{pmatrix}$$

and

$$P^{-1} = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix}. $$

We can then diagonalize $M_s$ as $D = P M_s P^{-1} = \begin{pmatrix} -n_H/n_L & 0 \\ 0 & -n_L/n_H \end{pmatrix}$ with the eigenvalues of $M_s$ on the diagonal. Doing this we see that $M_s = P^{-1} D P$ and

$$M_N = (M_s)^N = P^{-1} D^N P = \frac{1}{2} \begin{pmatrix} (n_H/n_L)^{2N} + 1 & (n_H/n_L)^{2N} - 1 \\ (n_H/n_L)^{2N} - 1 & (n_H/n_L)^{2N} + 1 \end{pmatrix}. $$

The reflectivity is

$$R_N = \left( \frac{M_{N,21}}{M_{N,11}} \right)^2 = \left[ \frac{(n_H/n_L)^{2N} - 1}{(n_H/n_L)^{2N} + 1} \right]^2.$$

We see that for $n_H > n_L$, $R_N \rightarrow 1$ as $N \rightarrow \infty$.

Various filter morphologies can be constructed with similar methods. Coatings that have bandpass characteristics, specific angular responses, etc. can be synthesized using multiple layers with different thicknesses and indices. As in the case of lens design this subject is an art, but there is much accumulated knowledge available in specialized books on the topic.
Chapter 3

Geometrical optics

The short wavelength limit of the Maxwell equations describes rays that propagate in straight lines in homogeneous media. We will justify this statement in Sec. 3.5. Until then we will explore the effect of planar and curved surfaces on light rays which propagate along straight lines between surfaces.

3.1 Reflection and refraction

Light rays are reflected and refracted at interfaces as seen in Fig. 3.1. On reflection the incident and reflected angles are equal, \( \theta_r = \theta_i \). On refraction we have Snell’s law

\[
n_1 \sin \theta_i = n_2 \sin \theta_t
\]

where the angles are measured from the interface normal. Snell’s law can be derived from the Maxwell equations or from Fermat’s principle of least time.

3.1.1 Paraxial Ray Matrices

Following the path of a ray through multiple surfaces becomes cumbersome due to the sin that appears in Snell’s law. For small angles \( \sin \theta \approx \theta \) and the algebra is greatly simplified.

![Figure 3.1: Reflection and transmission at an interface.](image)
Figure 3.2: Fractional error in approximating \( \sin(\theta) \) by \( \theta \).

This approximation corresponds to a near axial or \textit{paraxial} description. Using a paraxial approximation the propagation of rays is efficiently described using ray matrices.

A useful abstraction is the concept of an optical ray that propagates in a straight line, without diffracting. It is an idealized construction that corresponds to the mathematical limit of \( \lambda \to 0 \). In a plane perpendicular to the optical axis \( z \) a ray is described fully by its transverse coordinate \( x \) and angle with the \( z \) axis \( \theta \). Propagation through a length of homogeneous space \( l \) gives \( x(l) = x(0) + l \tan(\theta(0)) \) and \( \theta(l) = \theta(0) \). If we limit ourselves to rays that make a small angle with respect to the propagation axis then \( \tan \theta \approx \theta \) (this is referred to as paraxial propagation) and the transformation of the ray parameters is

\[
\begin{pmatrix}
  x \\
  \theta
\end{pmatrix}
\rightarrow
\begin{pmatrix}
  x' \\
  \theta'
\end{pmatrix}
= \begin{pmatrix}
  1 & l \\
  0 & 1
\end{pmatrix}
\begin{pmatrix}
  x \\
  \theta
\end{pmatrix}.
\tag{3.1}
\]

If we define a vector \( s = (x, \theta) \) propagation can be written as \( s(l) = M_l s(0) \) with

\[
M_l = \begin{pmatrix}
  A & B \\
  C & D
\end{pmatrix} = \begin{pmatrix}
  1 & l \\
  0 & 1
\end{pmatrix}.
\]

As indicated, the elements of ray matrices are conventionally referred to by the letters \( A, B, C, D \). The error incurred by the paraxial approximation is shown in Fig. 3.2. For angles of not more than 30 deg. from the optical axis the fractional error is less than 5%.

Paraxial propagation through any optical element can be characterized in terms of the \( ABCD \) matrix for that element. For example a thin lens of focal length \( f \) transforms rays as shown in Fig. 3.3. The matrix connecting the ray parameters directly before and after the lens leaves the position \( x \) unchanged so \( A = 1 \) and \( B = 0 \). When \( x = 0 \) the input and output angles are equal so \( D = 1 \). Finally, we see that when \( x = f \theta, \theta' = Cx + \theta = 0 \) so \( C = -1/f \). The thin lens transformation is therefore

\[
\begin{pmatrix}
  x' \\
  \theta'
\end{pmatrix}
= \begin{pmatrix}
  1 & 0 \\
  -1/f & 1
\end{pmatrix}
\begin{pmatrix}
  x \\
  \theta
\end{pmatrix}.
\tag{3.2}
\]

More complex systems are described by a transformation matrix \( M \) found from the matrix product of the individual matrices: \( M = \prod_i M_i \). For future reference the ray matrices for
some common optical elements are given in Fig. 3.4. Note that if the optical element is tilted with respect to the optical axis we need to distinguish between tangential and sagittal rays. Tangential rays propagate in the plane that is normal to the axis about which the optical element is rotated. Sagittal rays lie in the plane containing the axis of rotation. Attention should be drawn to the fact that some authors use a different convention for the ray vector that replaces $\theta$ by $n\theta$ where $n$ is the local refractive index. See the book, A. E. Siegman, Lasers, for details. Our definitions coincide with those in the original papers of Kogelnik and Li from 1965.

The elemental ray matrices have unit determinant, and hence any extended system is described by a matrix of unit determinant\(^1\). There are thus only three independent components of the ray matrix. It is an exceedingly useful result that the ray matrices found from consideration of the propagation of rays, without account of diffraction, can be used to describe the diffractive propagation of light through a wide class of systems[Anan’ev, Siegman]. This is discussed in Sec. 4.1.12.

\[^1\text{This is only true when the input and output planes are in media with the same refractive index. Otherwise the determinant is given by } n_1/n_2.\]
Geometrical optics

Figure 3.4: Ray matrices in the tangential and sagittal planes. The relative refractive index is \( n_r = n_2/n_1 \). In the fourth and fifth lines \( R > 0 \) for the interface concave to the right. The matrices for transmission through the tilted interface can be found in G.A. Massey and A. E. Siegman, Appl. Opt. 8, 975 (1969).

3.1.2 Ray matrix for a concave mirror

As an example of how to derive the ray matrix for a more complicated geometry consider the concave mirror in Fig. 3.5. The center of curvature of the mirror is at \( C \), the object is at \( P \) and the image is formed at \( Q \). The object distance is \( OP = d_o \) and the image distance is \( OQ = d_i \). Using the law of sines we have \( R/\sin(\theta_1) = (d_o - R)/\sin(\theta) \) and \( (R - d_i)/\sin(\theta') = R/\sin(\pi - \theta_2) = R/\sin(\theta_2) \). Reflection on the mirror gives \( \theta' = \theta \) from which it follows that

\[
\sin(\theta) = \frac{d_o - R}{R} \sin(\theta_1) = \frac{R - d_i}{R} \sin(\theta_2)
\]

or

\[
\frac{\sin(\theta_1)}{\sin(\theta_2)} = \frac{R - d_i}{d_o - R}.
\]
3.1 Reflection and refraction

Figure 3.5: Geometry for imaging with a concave mirror.

We then make the paraxial approximation and write this as or

\[
\frac{\theta_1}{\theta_2} = \frac{R - d_i}{d_o - R}.
\]

Then use \(d_o \theta_1 = d_i \theta_2\) so

\[
\frac{d_i}{d_o} = \frac{R - d_i}{d_o - R}.
\]

Rearranging gives \(2/R = 1/d_o + 1/d_i\) and the focal length of the mirror is \(f = R/2\). The ray matrix is thus

\[
M = \begin{pmatrix}
1 & 0 \\
-2/R & 1
\end{pmatrix}.
\]

A derivation of the more complex case where the mirror is tilted with respect to the optical axis so the matrix is different in the tangential and sagittal planes can be found in G. A. Massey and A. E. Siegman, Appl. Opt. 8, 975 (1969).

3.1.3 Imaging with a thin lens

We can use the matrices for linear propagation and a thin lens to derive the formula for imaging with a lens. An imaging system takes rays from an object plane at \(z = 0\) to an image plane at \(z = L\). The geometry is shown in Fig. 3.6. If the lens is at \(z = d_1\) and \(d_2 = L - d_1\) the transformation matrix is

\[
\begin{pmatrix}
x' \\
\theta'
\end{pmatrix} = M_l(d_2)M_{\text{lens}}(f)M_l(d_1)\begin{pmatrix}
x \\
\theta
\end{pmatrix}.
\]

Note that the matrices should be multiplied with the matrix of the first element on the right, progressing to the matrix of the last element on the left. Multiplying out gives a composite matrix

\[
M = \begin{pmatrix}
1 - \frac{d_2}{f} & d_1 + d_2 - \frac{d_1 d_2}{f} \\
-\frac{1}{f} & 1 - \frac{d_1}{f}
\end{pmatrix}.
\]

(3.3)

Image formation means that all rays that start at \(x\) end at the same position \(x'\) irrespective of the value of \(\theta\). This occurs when \(B = 0\) or

\[
\frac{1}{f} = \frac{1}{d_1} + \frac{1}{d_2}.
\]

(3.4)

March 8, 2016 M. Saffman
The magnification is given by \( \frac{x'}{x} = 1 - \frac{d_2}{f} = -\frac{d_2}{d_1} \).

The imaging condition can be written in a different form by introducing the principal foci shown in Fig. 3.6. The principal focus is the point which a parallel bundle of rays propagating along the optical axis are focused to. For a thin lens the principal foci are a distance \( f \) before and after the lens. The distances measured from the principal foci are \( z_1 = d_1 - f \) and \( z_2 = d_2 - f \). Introducing these quantities into (3.4) and rearranging we find

\[
z_1 z_2 = f^2
\]

which is often referred to as the Newtonian form of the lens equation.

### 3.1.4 Negative lens

A lens with index greater than 1 and concave surfaces has a negative focal length. An incident bundle of rays parallel to the optical axis will become divergent after passage through the lens so no real image is formed to the right. Nonetheless there is an image plane containing a virtual image and located to the left of the lens as shown in Fig. ??.

To analyze this situation we put \( f = -|f| \) so that Eq. (3.3) becomes

\[
M = \left( 1 + \frac{d_1}{|f|} \right) \left( \frac{d_1 + d_2 + \frac{d_1 d_2}{|f|}}{1 + \frac{d_1}{|f|}} \right).
\]

The image plane is at

\[
d_2 = -\frac{d_1 |f|}{d_1 + |f|}
\]

which is negative for \( d_1 > 0 \). In other words when the object is in front of the lens an image is formed to the left of the image. This is now a virtual image with positive magnification ...
3.1.5 Compound lens

We can form a compound lens by placing two lenses with focal lengths $f_1, f_2$ next to each other with separation $d$ as shown in Fig. 3.7. Propagation from object to image planes is described by a matrix

$$M = \begin{pmatrix} A & B \\ C & D \end{pmatrix}.$$ 

The matrix elements are

$$A = \frac{d(d_2 - f_2) + f_1 f_2 - d_2 (f_1 + f_2)}{f_1 f_2}, \quad (3.6a)$$

$$B = \frac{d(d_1 - f_1) (d_2 - f_2) + d_2 f_1 f_2 - d_1 [d_2 (f_1 + f_2) - f_1 f_2]}{f_1 f_2}, \quad (3.6b)$$

$$C = \frac{d - f_1 - f_2}{f_1 f_2}, \quad (3.6c)$$

$$D = \frac{d(d_1 - f_1) + f_1 f_2 - d_1(f_1 + f_2)}{f_1 f_2}. \quad (3.6d)$$

Setting $C = -1/f$ we find for the effective focal length of the compound lens

$$f = \frac{f_1 f_2}{f_1 + f_2 - d}. \quad (3.7)$$

Note that when $d = 0$ then $f = f_1 f_2/(f_1 + f_2)$ or $1/f = 1/f_1 + 1/f_2$. Thus adjacent lenses combine like parallel resistors as regards the composite focal length. The focusing power proportional to $1/f$ combines like resistors in series.

The image plane is located at the value of $d_2$ for which the $B$ element of the propagation matrix is zero. Solving $B = 0$ gives

$$d_{2,\text{image}} = \frac{(dd_1 - df_1 - d_1 f_1) f_2}{dd_1 - df_1 - d_1 f_1 - d_1 f_2 + f_1 f_2}.$$ 

When $d_1 = f_1$ then $d_{2,\text{image}} = f_2$, independent of the value of $d$. This is to be expected since when $d_1 = f_1$ the first lens creates a parallel bundle of rays from the object. The parallel bundle is then focused to an image at the principal focus located at $d_2 = f_2$. 

March 8, 2016 M. Saffman
We can also find the image magnification, which is given by the $A$ matrix element. Substituting $d_2 \rightarrow d_{2,\text{image}}$ we find

$$A = \frac{f_1 f_2}{d(d_1 - f_1) + f_1 f_2 - d_1 (f_1 + f_2)}.$$  

Note that $A \neq d_{2,\text{image}}/d_1$ which is given by

$$\frac{d_{2,\text{image}}}{d_1} = \frac{[dd_1 - (d + 2d_1)f_1] f_2}{d_1 [d(d_1 - f_1) + f_1 f_2 - d_1 (f_1 + f_2)]}.$$  

Although the compound lens acts as though it has an effective focal length given by (3.7) the simple relations between focal length, image and object distances, and magnification that described a thin lens no longer apply.

### 3.1.6 Principal Planes

The concept of principal planes allows us to describe a compound lens as an effective thin lens, provided we measure the object and image distances from the principal planes. Let the distances from the principal planes be $z_1, z_2$ as shown in Fig. 3.7, with the principal planes located at distances $s_1, s_2$ from the lenses. We wish to choose the position of the principal planes such that the transformation matrix for the compound lens takes the same form as the transformation matrix for a thin lens, as given by Eq. (3.3), except that $d_1, d_2$ are replaced by $z_1, z_2$. Explicitly we solve

$$\begin{pmatrix} A & B \\ -\frac{1}{f} & D \end{pmatrix} = \begin{pmatrix} 1 - \frac{z_2}{f} & z_1 + z_2 - \frac{z_1 z_2}{f} \\ -\frac{1}{f} & 1 - \frac{z_1}{f} \end{pmatrix},$$

with $z_1 = d_1 - s_1$, $z_2 = d_2 - s_2$. The matrix entries are given above and solving we find

$$s_1 = \frac{df_1}{d - f_1 - f_2}, \quad (3.8a)$$

$$s_2 = \frac{df_2}{d - f_1 - f_2}. \quad (3.8b)$$

Thus the compound lens acts on optical rays as an effective thin lens with the image and object distances measured from the principal planes $s_1, s_2$.

The real utility of this result arises from the fact that it is not limited to a compound lens, but is true for an arbitrary optical transformation described by an ABCD matrix. Surprisingly, any optical system can be described as an effective thin lens, as regards propagation measured from the principal planes. This result greatly simplifies the analysis of complex, multi-element optical systems.

To prove this consider the situation shown in Fig. 3.8. The matrix for propagation between the principal planes is

$$M_{pp-pp} = \begin{pmatrix} 1 & s_2 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} A & B \\ C & D \end{pmatrix} \begin{pmatrix} 1 & s_1 \\ 0 & 1 \end{pmatrix}$$

$$= \begin{pmatrix} A + s_2 C & s_1 A + B + s_1 s_2 C + s_2 D \\ C & s_1 C + D \end{pmatrix}.$$
3.1 Reflection and refraction

Figure 3.8: Principal plane analysis. A composite optical system (top) is equivalent to an effective thin lens with the propagation distances measured to and from the principal planes (bottom).

Setting this equal to $M_{\text{thin lens}} = \begin{pmatrix} 1 & 0 \\ -1/f & 1 \end{pmatrix}$ we find a solution for

$$f = \frac{-1}{C},$$

$$s_1 = \frac{1-D}{C},$$

$$s_2 = \frac{1-A}{C}.$$

Let’s verify that we reproduce the principal plane locations found for the compound lens. The ABCD matrix for the two lenses is

$$\begin{pmatrix} 1 & d \\ -1/f_2 & 1 \end{pmatrix} \begin{pmatrix} 1/f_1 & 0 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} 1 - d/f_1 & d \\ d - f_1 - f_2 & 1 - d/f_2 \end{pmatrix}.$$

Thus the principal planes are at

$$s_1 = \frac{df_1}{d - f_1 - f_2},$$

$$s_2 = \frac{df_2}{d - f_1 - f_2},$$

which agrees with Eqs. (3.8).
Another way of thinking about the principal planes is to recognize that an arbitrary ABCD ray transformation acts as though all of the ray refraction takes place at the principal planes.

### 3.1.7 Lens Maker’s formula

As another example of the use of the ray matrices consider the thick lens shown in Fig. 3.9. The first surface has radius of curvature $R_1 > 0$, the lens material has index $n$, and the second surface has radius of curvature $R_2 < 0$. Using the appropriate matrices from Fig. 3.4 the matrix for the lens is

\[
M = \begin{pmatrix}
1 & 0 & 0 & 0 \\
(n-1)/R_2 & n & 0 & 0 \\
1 & d & 1 & 0 \\
((1-n)/nR_1) & 1 & 1/n & 0
\end{pmatrix}
\]

The focal length is given by the lens maker’s formula

\[
\frac{1}{f} = (n-1) \left[ \frac{1}{R_1} - \frac{1}{R_2} + \frac{(n-1)d}{nR_1 R_2} \right].
\]

The principal planes are located at

\[
s_1 = \frac{dR_1}{(n-1)d - n(R_1 - R_2)},
\]

\[
s_2 = \frac{dR_2}{d(1-n) + n(R_1 - R_2)}.
\]

Let’s consider some numerical examples. For a symmetric bi-convex lens take $R_1 = 10$ cm, $R_2 = -10$ cm, $d = 0.2$ cm, and $n = 1.5$. These values give

\[
f = 10.03$ cm, \quad s_1 = -0.067$ cm, \quad s_2 = -0.067$ cm.
\]
We see that the principal planes are at negative distances, i.e. they are inside the lens, approximately 1/3 of the thickness from each surface. For a negative bi-convex lens take the same values except \( R_1 = -10 \) cm, \( R_2 = 10 \) cm. We find
\[
f = -9.97 \text{ cm}, \quad s_1 = -0.066 \text{ cm}, \quad s_2 = -0.066 \text{ cm}.
\]
For a plano-convex lens we could use \( R_1 = \infty \) cm, \( R_2 = -10 \) cm, \( d = 0.2 \) cm, \( n = 1.5 \) giving
\[
f = 20. \text{ cm}, \quad s_1 = -0.13 \text{ cm}, \quad s_2 = 0. \text{ cm}.
\]

### 3.1.8 Optical plates

As an example of the use of the ray matrices to analyze other optical elements consider the slab of index \( n_2 \) shown in Fig. 3.10. The thickness of the slab is \( t \), and it is rotated so that the axis of the slab is at an angle \( \alpha \) with respect to the propagation axis. The internal ray length is then
\[
L = t/\sqrt{1 - n_2^2 \sin^2(\theta + \alpha)/n_1^2}.
\]
Using the matrices given in Fig. 3.4 we can find the matrix describing propagation through the slab. The general expression is not compact. Of most interest is the situation where the beam is incident at Brewster’s angle for which \( \tan(\theta + \alpha) = n_2/n_1 \) and
\[
L = t\sqrt{1 + n_2^2/n_1^2}/(n_2/n_1).
\]
The matrix of the slab for tangential rays is then
\[
M_t = \begin{pmatrix}
\frac{n_2}{n_1} & 0 & 0 \\
0 & \frac{n_2^2}{n_1^2} & 0 \\
0 & 0 & \frac{n_2}{n_1}
\end{pmatrix}
\begin{pmatrix}1 & L & 0 \\
0 & 0 & 1
\end{pmatrix}
\begin{pmatrix}
\frac{n_2}{n_1} & 0 & 0 \\
0 & \frac{n_2^2}{n_1^2} & 0 \\
0 & 0 & \frac{n_2}{n_1}
\end{pmatrix}
= \begin{pmatrix}1 & \frac{L}{(n_2/n_1)} & 0 \\
0 & 0 & 1
\end{pmatrix},
\]
and for sagittal rays\(^2\)
\[
M_s = \begin{pmatrix}1 & 0 & 0 \\
0 & \frac{n_2}{n_1} & 0 \\
0 & 0 & \frac{n_1}{n_2}
\end{pmatrix}
\begin{pmatrix}1 & L & 0 \\
0 & 0 & 1
\end{pmatrix}
\begin{pmatrix}1 & 0 & 0 \\
0 & \frac{n_2}{n_1} & 0 \\
0 & 0 & \frac{n_1}{n_2}
\end{pmatrix}
= \begin{pmatrix}1 & L \frac{n_2}{n_1} & 0 \\
0 & 0 & 1
\end{pmatrix}.
\]

An optical element that is often used for optical pumping of laser or nonlinear optical cavities is a tilted, partially transmitting mirror as shown in Fig. 3.11. Application of the

---


---

![Figure 3.10](image-url)
Figure 3.11: A tilted mirror with radius of curvature \( R \), thickness \( t \), and substrate index \( n_2 \).

Ray matrices from Fig. 3.4 leads to

\[
M_t = \left( \begin{array}{c}
\frac{e_2 n_r}{e_1 R \cos \alpha} \\
\frac{e_1 n_r}{e_2 n_r^2 \cos \alpha} - 1
\end{array} \right) 1 + \frac{t}{e_2 n_r^2 R} \left( \sec \alpha - \frac{n_r}{e_1} \right)
\]

(3.12a)

\[
M_s = \left( \begin{array}{c}
\frac{1}{R (e_1 - n_r \cos \alpha)} \\
1 + \frac{t}{e_2 n_r R (e_1 - n_r \cos \alpha)}
\end{array} \right)
\]

(3.12b)

where \( n_r = n_2 / n_1 \), \( e_1 = \sqrt{1 - n_r^2 \sin^2 \alpha} \), and \( e_2 = \sqrt{1 - \sin^2 \alpha / n_r^2} \). An initially round beam entering from the left is transformed into an elliptically shaped astigmatic beam. Optimization of a nonlinear cavity where the pump beam passes through this type of tilted curved mirror requires careful matching to the cavity mode or precompensation of the pump beam to balance the mirror induced astigmatism.

3.1.9 Sign conventions

The sign conventions we use for ABCD matrices are as follows.

1) Rays travel from left to right.

2) Positions above (below) the optical axis have \( x > 0 \) (\( x < 0 \)).

3) Rays pointing towards positive \( x \) (negative \( x \)) have \( \theta > 0 \) (\( \theta < 0 \)).

4) Curved surfaces with center of curvature to the right (left) of the surface have \( R > 0 \) (\( R < 0 \)).

5) Object distances are positive (negative) when the object is to the left (right) of the optical element.

6) Image distances are positive (negative) when the image is to the right (left) of the optical element.

Note that these definitions are not universal and different sign conventions can be found in different texts.
3.1.10  Brightness

An important question is the possibility of increasing the brightness of an optical beam using a system of optical elements described by $ABCD$ matrices. We can think of an optical beam as being a bundle of rays with a characteristic width $\Delta x$ and angular divergence $\Delta \theta$. In one transverse dimension the brightness of a beam is defined as $B = P/(\Delta x)(\Delta \theta)$, where $P$ is the power.

Transformation by an optical system described by matrix $M$ results in $x' = Ax + B\theta$ and $\theta' = Cx + D\theta$. Therefore

\[
(\Delta x')^2 \equiv \langle (x' - \langle x' \rangle)^2 \rangle = \langle x'^2 \rangle - \langle x' \rangle^2 = A^2(\Delta x)^2 + B^2(\Delta \theta)^2 + 2AB(\langle x\theta \rangle - \langle x \rangle \langle \theta \rangle) \tag{3.13}
\]

and

\[
(\Delta \theta')^2 = C^2(\Delta x)^2 + D^2(\Delta \theta)^2 + 2CD(\langle x\theta \rangle - \langle x \rangle \langle \theta \rangle). \tag{3.14}
\]

We assume that the input beam is centered on the $x$ axis so $\langle x \rangle = 0$ and that the $x$ and $\theta$ distributions are uncorrelated so that $\langle x\theta \rangle = 0$. We then have, using $AD - BC = 1$,

\[
(\Delta x')^2(\Delta \theta')^2 = (A^2D^2 + B^2C^2)(\Delta x)^2(\Delta \theta)^2 + A^2C^2(\Delta x)^2 + B^2D^2(\Delta \theta)^2 = (1 + 2ABCD)(\Delta x)^2(\Delta \theta)^2 + A^2C^2(\Delta x)^2 + B^2D^2(\Delta \theta)^2 = (\Delta x)^2(\Delta \theta)^2 + (AC\Delta x + BD\Delta \theta)^2 \geq (\Delta x)^2(\Delta \theta)^2. \tag{3.15}
\]

Thus $\Delta x'\Delta \theta' \geq \Delta x\Delta \theta$ so $B' \leq B$. The brightness cannot be increased by a linear optical system. The brightness is maximized when $AC = 0$ and $BD = 0$. Since $AD - BC = 1$ it is impossible for $A&B$ or $C&D$ to simultaneously vanish so the condition is $A = D = 0$ or $B = C = 0$. We can identify the product $\Delta x\Delta \theta$ as the phase space density of the beam. The impossibility of reducing the phase space density is an example of a general result from classical mechanics known as Liouville’s Theorem (see for example Landau & Lifshitz, Mechanics).

3.2  Aberrations

The paraxial approximation is just that, an approximation. A more careful treatment that accurately follows the refraction at each interface without any approximation to $\sin(\theta)$ shows that even when the imaging condition is fulfilled rays originating from a single object point may not converge to a single point in the image plane. In addition the separation of object points may be distorted in the image plane. These aberrations become more pronounced as the ray angles with respect to the optical axis increase. The variation of refractive index with wavelength leads to chromatic aberrations, even for small ray angles.

Aberrations can be corrected for using multiple lenses with different types of glass designed to compensate both chromatic and other errors. The study of aberrations was originally based on analytical methods and there exist extensive classical results. With the advent of optical design software for ray tracing the study of aberrations for even very complex optical systems can now be quickly performed on computers.

March 8, 2016 M. Saffman
3.2.1 Primary monochromatic aberrations

Despite the availability of computer programs for ray tracing it is still useful to understand the basic types of aberration that occur. Aberrations that depend on the wavelength of the light due to dispersion of the optical refractive index, $n = n(\lambda)$ are called chromatic aberrations. Aberrations that occur for a single wavelength are called monochromatic aberrations. The most important monochromatic aberrations can be categorized analytically by approximating

\[
\sin(\theta) \simeq \theta - \theta^3/6, \quad \cos(\theta) \simeq 1 - \theta^2/2, \quad \tan(\theta) \simeq 1 + \theta^3/3.
\]

This approximation leads to five types of third order aberration. These were categorized in a series of 1856 papers by Seidel who considered a centered system of spherical surfaces\(^3\) and

\(^3\)L. Seidel, Astr. Nachr. 43, No. 1027, 289 (1856), No. 1028, 305 (1856), No. 1029, 321 (1856).
are known as the Seidel aberrations shown in Fig. 3.12. They are

- Spherical aberration. Rays from a point on the optical axis do not converge to a single point on axis.
- Coma. Rays from an off-axis image point do not converge to a single point. The distribution of the rays looks like a comet, hence the name coma.
- Astigmatism. Rays in the tangential and sagittal planes are focused at different axial locations.
- Field curvature. The surface on which rays are imaged is not planar but lies on a curved surface.
- Distortion. Rays from an object point converge to a single image point but the distance of the image point from the optical axis is not proportional to the distance of the object point.

**Spherical aberration calculation**

Consider the geometry of Fig. 3.13. A ray with angle $\theta_1$ refracts at a surface of curvature $R$ at height $x_1 = h$. The normal to the surface makes an angle $\alpha$ with the optical axis. The local normal to the surface can be written as $\mathbf{n} = -\cos(\alpha)\hat{z} + \sin(\alpha)\hat{x}$ with $\sin(\alpha) = h/R$. The ray starts in a medium with index $n_1$ and refracts into a medium with index $n_2$. After refraction the ray is at height $x_2 = h$ and has angle with respect to the surface normal of

$$\theta' = \sin^{-1}\left[\frac{n_1}{n_2} \sin(\theta_1 + \alpha)\right]. \quad (3.16)$$

The refracted ray meets the planar back surface of the lens at height $h'$ with angle from the normal of $\theta'' = \alpha - \theta'$. The height $h'$ is given by

$$h' = h - t' \tan(\theta'') \quad (3.17)$$

with $t' = t - R(1 - \cos(\alpha))$. The ray then refracts giving a ray heading towards the optical axis with angle

$$\theta_2 = \sin^{-1}\left[\frac{n_2}{n_1} \sin(\theta'')\right]$$
$$\quad = \sin^{-1}\left[\frac{n_2}{n_1} \sin\left(\alpha - \sin^{-1}\left[\frac{n_1}{n_2} \sin(\theta_1 + \alpha)\right]\right)\right]. \quad (3.18)$$

The ray crosses the optical axis at a distance after the front surface of the lens given by

$$d = t + \frac{h'}{\tan(\theta_2)} \quad (3.19)$$

where $t$ is the lens thickness.

March 8, 2016 M. Saffman
Combining Eqs. (3.16 - 3.19) we obtain an expression for the distance \(d_0\) as a function of the ray height \(h\). The paraxial approximation implies that the ratio \(\eta = h/R\) is a small parameter. After some algebra we find for the simplest case of \(\theta_1 = 0\) and \(t = 0\)

\[
d = \frac{R}{n_r - 1} + \frac{n_r^2(2 - n_r) - 2}{2n_r(n_r - 1)} R\eta^2 + \mathcal{O}(\eta^4)
\]

where the relative refractive index is \(n_r = n_2/n_1\). From the lens maker’s formula Eq. (3.9) the focal length is \(f = R/(n_r - 1)\) so

\[
\frac{d}{f} = 1 + \frac{n_r^2(2 - n_r) - 2}{2n_r} \eta^2 + \mathcal{O}(\eta^4).
\]  

(3.20)

The coefficient in front of \(\eta^2\) determines the amount of spherical aberration of the lens. The fractional shift of the focal position as a function of \(\eta\) is shown in Fig. 3.13. It is a happy coincidence that for the plano-convex lens the spherical aberration is minimized for \(n_r = 1.466\) which is close to the index of common glass.

The other Seidel aberrations can be analyzed by ray tracing in the same manner as we have done for the spherical aberration. The geometrical details are complicated and the results can be found in many optics books.

### 3.2.2 Chromatic aberration

In most glasses the refractive index increases towards shorter wavelengths. Thus blue light experiences greater bending than red light. This leads to the image location varying with the wavelength which results in colored fringes. The chromatic aberration can lead to axial variations in image position (referred to as axial color) or transverse variations (referred to as lateral color). To correct for chromatic aberration the lens designer combines glasses with different dispersion characteristics to compensate for this effect. A lens that is designed to have no chromatic aberration at two distinct wavelengths is called an achromat. Such lenses will have relatively good performance over a broad range of wavelengths. Lenses that are corrected at three distinct wavelengths require more complex designs and are called apochromats.
3.3 Image forming instruments

3.3.1 The eye

The eye is a precision optical instrument that focuses light onto the retina. Focusing is done by the combination of the cornea and the eye lens. The cornea has a fixed focal length while the eye lens responds to muscle tension by changing shape and focusing power in order to bring objects at different distance into focus on the retina. The image on the retina is inverted and the wiring from the retina to the brain corrects this.

The image distance from the eye lens to the retina is fixed and is typically $d_i \simeq 1.7$ cm.

The power of the lens is measured in diopters $D = 1/f$ which are defined as the inverse of the focal length in m. A typical person has an eye with a relaxed focusing power of $D_0 \simeq 58.8$ diopters of which 43 come from the cornea and the rest come from the eye lens which is variable. Muscle tension on the eye lens is used to increase the focusing power, an ability which decreases with age. A young person can typically achieve an additional 20 diopters, by age 25 about 10 diopters, and by age 50 about 1 diopter. This limits the ability to form an image of objects at short distance, hence the need for reading glasses.

In terms of the focusing power the object distance is

$$d_o = \frac{1}{D - 1/d_i}.$$  

Using $d_i = 0.017$ m the relaxed eye will focus a distant object onto the retina. Assuming an accomodation of $D_a = 6$ in an adult we get a total power of of $D = D_0 + D_a = 64.8$ giving $d_0 = 16.7$ cm. A nominal value of 15 cm is typically assumed as the shortest distance an eye can focus to.

The size of an object viewed by the eye depends on the magnification $M = |d_i/d_o|$. At the nominal closest viewing distance of $d_0 = 15$ cm the eye has a magnification of

---

To be precise the imaging equation should be modified to account for the index of the fluid inside the eye. We will sidestep this complication.

March 8, 2016 M. Saffman
$M = 1.7/15 = 0.11$. This can be improved on with a simple magnifying glass which is a lens with a power $D_{mg}$. If the magnifying glass is brought close to the eye the focusing power becomes $D = D_0 + D_{mg}$. Suppose the magnifying glass has power $D_{mg} = 40$ (this corresponds to $f = 2.5$ cm). Then a relaxed eye can image an object at distance $d = 1/(D-1/d_i) = 2.5$ cm giving a magnification of $d_i/d = 1.7/2.5 = 0.68$ which is an increase of $\times 6$ relative to the unaided eye. Note that the magnifying glass forms a virtual image in front of the lens at a distance which can be comfortably viewed by the eye.

Another way to express the improvement is that we get a magnification by a factor of $d_0/f$ where $d_0$ is the shortest object distance of the unaided eye. Since shorter $f$ requires thicker and more strongly curved lenses there is a practical limit of about $\times 25$ for a magnifying glass. This can be improved on using a microscope.

### 3.3.2 Microscope

The microscope as shown in Fig. 3.15 has two lenses, the objective lens which forms a magnified real image, followed by an eyepiece which gives an additional magnifying factor in the same way as the magnifying glass. Most modern microscopes are standardized to have a distance $L = 16$ cm between the objective and the intermediate real image. The objective is often a multi-element lens that has been designed to minimize aberrations. Focal lengths as short as $f_1 \approx 1$ mm are common. This give a magnification of 160 at the intermediate focus. With an eyepiece that gives an additional magnification of $\times 20$ total magnification factors of several thousand are possible.

The ultimate limit for the smallest object that can be viewed is not set by the magnification but by the wave nature of light and diffraction. As we will see later on the resolution limit of the objective lens is approximately $\lambda/(2NA)$ where $\lambda$ is the wavelength and $NA$ is the numerical aperture of the lens. The numerical aperture is $NA = n_o \sin(\theta)$ where $n_o$ is the refractive index surrounding the object and $\theta$ is the half opening angle of the lens. High resolution microscope objectives can have a numerical aperture $NA = 0.9$, or even greater than 1 for oil immersion devices. Thus objects smaller than the wavelength can be viewed.
3.3.3 Telescope

Telescopes are similar to microscopes, but designed to view very distant objects. In typical usage the object distance $d_o$ is very large compared to the focal length of the objective lens $f_o$. Thus the intermediate image is formed close to the back focal plane of the lens. The eyepiece and the objective are separated by the sum of their focal lengths $f_o + f_e$ and in this configuration the ray matrix from the front of the objective lens to the back of the eyepiece is

$$M_{tel} = \begin{pmatrix} 1 & 0 \\ -1/f_e & 0 \end{pmatrix} \begin{pmatrix} 1 & f_o + f_e \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ -1/f_o & 0 \end{pmatrix} = \begin{pmatrix} -f_e/f_o & f_e + f_o \\ 0 & -f_o/f_e \end{pmatrix}.$$ 

We see that the effective focal length is infinite so the focusing power is zero. Such an instrument is called afocal. Parallel rays are transformed into parallel rays, but at a larger angle to the optical axis. In other words the telescope provides angular magnification by an amount

$$M_\theta = \frac{\theta_{out}}{\theta_{in}} = -\frac{f_o}{f_e}.$$ 

The image is inverted. This can be corrected using the Galilean version of the telescope which uses a negative lens for the eyepiece and forms an upright image.

3.3.4 Binoculars

Binoculars are essentially two telescopes placed next to each other, one for each eye. The image inversion is corrected using a pair of prisms as in Fig. 3.16.

March 8, 2016 M. Saffman
3.4 Solid Angle

The amount of light collected by an imaging instrument is important for obtaining detailed images in as short a time as possible. The collection efficiency is determined by the solid angle of the first lens of the instrument. Consider the geometry of Fig. 3.17 where a lens or window of radius $a$ has its center a distance $z$ from the origin. The solid angle subtended by the lens at the origin is given by

$$\Omega = \int_A \frac{\hat{R} \cdot \hat{n}}{R^2} da = \int_0^{2\pi} d\phi \int_0^a d\rho \frac{\cos(\tan^{-1} \frac{\rho}{z^2 + \rho^2})}{\rho}$$

$$= \frac{2\pi}{z^2} \int_0^a d\rho \frac{\rho}{\left(1 + \frac{\rho^2}{z^2}\right)^{3/2}}$$

$$= 2\pi \left(1 - \frac{1}{\sqrt{1 + a^2/z^2}}\right).$$

(3.21)

The f-number of the lens is $f_\# = z/(2a)$ so the solid angle can be written as

$$\Omega = 2\pi \left(1 - \frac{1}{\sqrt{1 + 1/(2f_\#)^2}}\right) \approx \frac{\pi}{4f_\#^2}$$

where the last equality holds for large $f_\#$. Camera lens often specify the $f_\#$. The speed of a lens is proportional to the light gathering power which is proportional to the inverse of the $f_\#$.

It is also customary to specify optical lenses in terms of the numerical aperture defined as $NA = n \sin(\theta)$ where $n$ is the index of refraction of the medium the lens is embedded in, and $\theta = \tan^{-1}(a/z)$ is the cone angle of the lens. When $n = 1$ we have

$$\Omega = 2\pi \left(1 - \sqrt{1 - NA^2}\right) \approx \pi NA^2,$$

---

It is common notation to write the f-number as $f/\#$. 

---

$^5$It is common notation to write the f-number as $f/\#$. 

---
where the last equality holds for small $NA$.

It is convenient to compare the fractional solid angle as a function of the cone angle $\theta$, the $f\#$, and the NA. We have for $\tilde{\Omega} = \Omega/4\pi$,

$$
\tilde{\Omega}_\theta = \frac{1}{2} \left[ 1 - \frac{1}{\sqrt{1 + \tan^2(\theta)}} \right],
\tilde{\Omega}_{f\#} = \frac{1}{2} \left[ 1 - \frac{1}{\sqrt{1 + (2f\#)^2}} \right],
\tilde{\Omega}_{NA} = \frac{1}{2} \left( 1 - \sqrt{1 - NA^2} \right).
$$

(3.22)

Note that for a cone angle of 45 deg. $\theta = \pi/4$ and $f\# = 0.5$, and $NA = .707$. Finally, for convenience in converting from $f\#$ to $NA$ we note that

$$
NA = \frac{1}{\sqrt{1 + 4f\#^2}}, \quad \text{and} \quad f\# = \frac{\sqrt{1 - NA^2}}{2NA}.
$$

### 3.5 Eikonal equation

So far we have considered optical rays that propagate in straight lines but are reflected or refracted at surfaces separating media with different indices of refraction. In many cases there can be a continuous variation of the optical properties of a medium. For example in gases the refractive index is proportional to the density of atoms or molecules. In an ideal gas the density is in turn inversely proportional to the temperature. Temperature gradients therefore lead to index gradients which change the ray path in a continuous fashion. Another example is an optical fiber, or a lens, with a gradient index profile. The index profile is imprinted into the glass in the manufacturing process and is frozen in place.

We can describe the ray path in a continuously varying medium with the geometry of Fig. 3.19. Let the ray path be given by the vector $\mathbf{r}(s)$ with $s$ a scalar quantity which parameterizes distance along the ray. The local tangent vector to the ray is $\mathbf{\hat{t}} = \frac{d\mathbf{r}}{ds}$. The time it takes for light to follow a ray path $\mathcal{C}$ is

$$
t = \int_{\mathcal{C}} \frac{ds}{v(\mathbf{r})} = \frac{1}{c} \int_{\mathcal{C}} ds n(\mathbf{r}).
$$
The optical path length $\mathcal{L}$ is defined as

$$\mathcal{L} = ct = \int_{c} ds n(r) .$$ \hspace{1cm} (3.23)

The differential optical path can be expressed in two different ways. From (3.23) we have

$$d\mathcal{L} = n(r) ds = n(r) \hat{e}_t \cdot dr$$

and formally considering $\mathcal{L} = \mathcal{L}(r)$ we can write

$$d\mathcal{L} = \frac{\partial \mathcal{L}}{\partial r} \cdot dr = \nabla \mathcal{L} \cdot dr .$$

Equating the two expressions gives

$$\nabla \mathcal{L} = n(r) \hat{e}_t .$$

Squaring this relation we arrive at

$$\left( \nabla \mathcal{L} \right)^2 = n^2(r) ,$$

which is known as the eikonal equation.

It is instructive to derive this equation in a different way. Assume time harmonic fields at frequency $\omega$, $\vec{E}(r, t) = \vec{E}(r)e^{-i\omega t}$, $\vec{H}(r, t) = \vec{H}(r)e^{-i\omega t}$. Without loss of generality we factor the spatial dependence of the fields into a slowly varying part and a phase depending on the optical path length as

$$\vec{E}(r) = \vec{E}_0(r)e^{ik_0 \mathcal{L}(r)} ,$$

$$\vec{H}(r) = \vec{H}_0(r)e^{ik_0 \mathcal{L}(r)} .$$

Here $k_0 = \omega/c$ is the wavenumber of the light in vacuum. Taking the curl we get

$$\nabla \times \vec{E} = [\nabla \times \vec{E}_0 - ik_0(\nabla \mathcal{L} \times \vec{E}_0)]e^{ik_0 \mathcal{L}(r)} = i\omega \mu \vec{H} ,$$

$$\nabla \times \vec{H} = [\nabla \times \vec{H}_0 - ik_0(\nabla \mathcal{L} \times \vec{H}_0)]e^{ik_0 \mathcal{L}(r)} = -i\omega \varepsilon \vec{E} .$$

The last equality on each line follows from the Maxwell equations. Thus

$$\nabla \mathcal{L} \times \vec{H}_0 = -\frac{i}{k_0} \nabla \times \vec{H}_0 + \frac{\omega}{k_0} \varepsilon \vec{E}_0$$

Figure 3.19: Ray propagation in a medium with continuously varying index $n(r)$.
or
\[ \nabla \mathcal{L} \times \vec{H}_0 - c \varepsilon \vec{E}_0 = -\frac{i}{k_0} \nabla \times \vec{H}_0. \]

Similarly we find
\[ \nabla \mathcal{L} \times \vec{E}_0 + c \mu \vec{H}_0 = -\frac{i}{k_0} \nabla \times \vec{E}_0. \]

The ray approximation is that \( \lambda_0 \to 0 \) or \( k_0 \to \infty \) so the right hand sides can be neglected leaving
\[ \nabla \mathcal{L} \times \vec{H}_0 - c \varepsilon \vec{E}_0 = 0, \]
\[ \nabla \mathcal{L} \times \vec{E}_0 + c \mu \vec{H}_0 = 0. \]

Eliminating \( \mathcal{H}_0 \) this becomes
\[ \nabla \mathcal{L} \times (\nabla \mathcal{L} \times \vec{E}_0) + n^2 \vec{E}_0 = 0, \]
where we have used \( c^2 \varepsilon \mu = \frac{\varepsilon \mu}{\varepsilon_0 \mu_0} = n^2 \). Expanding the triple product gives
\[ (\nabla \mathcal{L} \cdot \vec{E}_0) \nabla \mathcal{L} - (\nabla \mathcal{L})^2 \vec{E}_0 + n^2 \vec{E}_0 = 0. \]

The first factor can be shown to vanish in the ray limit\(^6\) which leaves \(-(\nabla \mathcal{L})^2 \vec{E}_0 + n^2 \vec{E}_0 = 0\) or
\[ (\nabla \mathcal{L})^2 = n^2 \]
which is the eikonal equation.

### 3.5.1 Ray equation

Using the eikonal equation we can derive an equation for the ray path \( \mathbf{r}(s) \). We have
\[
\frac{d}{ds} \nabla \mathcal{L} = \nabla(\nabla \mathcal{L}) \cdot \frac{d\mathbf{r}}{ds} = \frac{d\mathbf{r}}{ds} \cdot \nabla(\nabla \mathcal{L}) = \hat{\mathbf{e}}_t \cdot \nabla(\nabla \mathcal{L}) = \frac{\nabla \mathcal{L}}{n(\mathbf{r})} \cdot \nabla(\nabla \mathcal{L}) = \frac{\nabla [(\nabla \mathcal{L})^2]}{2n(\mathbf{r})} = \nabla n(\mathbf{r}).
\]

We can also write
\[
\frac{d}{ds} \nabla \mathcal{L} = \frac{d}{ds} \left[ n(\mathbf{r}) \hat{\mathbf{e}}_t \right] = \frac{d}{ds} \left[ n(\mathbf{r}) \frac{d\mathbf{r}}{ds} \right].
\]

Equating the expressions for \( \frac{d}{ds} \nabla \mathcal{L} \) gives
\[
\frac{d}{ds} \left[ n(\mathbf{r}) \frac{d\mathbf{r}}{ds} \right] = \nabla n(\mathbf{r}). \quad (3.24)
\]

This is the equation for the optical ray path in an inhomogeneous medium which can be thought of as a generalization of Snell’s law. When \( n(\mathbf{r}) = n_0 \) is a constant we get immediately
\[
\frac{d}{ds} \left( n_0 \frac{d\mathbf{r}}{ds} \right) = 0 \quad \rightarrow n_0 \frac{d\mathbf{r}}{ds} = \text{constant}.
\]
The solution is \( \mathbf{r} = c_1 + sc_2 \), with \( c_1, c_2 \) constant vectors. We see that rays in a homogeneous optical medium indeed travel in straight lines.

Let’s use the ray equation to calculate the path of rays in a medium with a continuously varying index. A common and important example of this is a graded index optical fiber as shown in Fig. 3.20. The index is a function of the radial coordinate, \( n(x) \) and is a maximum on the axis of the fiber. This geometry supports bound rays that propagate along the fiber with a maximum value of \( x = x_{\text{max}} \) for all \( z \). There are two types of rays: meridional rays that pass through the fiber axis, and skew rays that start off-axis and follow a spiral path.

Let’s consider a meridional ray propagating mainly along \( \hat{z} \) for which the motion can be described in terms of a 2D geometry, \( \mathbf{r} = x(z)\hat{x} + z\hat{z} \). We parameterize the transverse coordinate \( x = x(z) \) by the axial coordinate \( z \). At a distance \( s \) along the ray the unit tangent vector makes an angle \( \theta(x) \) with the \( \hat{z} \) axis, i.e \( \hat{\epsilon}_t = \sin \theta \hat{x} + \cos \theta \hat{z} \). Using \( \hat{\epsilon}_t = \frac{d\mathbf{r}}{ds} \) we see that \( \frac{dx}{ds} = \sin \theta \) and \( \frac{dz}{ds} = \cos \theta \). The two components of (3.24) are then

\[
\frac{d}{ds} [n(x) \sin \theta(x)] = \frac{dn(x)}{dx}, \quad \frac{d}{ds} [n(x) \cos \theta(x)] = 0.
\]

The second equation can be integrated to give

\[
n(x) \cos \theta(x) = \bar{n} = n(0) \cos \theta(0),
\]

with \( \bar{n} \) a constant. If \( n \) has a maximum at \( x = 0 \) and decreases monotonically to both sides away from the axis then the ray is confined and has a maximum value \( x_{\text{max}} \) which occurs when \( \cos \theta(x_{\text{max}}) = 1 \). The index of refraction at \( x_{\text{max}} \) is

\[
\bar{n} = n(x_{\text{max}}) \cos \theta(x_{\text{max}}) = n(0) \cos \theta(0).
\]

The actual value of \( x_{\text{max}} \) depends on the initial angle and the shape of the index profile and is found from solving \( n(x_{\text{max}}) = n(0) \cos \theta(0) \).

To find the ray path rewrite we can rewrite (3.25) as \( n \frac{dx}{ds} = \bar{n} \) so \( \frac{d}{ds} = \frac{\bar{n}}{n} \frac{d}{dz} \). The equation for the \( x \) component of the ray path can then be written as

\[
\frac{d}{ds} [n(x) \sin \theta(x)] = \frac{\bar{n}}{n} \frac{d}{dz} [n(x) \sin \theta(x)] = \frac{\bar{n}}{n} \frac{dx}{dz} \left[ n(x) \frac{dx}{ds} \right] = \frac{\bar{n}}{n} \frac{d}{dz} \left[ \frac{dx}{dz} \right] = \frac{dn(x)}{dx}.
\]

---

Figure 3.20: Periodic ray trajectories in a gradient index fiber. The numerical solutions are for \( n(x) = 1.2 \text{sech}(x) \), and \( \theta(0) = 0.1, 0.79 \).
or
\[ \bar{n}^2 \frac{d^2 x}{dz^2} = \frac{1}{2} \frac{d}{dx} n^2. \]

This has the same form as for the motion of a particle in a potential. As in mechanics problems we can integrate by multiplying with \( \frac{dx}{dz} \) to get
\[ \bar{n}^2 \frac{d^2 x}{dz^2} dx = \frac{1}{2} \frac{d}{dz} dx n^2 = \frac{1}{2} \frac{d}{dz} \bar{n}^2, \]

which integrates to
\[ \bar{n}^2 \left( \frac{dx}{dz} \right)^2 = n^2 + c_1. \]

At \( x = 0 \) the slope is \( \frac{dx}{dz}|_{x=0} = \frac{n(0)}{\bar{n}} \frac{dx}{ds}|_{x=0} = \frac{n(0)}{\bar{n}} \sin \theta(0) \) and
\[ c_1 = \bar{n}^2 \frac{n^2(0)}{\bar{n}^2} \sin^2 \theta(0) - n^2(0) = -n^2(0) \cos^2 \theta(0) = -\bar{n}^2. \]

Thus the ray path is given by
\[ \bar{n}^2 \left( \frac{dx}{dz} \right)^2 = n^2 - \bar{n}^2. \]

To solve for the trajectory \( z(x) \) we write
\[ \bar{n} \frac{dx}{dz} = \sqrt{n^2 - \bar{n}^2} \]

and invert to get
\[ z(x) = \bar{n} \int_0^x \frac{dx'}{\sqrt{n^2(x') - \bar{n}^2}} \]
assuming the ray begins at \( x = 0 \) with angle \( \theta(0) \).

For some choices of the profile \( n(x) \) this can be solved analytically. For example using \( n(x) = n(0) \text{sech}(\alpha x) \) with \( \alpha \) a constant the solution is
\[ x = \frac{1}{\alpha} \sinh^{-1} [\sinh(\alpha x_{\max}) \sin(\alpha z)] \]
where \( \bar{n} = n(x_{\max}) = n(0) \cos \theta(0) \). A few trajectories for different values of \( \theta(0) \) are shown in Fig. 3.20. As the initial angle is increased the ray reaches larger values of \( x_{\max} \) while the axial period of the trajectory is unchanged.
Chapter 4

Fourier optics

Fourier optics relies on spectral decomposition of optical fields as a route to calculating field changes under propagation. The spatial amplitude $A(x, y)$ can be represented by a Fourier transform as $\tilde{A}(k_x, k_y)$ where $k_x, k_y$ are the transverse components of the wavevector. Propagation of $\tilde{A}$ is efficiently calculated by multiplying with a quadratic phase factor. Since different transverse wavevectors pick up different propagation phases the spatial distribution found from transforming back to $A(x, y)$ changes. In this way, even though rays propagate in a straight line in a homogeneous medium, optical fields diffract. Diffractive spreading of localized fields is an important concept in optics and leads to resolution limits in imaging and spectroscopy.

A physical picture of diffraction can be based on the Huygens’ construction$^1$ as depicted in Fig. 4.1. Each point on the surface of a wavefront acts as a source for expanding spherical waves that combine to give a new downstream wavefront. A mathematical formulation of Huygens’ construction suitable for quantitative calculations was provided by Fresnel in 1816$^2$. Subsequent work showed that some modification to the Huygens-Fresnel theory was necessary to make it fully mathematically self-consistent. In this chapter we will not follow the historical development but will start with a paraxial theory of diffraction which follows from the Maxwell equations in Sec. 4.1. After developing the Fresnel and Fraunhofer approximations we will show the connection to the more accurate Huygens-Fresnel theory,

---

$^1$C. Huygens, Traite de la Lumiere, 1690. An English translation is available at www.gutenberg.org.


Figure 4.1: Huygens’ construction for describing propagation of wavefronts.
4.1 Paraxial propagation and diffraction

In a uniform medium with linear electric and magnetic properties and no free charges the Maxwell equations are

\[ \nabla \times \mathbf{E} = -\frac{\partial \mathbf{B}}{\partial t}, \quad \nabla \times \mathbf{B} = \epsilon \mu \frac{\partial \mathbf{E}}{\partial t}, \]
\[ \nabla \cdot \mathbf{E} = 0, \quad \nabla \cdot \mathbf{B} = 0. \]

Assuming a time dependence \( e^{-i\omega t} \) we find

\[ \nabla \times \nabla \times \mathbf{E} = \epsilon \mu \omega^2 \mathbf{E}. \]

Using \( \epsilon \mu \omega^2 = \frac{\epsilon \mu \omega^2}{\epsilon_0 \mu_0 c^2} = k^2 n^2 \) with \( k = 2\pi/\lambda_{\text{vac}} \) and \( \nabla \times \nabla \times \mathbf{E} = \nabla (\nabla \cdot \mathbf{E}) - \nabla^2 \mathbf{E} = -\nabla^2 \mathbf{E} \) we get

\[ \nabla^2 \mathbf{E} + k^2 n^2 \mathbf{E} = 0. \]

The same equation holds for \( \mathbf{B} \). Each component of the electric or magnetic field thus satisfies the scalar Helmholtz equation

\[ \nabla^2 U + k^2 n^2 U = 0. \]

The effective wavenumber inside the medium is given by \( kn = \omega n/c \) where \( \omega \) is the angular frequency of the field, \( c \) is the speed of light, and \( n \) is the index of refraction. In the following we will mostly assume \( n = 1 \). Although the Helmholtz equation is an exact description of the field evolution in a homogeneous medium it is only an approximation when the refractive index is spatially varying. Provided the length scales of the variation are large compared to the wavelength of light the Helmholtz equation provides a good description. In the opposite regime of nano-optics other methods such as finite difference time domain (FDTD) numerical solutions of the Maxwell equations are required. With the availability of fast computers FDTD simulations have become widely used\(^3\).

The scalar Helmholtz equation admits exact solutions including plane waves \( U = A e^{ikr} \) and spherical waves \( U = (A/r)e^{ikr} \) with \( A \) constant. In propagation problems we are interested in describing the transverse distribution of a light beam that propagates in a definite direction as shown in Fig. 4.2. Choosing \( \hat{z} \) as the propagation direction, and seeking a solution of the form \( U = A(x, y, z)e^{ikz} \) gives

\[ \left( \nabla^2 A + i2k \frac{\partial A}{\partial z} - k^2 A + k^2 A \right) e^{ikz} = 0 \]

which simplifies to

\[ \nabla_\perp^2 A + \frac{\partial}{\partial z} \left( \frac{\partial A}{\partial z} + i2kA \right) = 0 \tag{4.1} \]

where $\nabla^2 A = \partial^2 A/\partial x^2 + \partial^2 A/\partial y^2$. Solutions of Eq. (4.1) are exact solutions of the Helmholtz equation. We now introduce a slowly varying envelope approximation by assuming $(\partial A/\partial z) \ll 2kA$ which results in

$$\frac{\partial A}{\partial z} - \frac{i}{2k} \nabla^2_\perp A = 0. \quad (4.2)$$

All of Fresnel and Fraunhofer diffraction theory for scalar fields are described by the solutions of Eq. (4.2) which is called the paraxial wave equation. It is a partial differential equation of parabolic type. Given an appropriate set of initial conditions such as $A(x, y, 0)$ the solution is uniquely determined for all $A(x, y, z > 0)$.

We can develop a physical picture of diffractive propagation by decomposing the field $A(x, y)$ into a superposition of plane waves with transverse components $\mathbf{k}_\perp = k_x \hat{x} + k_y \hat{y}$ and amplitudes $A(k_x, k_y)$. A plane wave $\mathbf{k}_\perp = (k_x, k_y)$ picks up a propagation phase from $r_1 = (0, 0, z_1)$ to $r_2 = (0, 0, z_2)$ of

$$e^{i\phi(\mathbf{k}_\perp)} = e^{i\mathbf{k}(r_2-r_1)} = e^{ik_z (z_2-z_1)}.$$

The phase difference compared to a plane wave propagating along $\hat{z}$ is

$$\delta\phi(\mathbf{k}_\perp) = \phi(\mathbf{k}_\perp) - \phi(0) = (k_z - k)(z_2 - z_1).$$

Making a paraxial approximation we write

$$k_z = \sqrt{k^2 - k^2_\perp} \simeq k - \frac{k^2_\perp}{2k}.$$

Therefore off-axis waves pick up a differential phase shift compared to an axial wave given by

$$\tilde{A}(\mathbf{k}_\perp, z_1) \rightarrow \tilde{A}(\mathbf{k}_\perp, z_2) = \tilde{A}(\mathbf{k}_\perp, z_1)e^{-i\frac{k^2_\perp}{2k}z}.$$

The relative phase of the spectral components proportional to $k^2_\perp$ results in a new field distribution $A(x, y, z_2)$ which is found from combining the plane wave components with the new phases. This results in the physical phenomena of diffraction and spreading of localized wave packets, even though each individual plane wave propagates along a straight line.
4.1 Paraxial propagation and diffraction

4.1.1 Fresnel Diffraction

The above discussion can be put in a formal framework using the mathematics of Fourier transforms. The spatial spectrum of \( A(x, y) = A(\rho) \) is given by

\[
\tilde{A}(k_{\perp}) = \frac{1}{2\pi} \int_{-\infty}^{\infty} d\rho \ A(\rho) e^{-i\rho \cdot k_{\perp}},
\]

and the field is found from the inverse transform

\[
A(\rho) = \frac{1}{2\pi} \int_{-\infty}^{\infty} dk_{\perp} \tilde{A}(k_{\perp}) e^{i\rho \cdot k_{\perp}}.
\]

A useful representation of the Dirac delta function is given by the transform of unity:

\[
\delta(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} dk \ x e^{ikx}.
\]

Using the shorthand notation \( F[\ldots] = \frac{1}{2\pi} \int_{-\infty}^{\infty} d\rho \ \ldots e^{-ik_{\perp} \cdot \rho} \) we have \( F[\partial A/\partial x] = -ik_{x} F[A] \) and we get

\[
\frac{\partial \tilde{A}}{\partial z} = -(i/2k)k_{\perp}^{2} \tilde{A}.
\]

This can be solved as \( \tilde{A}(k_{\perp}, z_{2}) = \tilde{A}(k_{\perp}, z_{1}) e^{-i(1/2k)k_{\perp}^{2}z} \) where \( k_{\perp}^{2} = k_{x}^{2} + k_{y}^{2} \) and \( z = z_{2} - z_{1} \). Transforming back to the spatial domain we find

\[
A_{2}(\rho_{2}) = F^{-1} [F[A_{1}(\rho_{1})] H(k_{\perp}, z)]
\]

\[
= \frac{1}{4\pi^{2}} \int_{-\infty}^{\infty} dk_{\perp} e^{ik_{\perp} \cdot \rho_{2}} e^{-i\frac{k_{\perp}^{2}}{2k}(z_{2} - z_{1})} \int_{-\infty}^{\infty} d\rho_{1} A_{1}(\rho_{1}) e^{-i\rho_{1} \cdot \rho_{2}}
\]

(4.5)

where the transfer function of free space is

\[
H(k_{\perp}, z) = e^{-i(k_{\perp}^{2}/2k)z}.
\]

(4.6)

As discussed above the transfer function \( H \) is just the phase shift picked up by a wave propagating with a transverse wavenumber \( k_{\perp} \). While the integral in (4.5) extends to infinite values of \( k_{x}, k_{y} \) it is important to recognize that contributions from wavenumbers with \( k_{\perp}^{2} > k^{2} \) represent nonpropagating or evanescent waves. The propagation factor for an oblique wave is

\[
e^{i\sqrt{k^{2} - k_{\perp}^{2}} z} = e^{-\sqrt{k^{2} - k_{\perp}^{2}} z}.
\]

When \( k_{\perp}^{2} > k^{2} \) this is a decaying exponential. Although such waves do not propagate, and carry energy along \( \hat{z} \) they must be included in the Fourier integral to correctly describe the small scale features of the diffraction pattern. We can estimate the minimum spatial feature size as the interference pattern period created by two waves with \( k_{\perp} = \pm k\hat{\chi} \). The included angle is \( \pi \) and the interference period is \( \Lambda = \lambda/2 \). This simple estimate underpins the notion that the resolution limit of optics is of order the wavelength. Resolution below a wavelength is possible using nonlinear material response or using super-resolution techniques which trade off better resolution against reduced energy throughput.

As an alternative to the Fourier space transfer function we can use the Fourier convolution theorem to write the solution in the form of a convolution with the impulse response of free
Fourier optics

Figure 4.3: Paraxial evolution can be solved by convolution with the impulse response in the spatial domain or multiplication by the transfer function in the Fourier domain.

space or Green function. The Fourier convolution theorem says that if \( \tilde{A} = \tilde{B} \tilde{C} \) then \( A = \mathcal{F}^{-1}[\tilde{A}] = \frac{1}{2\pi} \mathcal{F}^{-1}[\tilde{B}] * \mathcal{F}^{-1}[\tilde{C}] \) so

\[
A_2(\rho_2) = \frac{1}{2\pi} A_1 * h \\
= \frac{-i k}{2\pi z} \int_{-\infty}^{\infty} d\rho_1 A_1(\rho_1) e^{i(k/2z)|\rho_2-\rho_1|^2} \tag{4.7}
\]

where the impulse response, or Green function, for free space propagation is

\[
h(\rho) = \mathcal{F}^{-1}[H] = \frac{-i k}{z} e^{i\rho^2/2\pi} \tag{4.8}
\]

It can be verified that Eqs. (4.5,4.7) are mathematically equivalent. The two methods for solving Eq. (4.2) are shown pictorially in Fig. 4.3. Either the real space or Fourier space approach can be used as they give identical results. The choice of how to perform a particular calculation is a matter of convenience and numerical efficiency. The availability of fast discrete transforms using the FFT algorithm often makes the Fourier space approach preferable. It is worth noting however that the impulse response form of Fresnel diffraction can be cast as a Fourier transform by expanding the exponential in (4.7) to get

\[
A_2(\rho_2) = \frac{-i k}{2\pi z} e^{i(k/2z)\rho_2^2} \int_{-\infty}^{\infty} d\rho_1 \left[A_1(\rho_1) e^{i(k/2z)\rho_1^2}\right] e^{-i(k/z)\rho_2 \cdot \rho_1} \\
= \frac{-i k}{z} e^{i(k/2z)\rho_2^2} \mathcal{F}\left[A_1(\rho_1) e^{i(k/2z)\rho_1^2}\right]_{k_{\perp,1}=(k/z)\rho_2} \tag{4.9}
\]

Thus, at the cost of multiplication by a quadratic phase, Fresnel diffraction can be calculated using a Fourier transform.

The solution given by Eqs. (4.5,4.7) is known as the Fresnel approximation or Fresnel diffraction. We have found an exact solution to the paraxial wave equation (4.2) which was
derived using the approximation $\partial A/\partial z \ll 2kA$. The Fresnel impulse response $h(\rho)$ is a paraxial approximation to an expanding spherical wave since

\[
(e^{-ikz}) \frac{1}{r} e^{-ikr} = e^{-ikz} \frac{1}{\sqrt{x^2 + y^2 + z^2}} e^{ik \sqrt{x^2 + y^2 + z^2}} \\
\approx \frac{1}{z} \left( 1 - \frac{x^2 + y^2}{2z^2} + \frac{3(x^2 + y^2)^2}{8z^4} \right) e^{ik \frac{x^2 + y^2}{2z}} e^{-ik \frac{x^2 + y^2}{2z}}. \tag{4.10}
\]

The prefactor of $e^{-ikz}$ accounts for the fact that we are solving for $A$ and $A = e^{-ikz} U$ where $U$ is the function that solves the Helmholtz equation. The paraxial approximation we have made above in arriving at Eq. (4.8) amounts to neglecting $(x^2 + y^2)/2z^2$ in the amplitude and $k(x^2 + y^2)^2/8z^3$ in the phase of a true spherical wave. This is valid provided

\[
k(x^2 + y^2)^2/8z^3 \ll \pi \text{ or } z \gg z_{\text{Fresnel}} = \left[ \frac{k}{8\pi} [(x_2 - x_1)^2 + (y_2 - y_1)^2]^{1/2} \right]^{1/3}. \tag{4.11}
\]

This limit is referred to as the Fresnel approximation. It is a sufficient condition for the use of Fresnel diffraction theory, but it may be overly restrictive in particular cases. For $\lambda = 1 \mu\text{m}$ and $|x_2 - x_1|_{\text{max}} = |y_2 - y_1|_{\text{max}} = 1 \text{ mm}$ we get $z_{\text{Fresnel}} \gg 10. \text{ mm}$.

### 4.1.2 Fraunhofer Diffraction

If we are willing to limit ourselves to even longer distances $z \gg z_{\text{Fresnel}}$ corresponding to smaller off-axis angles we can further simplify Eq. (4.7). For large $z$ we approximate the exponential factor under the integral by

\[
e^{i(k/2z)[(x_2 - x_1)^2 + (y_2 - y_1)^2]} = e^{i(k/2z)(x_1^2 + x_2^2 + y_1^2 + y_2^2)} e^{-i(k/z)x_1x_2 + y_1y_2} \\
\approx e^{i(k/2z)(x_2^2 + y_2^2)} e^{-i(k/z)(x_1x_2 + y_1y_2)}.
\]

Here we have discarded the quadratic phase in plane 1. This implies that the size of the source aperture is not too large. With this approximation we find

\[
A_2(\rho_2) = -\frac{i k}{2\pi z} e^{\frac{ik}{2z} \rho^2} \int_{-\infty}^{\infty} d\rho_1 A_1(\rho_1) e^{-i(k/z)\rho_1 \cdot \rho_2} \\
= -\frac{i k}{z} e^{\frac{ik}{2z} \rho^2} \mathcal{F} [A_1(\rho_1)]_{k_{\perp,1} = (k/z)\rho_2}. \tag{4.12}
\]

This is known as the Fraunhofer approximation. In this limit we see that apart from a quadratic phase factor, $A_2(x_2, y_2)$ is proportional to the Fourier transform of the input field. The ease with which optical systems calculate Fourier transforms has wide ranging consequences for optical signal processing, and we will return to this topic later on.

The Fraunhofer approximation is valid for

\[
z \gg z_{\text{Fraunhofer}} = (k/2\pi) \left( x_1^2 + y_1^2 \right)_{\text{max}}. \tag{4.13}
\]
For $\lambda = 1 \mu m$ and $|x_1|_{\text{max}} = |y_1|_{\text{max}} = 1 \text{ mm}$ we get $z_{\text{Fraunhofer}} \gg 2$. This approximation is two orders of magnitude more restrictive in the range of $z$ allowed than the Fresnel approximation. It will often turn out that both Fresnel and Fraunhofer approximations work quite well outside their strict validity limits. However, it is always necessary to check the results obtained in such a case. For the Fraunhofer approximation we can check by comparing the results with a Fresnel diffraction calculation. For Fresnel diffraction we would have to check against a more accurate approach, such as the Rayleigh-Sommerfeld formulation of diffraction discussed in Sec. 4.1.4 below.

### 4.1.3 Diffraction from an aperture

Let’s use these results to calculate the diffracted field after a rectangular aperture. For simplicity consider a one-dimensional geometry where the aperture has unit transmission for $-L/2 < x < L/2$ and is infinitely extended in $y$. Let the incident field be a uniform plane wave $A(\rho_1) = A_0$. The Fresnel diffraction pattern is

$$A(x_2) = -iA_0 \left( \frac{i}{2} \right)^{1/2} \int_{-L/2}^{L/2} dx \ e^{i \frac{\pi}{2} (x_2-x)^2}$$

where we have made the change of variable to

$$u^2 = \frac{k}{\pi z} (x-x_2)^2 , u_1 = \sqrt{k/(\pi z)} (-L/2-x_2), u_2 = \sqrt{k/(\pi z)} (L/2-x_2).$$

The remaining integral over $u$ can be written in terms of what are known as Fresnel integrals

$$C(s) = \int_0^s du \cos \left( \frac{\pi}{2} u^2 \right) , \ S(s) = \int_0^s du \sin \left( \frac{\pi}{2} u^2 \right).$$

These functions are plotted in Fig. 4.4. A parametric plot of $C(s)$ vs. $S(s)$ gives a spiral, known as the Cornu spiral, with accumulation points at $(0,0.5), (-0.5,0.5) \text{ for } s \to \pm \infty$.

The Cornu spiral was introduced to provide a graphical means of calculating the integrals, although with computers this is no longer needed.

In terms of Fresnel integrals the diffracted field is

$$A(x_2) = -iA_0 \left( \frac{i}{2} \right)^{1/2} \{ [C(u_2) - C(u_1)] + i[S(u_2) - S(u_1)] \} , \quad (4.14)$$

and the intensity is

$$I(x_2) = \frac{\varepsilon_0 c}{2} \frac{|A_0|^2}{2} \{ [C(u_2) - C(u_1)]^2 + [S(u_2) - S(u_1)]^2 \} .$$

Figure 4.5 shows the intensity at different $z$ values ranging from $5.4 - 540 \mu m$. These values can be compared with the Fresnel diffraction distance of (4.11) which we estimate,
4.1 Paraxial propagation and diffraction

Figure 4.4: Fresnel integrals $C(s)$, $S(s)$ and Cornu spiral.

using the parameters in the figure, to be

$$z_{\text{Fresnel}} = \left[ \frac{k}{8\pi} \left[ (x_2 - x_1)^2 + (y_2 - y_1)^2 \right]_{\text{max}} \right]^{1/3} \simeq \left[ \frac{k}{8\pi} \left[ L^2 + L^2 \right]_{\text{max}} \right]^{1/3} = 54. \, \mu\text{m.}$$

We see that all plots in the figure are safely inside the region of validity of the Fresnel approximation. The Fraunhofer approximation is valid for $z$ larger than

$$z_{\text{Fraunhofer}} = \left( \frac{k}{2\pi} \right) (x_1^2 + y_1^2)_{\text{max}} \simeq \left( \frac{k}{2\pi} \right) L^2 = 400. \, \mu\text{m.}$$

The last plot in the figure is in the Fraunhofer limit and indeed we see a smooth intensity profile which is close to the Fourier transform of the aperture function squared. Let’s check that the Fresnel and Fraunhofer expressions agree in this limit. The field in the Fraunhofer approximation is from (4.12)

$$A(x_2, y_2) = \frac{-ik}{2\pi z} A_0 e^{i\frac{k}{2\pi} (x_2^2 + y_2^2)} \int_{-\infty}^{\infty} dy \int_{-L/2}^{L/2} dx \, e^{-i(k/z)(xx_2 + yy_2)}$$

$$= \frac{-ik}{2\pi z} A_0 e^{i\frac{k}{2\pi} (x_2^2 + y_2^2)} \int_{-\infty}^{\infty} dy e^{-i(k/z)yy_2} \int_{-L/2}^{L/2} dx e^{-i(k/z)xx_2}.$$ 

The $y$ integral gives a $\delta$ function which is singular. To bypass this complication we write the field as

$$A(x_2, y_2) = A_0 \left[ \left( \frac{-ik}{2\pi z} \right)^{1/2} e^{i\frac{k}{2\pi} y_2^2} \int_{-\infty}^{\infty} dy e^{-i(k/z)yy_2} \right] \left[ \left( \frac{-ik}{2\pi z} \right)^{1/2} e^{i\frac{k}{2\pi} x_2^2} \int_{-L/2}^{L/2} dx e^{-i(k/z)xx_2} \right].$$

The first square bracket has all the $y, y_2$ dependence and for an aperture that is infinitely extended along $y$ this dependence vanishes. We will therefore write the effective one-
Figure 4.5: Diffracted intensity using Fresnel diffraction (blue solid line) and Fraunhofer diffraction (green dashed line) from a slit of width $L = 20. \mu\text{m}$, with $\lambda = 1. \mu\text{m}$ at different distances $z/z_{\text{Fresnel}} = 0.1, 0.5, 1, 5, 10$. The intensity has been normalized to the incident intensity at the aperture plane. The aperture parameters give $z_{\text{Fresnel}} = 54. \mu\text{m}$ and $z_{\text{Fraunhofer}} = 400. \mu\text{m}$.

dimensional solution as

$$A(x_2) = A_0 \left( \frac{-ik}{2\pi z} \right)^{1/2} e^{\frac{k}{2\pi z} x_2^2} \int_{-L/2}^{L/2} dx e^{-i(k/z)xx_2}$$

$$= A_0 e^{-i\pi/4} \left( \frac{2z}{\pi k} \right)^{1/2} \frac{e^{\frac{k}{2\pi z} x_2^2} \sin(kLx_2/2z)}{x_2}. \quad (4.15)$$

We can verify that this is correct by comparing with (4.14) in the limit of $z \to \infty$ where Fresnel and Fraunhofer diffraction calculations must agree. In this limit $u_1, u_2 \to 0$ and we can use the expansions $C(u) \simeq u, S(u) \sim \frac{\pi u^3}{6}$ to get

$$A_{\text{Fresnel}}(x_2) \simeq -iA_0 \left( \frac{i}{2} \right)^{1/2} \left( u_2 - u_1 \right) + i \left( \frac{\pi u_2^3}{6} - \frac{\pi u_1^3}{6} \right),$$

$$\simeq A_0 e^{-i\pi/4} \left( \frac{k}{2\pi z} \right)^{1/2} L,$$

where we have kept only the leading term in inverse powers of $z$. Expanding (4.15) we get for the Fraunhofer calculation

$$A_{\text{Fraunhofer}}(x_2) \simeq A_0 e^{-i\pi/4} \left( \frac{k}{2\pi z} \right)^{1/2} L,$$

which agrees with the Fresnel result. Comparing the solid and dashed lines in Fig. 4.5 we see that for $z \lesssim z_{\text{Fraunhofer}}/2$ the Fresnel and Fraunhofer calculations give very different results, while for large $z$ we get the same result from both approximations.
Beyond Fresnel diffraction

Figure 4.5 clearly shows that the results of Fresnel and Fraunhofer diffraction theory differ for $z \ll z_{\text{Fraunhofer}}$. On the other hand for $z \ll z_{\text{Fresnel}}$ we have no way of knowing how accurate the calculated intensity pattern is. To understand the limits of Fresnel diffraction theory we need a more accurate theory to compare it with. A more accurate theory of diffraction, although still limited to scalar fields, can be based on the Huygens-Fresnel theory whereby each point on the source wavefront acts as a source of spherical waves which interfere to create the observed diffraction field. A self-consistent and mathematically correct description of this basic idea turns out to be a remarkably difficult and subtle problem. Details can be found in M. Born & E. Wolf, *Principles of Optics*, or J. W. Goodman, *Introduction to Fourier Optics*.

Instead of pursuing a formal, more accurate solution of the Helmholtz equation let’s try and modify what we have so far to achieve better accuracy. Within the framework of the paraxial wave equation we arrived at an impulse response given by Eq. (4.8),

$$h(\rho) = -ik \frac{e^{ik\rho^2/2z}}{2z}.$$  

This impulse response differs from a true spherical wave by neglect of higher order terms in $\rho^2/z^2$ as was shown in (4.10). A first guess at an improved theory would be to simply replace our paraxial approximation to $h$ by

$$h_r = -ik \frac{e^{ikr}}{r}$$

where $r = |r_2 - r_1|$ is the three-dimensional distance from object point $r_1$ to image point $r_2$. Using this in (4.7) we get

$$A_2(\rho_2) = \frac{1}{2\pi} A_1 \ast h_r$$

$$= \frac{-ik}{2\pi} \int_\infty^{-\infty} d\rho_1 A_1(\rho_1) \frac{e^{ik|\rho_2 - \rho_1|}}{|\rho_2 - \rho_1|}$$

(4.16)

We are still assuming $A_1(\rho_1), A_2(\rho_2)$ to be planar fields but include the full three-dimensional distance $r$ in the integration. It turns out that (4.16) agrees with more accurate theories of diffraction apart from a missing obliquity factor $\chi$. In the Kirchhoff theory $\chi = \frac{1 + \cos(\hat{r}_{12} \cdot \hat{z})}{2}$ where $\hat{z}$ defines the optical axis, and $\hat{r}_{12}$ is a unit vector from object to image points. For small angles $\hat{r}_{12} \cdot \hat{z} \simeq 1$ so this factor is only important for large angular spreads. The Kirchhoff theory is not mathematically self-consistent and an improved approach due to Rayleigh and Sommerfeld yields $\chi = \cos(\hat{r}_{12} \cdot \hat{z})$. In practice it turns out that these two forms of the obliquity factor yield very similar results.

With the Rayleigh and Sommerfeld obliquity factor the diffraction integral reads

$$A_2(\rho_2) = \frac{1}{2\pi} A_1 \ast h_r$$

$$= \frac{-ik}{2\pi} \int_\infty^{-\infty} d\rho_1 A_1(\rho_1) \frac{e^{ik|\rho_2 - \rho_1|}}{|\rho_2 - \rho_1|} \cos(\hat{r}_{12} \cdot \hat{z}).$$

(4.17)

---

Figure 4.6 compares Fresnel diffraction with Eq. (4.17) for a square aperture. We see that at \( z = z_{\text{Fresnel}} \) there is perfect agreement in the intensity profile, at \( z = 0.5z_{\text{Fresnel}} \) small deviations are apparent, and at \( z = 1z_{\text{Fresnel}} \) large deviations appear. The Fresnel calculation over emphasizes the off-axis peaks at small \( z \) compared to the more accurate approach. This can be attributed to the missing obliquity factor which suppresses the large angle diffraction.

If we wish to improve the accuracy of the calculation even further it is necessary to include vectorial effects. This can be done rigorously using FDTD numerical simulations of Maxwell equations or approximately with analytical approaches due to Debye and Wolf\(^5\). A discussion of these methods is beyond what we wish to discuss here.

### 4.1.5 Fourier transforming properties of lenses

A lens of focal length \( f \) can be described as a phase plate with transmission function \( t = e^{-i(k/2f)(x_f^2+y_f^2)} \) where \( x_f, y_f \) are the coordinates in the plane of the lens. Let us place a lens at a distance \( L_1 \) after the input plane and calculate the output field a distance \( L_2 \) after the lens as shown in Fig. 4.7. Using Fresnel diffraction theory we find

\[
A_2(\rho_2) = \left( \frac{-i k}{2 \pi L_2} \right) \int \! d\rho_f A(\rho_f) e^{-i(k/2f)\rho_f^2} e^{i(k/2L_2)|\rho_2 - \rho_f|^2} \\
= \left( \frac{-i k}{2 \pi L_2} \right) \int \! d\rho_f e^{-i(k/2f)\rho_f^2} e^{i(k/2L_2)|\rho_2 - \rho_f|^2} \left( \frac{-i k}{2 \pi L_1} \right) \int \! d\rho_1 A(\rho_1) e^{i(k/2L_1)|\rho_f - \rho_1|^2},
\]

4.1 Paraxial propagation and diffraction

Figure 4.7: Field transformation with a lens.

with \( \rho = (x, y) \). Reversing the order of integration, we have for the integral over the lens coordinates

\[
\int d\rho f e^{-i(k/2)f}\rho^2 e^{i(k/2L_2)|\rho_2-\rho_f|^2} e^{i(k/2L_1)|\rho_f-\rho_1|^2} = e^{i(k/2L_2)\rho_2^2} e^{i(k/2L_1)\rho_1^2} \int d\rho e^{i\frac{1}{2}\rho^2 \left( \frac{1}{L_1} + \frac{1}{L_2} \right)} e^{-i\rho f \left( \frac{1}{L_1} + \frac{1}{L_2} \right)}
\]

\[
= i \frac{2\pi}{k \left( \frac{1}{L_1} + \frac{1}{L_2} - \frac{1}{f} \right)} \exp \left[ -\frac{ik}{2} \frac{\rho_2^2}{L_1} + \frac{\rho_2^2}{L_2} - f |\rho_1 - \rho_2|^2 / (L_1 L_2) \right].
\]

The output field is thus

\[
A(\rho_2) = -i \frac{k}{2\pi (L_1 + L_2 - L_1 L_2 / f)} \int d\rho_1 A(\rho_1) \exp \left[ -\frac{ik}{2} \frac{\rho_2^2}{L_1} + \frac{\rho_2^2}{L_2} - f |\rho_1 - \rho_2|^2 / (L_1 L_2) \right].
\]

This expression simplifies in the back focal plane at a distance \( L_2 = f \) after the lens where

\[
A_2(\rho_2) = -\frac{ik}{2\pi f} e^{\frac{ik^2}{2f}(1-L_1/f)} \int d\rho_1 A(\rho_1) e^{-i\frac{k}{f}\rho_1 \cdot \rho_2} = -\frac{ik}{f} e^{\frac{ik^2}{2f}(1-L_1/f)} \mathcal{F}[A(\rho_1)]_{k_\perp = \frac{k}{f} \rho_2}.
\]

Hence the output intensity is proportional to the spatial spectrum of the Fourier transform of the input field. When the input field is in the front focal plane of the lens \( (L_1 = f) \) the quadratic phase factor vanishes and

\[
A_2(\rho_2) = -\frac{ik}{f} \mathcal{F}[A(\rho_1)]_{k_\perp = \frac{k}{f} \rho_2} = -\frac{ik}{f} \int \int d\rho_1 A_1(\rho_1) e^{i\frac{k}{f}\rho_1 \cdot \rho_2}. \tag{4.18}
\]

We see that a single lens can be used to create an exact Fourier transform of an arbitrary paraxial field. This result is valid for both the Fresnel and Fraunhofer approximations.

March 8, 2016 M. Saffman
4.1.6 Fields with radial symmetry

When the field has radial symmetry the Fourier transform relations (4.3,4.4) can be simplified to one-dimensional integrals. Put \( x = \rho \cos \theta, \quad y = \rho \sin \theta, \quad dxdy = \rho d\rho d\theta \) and \( k_x = k_\perp \cos \phi, \quad k_y = k_\perp \sin \phi, \quad dk_xdk_y = k_\perp dk_\perp d\phi \) to get

\[
\tilde{A}(k_\perp, \phi) = \frac{1}{2\pi} \int_0^\infty d\rho \rho A(\rho) \int_0^{2\pi} d\theta e^{-i\rho k_\perp (\cos \phi \cos \theta + \sin \phi \sin \theta)} \]

\[
= \frac{1}{2\pi} \int_0^\infty d\rho \rho A(\rho) \int_0^{2\pi} d\theta e^{-i\rho k_\perp \cos(\phi - \theta)}. \]

The angular integral can be expressed as

\[
\int_0^{2\pi} d\theta e^{-i\rho k_\perp \cos(\phi - \theta)} = \int_0^{2\pi} d\theta e^{-i\rho k_\perp \cos(\theta)} = 2\pi J_0(\rho k_\perp). \]

We see that \( \tilde{A} \) depends only on \( k_\perp \) and can be written as

\[
\tilde{A}(k_\perp) = \int_0^\infty d\rho \rho A(\rho) J_0(\rho k_\perp) \tag{4.19} \]

with the inverse transform

\[
A(\rho) = \int_0^\infty dk_\perp k_\perp \tilde{A}(k_\perp) J_0(\rho k_\perp). \tag{4.20} \]

Thus the Fourier transform for problems with radial symmetry has a Bessel function kernel. This is referred to as a Hankel transform.

The analog of (4.18) for the field transformation by a lens is

\[
A_2(\rho_2) = -\frac{ik}{f} \mathcal{F}[A_1(\rho_1)]_{k_\perp = \frac{k_\perp}{f}} \]

\[
= -\frac{ik}{f} \int d\rho_1 \rho_1 A_1(\rho_1) J_0(\rho_1 \rho_2 k/f). \tag{4.21} \]

4.1.7 Airy disk

A first example of the use of the Hankel transform arises in computing the focal plane pattern formed by uniform illumination of a circular lens of diameter \( d \). Using (4.21) for a circularly symmetric field we find

\[
A(\rho_2) = -\frac{ik}{f} \int_0^{d/2} d\rho_1 \rho_1 J_0 \left( \frac{k \rho_1 \rho_2}{f} \right) \]

\[
= -id \frac{J_1 \left( \frac{dk \rho_2}{2f} \right)}{2\rho_2}. \tag{4.22} \]
4.1 Paraxial propagation and diffraction

The $J_1(x)$ Bessel function has a first zero away from the origin at $x_1 = 3.83$ so the field in the back focal plane vanishes at

$$\rho_2 = 7.66 \frac{f}{d} = 1.22 \frac{\lambda f}{d}.$$ 

This result was first obtained by Airy in 1835\textsuperscript{6}. The intensity which is proportional to $|A(\rho_2)|^2$ is plotted in Fig. 4.8. The total power in the Airy diffraction pattern is

$$P = \frac{\varepsilon_0 c}{2} \int_0^\infty d\rho_2 2\pi \rho_2 A(\rho_2)^2 = \frac{\varepsilon_0 c \pi d^2}{4},$$

which is just the intensity intercepted by the lens. The power inside the first zero is

$$P_0 = \frac{\varepsilon_0 c}{2} \int_0^{2f x_1 / dk} d\rho_2 2\pi \rho_2 A(\rho_2)^2 = \frac{\varepsilon_0 c}{2} 0.66d^2.$$

A fraction of $0.66/(\pi/4) = 0.84$ of the energy is focused in the central spot.

For an annulus of inner and outer radii $a_1, a_2$ we get

$$A(\rho_2) = -ia_2 \frac{J_1 \left( \frac{a_2 \kappa \rho_2}{f} \right)}{\rho_2} + ia_1 \frac{J_1 \left( \frac{a_1 \kappa \rho_2}{f} \right)}{\rho_2}$$

$$= a_1 J_1 \left( \frac{a_1 \kappa \rho_2}{f} \right) - a_2 J_1 \left( \frac{a_2 \kappa \rho_2}{f} \right) \frac{i}{\rho_2}. $$


March 8, 2016 M. Saffman
Figure 4.9: Spot of Arago calculated for $\lambda = 1$ $\mu$m, $a = 10$ $\mu$m, giving $z_{\text{Fresnel}} = 54$ $\mu$m. The diffracted intensity normalized to the incident intensity is shown for $z/z_{\text{Fresnel}} = 1, 2, 10, 20$.

Referring to Fig. 4.8 we see that the annulus has the effect of suppressing the central intensity and putting relatively more energy into the rings.

The ring structure sets the diffraction limited angular resolution of a lens. Suppose we use a lens to view two distant stars with an angular separation of $\theta$. The images in the back focal plane will be separated by $\delta \rho_2 = \sin(\theta) f$. If we say that the first zero of one image coincides with the peak of the adjacent image then there will be a visible dip of the total intensity in between the images. This is shown in the inset to the figure and is referred to as the Rayleigh resolution limit. This limit occurs when $\sin(\theta) f = 1.22 \lambda f / d$ or

$$\sin(\theta) = 1.22 \frac{\lambda}{d}.$$ 

This expression gives the angular resolution limit of a telescope according to the Rayleigh criterion. This is not the ultimate resolution limit of a telescope since the dip is clearly visible at the Rayleigh separation.

### 4.1.8 Spot of Arago

An interesting phenomenon appears if we consider the diffracted field without a focusing lens behind a uniformly illuminated circular plate which blocks the light for $0 \leq \rho \leq a$. With circular symmetry Eq. (4.7) for Fresnel diffraction becomes

$$A(\rho_2) = \frac{-ik}{2\pi z} \int_{-\infty}^{\infty} d\rho A(\rho)e^{i(k/2z)(\rho_2 - \rho)^2}$$

$$= \frac{-i k}{z} e^{\frac{k\rho_2^2}{2z}} \int_{0}^{\infty} d\rho A(\rho)\rho e^{i\frac{k\rho^2}{2z}} J_0 \left( \frac{k\rho_2\rho}{z} \right).$$

(4.23)

For the uniformly illuminated circular plate we get

$$A(\rho_2) = \frac{-i k}{z} e^{\frac{k\rho_2^2}{2z}} \int_{0}^{\infty} d\rho \rho e^{i\frac{k\rho^2}{2z}} J_0 \left( \frac{k\rho_2\rho}{z} \right)$$

$$= \frac{-i k}{z} e^{\frac{k\rho_2^2}{2z}} \int_{0}^{\infty} d\rho \rho e^{i\frac{k\rho^2}{2z}} J_0 \left( \frac{k\rho_2\rho}{z} \right) + \frac{i k}{z} e^{i\frac{k\rho^2}{2z}} \int_{0}^{a} d\rho \rho e^{i\frac{k\rho^2}{2z}} J_0 \left( \frac{k\rho_2\rho}{z} \right).$$

The infinite range integral is evaluated using $\int_{0}^{\infty} dx x e^{ax^2} J_0(bx) = \frac{i}{2a} e^{-ib^2/4a}$ to give

$$A(\rho_2) = 1 + \frac{i k}{z} e^{\frac{k\rho_2^2}{2z}} \int_{0}^{a} d\rho \rho e^{i\frac{k\rho^2}{2z}} J_0 \left( \frac{k\rho_2\rho}{z} \right).$$
4.1 Paraxial propagation and diffraction

When there is no blocking plate, \( a = 0 \) and \( A(\rho_2) = 1 \) as expected. Surprisingly when there is a plate the intensity is always nonzero at the origin. We can see this by calculating

\[
A(0) = 1 + \frac{ik}{z} \int_0^a d\rho \rho e^{i k \rho^2/2z} = e^{i k a^2/2z}.
\]

For all values of \( z \) there is unit intensity on axis behind the blocking disk. This is known as the spot of Arago and is shown in Fig. 4.9. Due to diffractive spreading the spot is only clearly visible for \( z \) up to about \( 10z_{\text{Fresnel}} \).

4.1.9 Fresnel zones

A phenomenon related to that of the previous section is the possibility of focusing without using a lens. Consider a uniformly illuminated source plane with radial coordinate \( \rho \) and an image plane a distance \( z \) after it with radial coordinate \( \rho_2 \). On axis in the image plane \( \rho_2 = 0 \) and Eq. (4.23) predicts a field

\[
A(0, z) = \frac{ik}{z} A_0 \frac{1}{2\pi} \int_0^\infty d\rho 2\pi \rho e^{i k \rho^2/2z}
\]

where \( A_0 \) is the field at the source plane. The integrand has a phase which starts at 0 and grows proportional to \( \rho^2 \). The phase reaches the value \( p\pi \) for \( a_p = \sqrt{p\pi z/k} = \sqrt{p\lambda z} \). The source area between radii \( a_{p-1} \) and \( a_p \) is

\[
\sigma = \pi a_p^2 - \pi a_{p-1}^2 = \pi \lambda z[p - (p - 1)] = \pi \lambda z,
\]

which is constant. We see that the source plane can be divided into zones of equal area, with the diffracted phase at the image plane varying by \( \pi \) from zone to zone. These are known as Fresnel zones.

The contributions to the diffracted field on axis from neighboring zones tend to cancel out since they have opposite signs. However, if we block every other zone, as in Fig. 4.10, the zones which are transmitted all add constructively and a large intensity will be seen on-axis. This construction is known as a Fresnel zone plate.
We can estimate the on-axis intensity as follows. The \( p \)th zone contributes a power
\[
\frac{\varepsilon_0 c}{2} |A_0|^{2} \sigma |\chi_p|^2
\]
where
\[
\chi_p = \frac{1}{\sigma} \int_{a_{p-1}}^{a_p} d\rho 2\pi \rho e^{i k \rho^2} \\
= \frac{2i}{\pi} e^{i \pi p}
\]
so \( |\chi_p|^2 = 4/\pi^2 \simeq 0.405 \). Due to the phase averaging within each zone the lens is only about 40% efficient. The integrated intensity from \( N \) zones is thus
\[
I_N = \frac{\varepsilon_0 c}{2} |A_0|^2 \frac{4N \sigma}{\pi^2 \sigma_f} 
\]
where \( \sigma_f \) is the area of the focused spot. Relative to the incident intensity \( I_0 = \frac{\varepsilon_0 c}{2} |A_0|^2 \) we find
\[
\frac{I_N}{I_0} = N \frac{4\lambda z}{\pi \sigma_f}
\]
A calculation of \( \sigma_f \) can be made using Fresnel integrals as in the spot of Arago calculation. Assuming \( k p \rho / z \ll 1 \) we expand the Bessel function as \( J_0(x) \simeq 1 - x^2/4 \) and find that the field from the \( p \)th zone vanishes at \( \rho_2 \simeq \sqrt{\lambda z / (\pi^2 p)} \) for \( p \gg 1 \). The focal spot size decreases at large \( p \). The limit on the focusing is set by \( p = 1 \) for which case the field vanishes at \( \rho_2 = \sqrt{2\lambda z / \pi^2} \) and we can estimate the focal spot area as \( \sigma_f \sim \pi(2\lambda z / \pi^2) = (2/\pi)(\lambda z) \)
\[
\frac{I_N}{I_0} \sim 2N
\]
for \( N \) alternating zones added together. While the zone plate does not produce better focusing than a refractive lens it is useful for wavelengths where it is difficult to produce materials with desired refractive properties. This is the case in x-ray optics where Fresnel zone plates are commonly used for focusing.

For a given aperture of radius \( a \) the Fresnel number \( N_{\text{Fresnel}} \) is defined as the number of Fresnel zones seen in the image plane. This is given by \( \sqrt{N_{\text{Fresnel}} \lambda z} = a \) or
\[
N_{\text{Fresnel}} = \frac{a^2}{\lambda z}
\]
Note that \( N_{\text{Fresnel}} = 1 \) corresponds to \( z = a^2/\lambda \sim z_{\text{Fraunhofer}} \) (see Eq. (4.13)). Put another way if the source aperture has \( N_{\text{Fresnel}} \leq 1 \) then Fraunhofer diffraction is adequate to calculate the image field. Conversely if \( N_{\text{Fresnel}} > 1 \) we must use Fresnel diffraction theory.

4.1.10 Talbot effect

Another interesting piece of Fourier optics is the Talbot effect which results in the phenomenon of lensless imaging of periodic objects. Consider an optical field \( A(x,z) \) that is periodic along \( x \) with period \( L_x \). This could be a field with periodic amplitude, phase, or both. Periodicity implies that the field can be written as
\[
A(x,z_0) \sim \cos(2\pi x/L_x) \sim e^{2\pi x/L_x} + e^{-i2\pi x/L_x}.
\]
From Eq. (4.24) we identify the field as being due to the interference of plane waves with transverse wave numbers \( k_x = \pm 2\pi/L_x \). Transverse wave numbers that are a factor \( m \) larger, with \( m \) an integer, will also synthesize a field with period \( L_x \) so a general periodic field can be written as

\[
A(x, z_0) = \sum_{m=1}^{\infty} c_m e^{imk_x x} + c_m^* e^{-imk_x x} \tag{4.25}
\]

with the \( c_m \) complex amplitudes.

Propagation of the field (4.25) an axial distance \( L_z \) will add quadratic phases to the plane wave components giving a new field

\[
A(x, z_0 + L_z) = \sum_{m=1}^{\infty} (c_m e^{imk_x x} + c_m^* e^{-imk_x x}) e^{i m^2 k_x^2 x^2 / 2k L_z} \tag{4.26}
\]

The new field will be identical to the original field when \( m^2 k_x^2 / 2k L_z = m^2 2\pi \) or

\[
L_z = \frac{4\pi k}{k_x^2} = \frac{2L_x^2}{\lambda} \equiv L_{\text{Talbot}}.
\]

We see that the periodic field distribution exactly reproduces itself at multiples of the Talbot length \( L_{\text{Talbot}} \). This phenomenon was discovered by Talbot\(^7\), and has been studied with both optical and matter waves.

It is instructive to rewrite the field with \( z_0 = 0 \) as

\[
A(x, z) = \sum_{m=1}^{\infty} (c_m e^{imk_x x} + c_m^* e^{-imk_x x}) e^{i 2\pi m^2 z / L_{\text{Talbot}}} \tag{4.27}
\]

which makes explicit the periodicity at the Talbot length. At subharmonics of the Talbot length there are other effects. At \( z = L_{\text{Talbot}}/2 \) the pattern is shifted half a period along \( x \). In addition if the field is weakly amplitude modulated and the modulation is an even function of \( x \) then at \( z = L_{\text{Talbot}}/4 \) the amplitude modulation is converted into a pure phase modulation. At other subharmonic distances higher frequency copies of the original field are observed. A more extensive discussion can be found in review articles\(^8\),\(^9\).

### 4.1.11 Bottle beam lenses

As another example of the use of the Hankel transform let’s consider a thin lens of focal length \( f \) with radius \( a \) and an inner region of radius \( b < a \) which has a \( \pi \) phase shift relative to the outer as shown in Fig. 4.11. We will illuminate the lens with a Gaussian beam with waist \( w_0 \) in the front focal plane. This configuration can be used to create a Bottle Beam optical trap (BBT). We can develop an analytical description of the field in the focal region using Fresnel diffraction theory.


Call the field in the front focal plane $A_0$ then at the lens the field is
\[ A_1 = \mathcal{F}^{-1}[\mathcal{F}[A_0(\rho)]H(k_\perp, f)] \]
with $H(k_\perp, z) = e^{-i\frac{k^2z^2}{2f}}$. Passage through the thin lens with two annular regions is accounted for by multiplying with the transmission function
\[ t(\rho) = e^{-i\frac{\rho^2}{2f}} [1 - 2 \text{circ}(\rho/b)] \]
where we have introduced the radial step function $\text{circ}(\rho) = 1$ for $\rho \leq 1$ and $\text{circ}(\rho) = 0$ for $\rho > 1$. We will assume the field does not extend to the outer boundary of the lens and ignore the radius $a$. We thus get $A_2(\rho) = tA_1(\rho)$ and the field in the output plane a distance $z$ after the lens is given by
\[ A_3(\rho, z) = \mathcal{F}_{32}^{-1}[\mathcal{F}_{22}[A_2(\rho_2)]H(k_{\perp 2}, z)] \]
\[ = \mathcal{F}_{32}^{-1}[\mathcal{F}_{22}[t(\rho_1)\mathcal{F}_{10}^{-1}[\mathcal{F}_{00}[A_0(\rho_0)]H(k_{\perp 0}, f)]H(k_{\perp 2}, z)]. \] (4.28)
Subscripts on the variables and operators have been introduced to indicate different transverse coordinates corresponding to the different planes in Fig. 4.11. Writing out (4.28) explicitly the output field can be expressed as
\[ A_3(\rho, z_3) = \int dk_{\perp 1} k_{\perp 1} J_0(\rho_3 k_{\perp 1})e^{-i\frac{k_{\perp 1}^2z_3}{2f}} \int dp_1 p_1 J_0(p_1 k_{\perp 1})e^{-i\frac{p_1^2}{2f}} [1 - 2 \text{circ}(\rho_1/b)] \]
\[ \times \int dk_{\perp 0} k_{\perp 0} J_0(\rho_0 k_{\perp 0})e^{-i\frac{k_{\perp 0}^2}{2f}} \int dp_0 p_0 J_0(p_0 k_{\perp 0})A_0(\rho_0). \] (4.29)

The factor $[1 - 2 \text{circ}(\rho_1/b)]$ splits the result into two terms. The first one, which is independent of $b$, results in a Gaussian beam with waist $w_3 = \lambda f/(\pi w_0)$ at $z_3 = f$. The second term is more complicated and can be written as
\[ A_{3b} = -\frac{2A_0}{1 + i w_3/w_0} \int_0^\infty dk_{\perp 1} k_{\perp 1} J_0(\rho_3 k_{\perp 1})e^{-i\frac{w_3 w_0 k_{\perp 1}^2}{4}} \int_0^b dp_1 p_1 J_0(k_{\perp 1} p_1)e^{-i\rho_1^2} \]
with \( h = k^2 w_0^2 / (4f^2 - 2fkw_0^2) = 1/(w_0^2 - iw_0 w_3) \). We have also introduced a normalized axial coordinate \( s = z_3 / f \), so that \( s = 1 \) corresponds to the back focal plane. Reversing the order of integration gives
\[
\int_0^\infty dk_{1\perp} k_{1\perp} J_0(\rho_3 k_{1\perp}) e^{-\frac{\omega_3 w_{3z}k_{1\perp}^2}{4}} \int_0^b d\rho_1 \rho_1 J_0(k_{1\perp}\rho_1) e^{-h\rho_1^2} = \int_0^b d\rho_1 \rho_1 e^{-h\rho_1^2} \int_0^\infty dk_{1\perp} k_{1\perp} J_0(\rho_3 k_{1\perp}) J_0(k_{1\perp}\rho_1) e^{-\frac{\omega_3 w_{3z}k_{1\perp}^2}{4}}. \]

To evaluate the integral over \( dk_{1\perp} \) use (Watson, p.395)
\[
\int_0^\infty dz z J_\nu(az) J_\nu(bz) e^{-p z^2} = \frac{1}{2p} e^{-\frac{a^2+b^2}{4p}} I_\nu(\frac{ab}{2p}). \tag{4.30}
\]
Here \( I_\nu \) is a modified Bessel function and the result is valid provided \( \text{Re}(\nu) > -1/2 \) and \( \text{Re}(p) > 0 \). In our case \( p = i\omega_0 w_3 s/4 \) and \( \text{Re}(p) = 0 \) which violates the condition on \( p \) but (4.30) is still valid since we have\(^{10} \) \( \text{Re}(a) > 0, \text{Re}(b) > 0 \). Proceeding with \( a = \rho_1, b = \rho_3, \) and \( p = i\omega_0 w_3 s/4 \) we find
\[
A_{3b} = 4i \frac{A_0}{w_3 s(w_0 + i w_3)} e^{i \frac{\omega_3^2}{w_0 w_3} s} \int_0^b d\rho_1 \rho_1 e^{(-h + \frac{\omega_0 w_{3s}}{w_0 w_3})\rho_1^2} J_0 \left( \frac{2\rho_1 \rho_3}{w_0 w_3} \right).
\]
This last integral arises in the problem of diffraction from a finite circular aperture and can be expressed in terms of Lommel functions of two variables (Watson p.540)
\[
U_1(u, v) = u \int_0^1 dt t J_0(\sqrt{u^2 - t^2}) \tag{4.31a}
\]
\[
U_2(u, v) = u \int_0^1 dt t J_0(\sqrt{u^2 - t^2}) \sin \left( \frac{u}{2} (1 - t^2) \right). \tag{4.31b}
\]
The field is then
\[
A_{3b} = 4i \frac{A_0 b^2 e^{-u/2}}{w_3 s(w_0 + i w_3) u} e^{i \frac{\omega_3^2}{w_0 w_3} s} [U_1(u, v) + i U_2(u, v)]
\]
\(^{10}\)To show this use (G \& R 6.729, 1,2)
\[
\int_0^\infty dz z J_\nu(az) J_\nu(bz) \cos(p' z^2) = \frac{1}{2p'} \sin(\frac{a^2 + b^2}{4p'} - \frac{\nu \pi}{2}) J_\nu(\frac{ab}{2p'}),
\]
\[
\int_0^\infty dz z J_\nu(az) J_\nu(bz) \sin(p' z^2) = \frac{1}{2p'} \cos(\frac{a^2 + b^2}{4p'} - \frac{\nu \pi}{2}) J_\nu(\frac{ab}{2p'}),
\]
provided \( \text{Re}(\nu) > -2, \text{Re}(a) > 0, \text{Re}(b) > 0, \text{Re}(p') > 0 \). These can be combined to give
\[
\int_0^\infty dz z J_\nu(az) J_\nu(bz) e^{-p' z^2} = -\frac{i}{2p'} e^{(\frac{a^2+b^2}{4p'} - \frac{\nu \pi}{2})} J_\nu(\frac{ab}{2p'}),
\]
and putting \( p = ip' \) we get
\[
\int_0^\infty dz z J_\nu(az) J_\nu(bz) e^{-p z^2} = \frac{1}{2p'} e^{\frac{a^2+b^2}{4p'} - \frac{\nu \pi}{2}} J_\nu(\frac{ab}{2p'}) = \frac{1}{2p'} e^{\frac{a^2+b^2}{4p'} - \frac{\nu \pi}{2}} I_\nu(\frac{ab}{2p'}),
\]
which is the same as (4.30) even though \( \text{Re}(p) = 0 \). Thus (4.30) can be extended to the case of \( \text{Re}(p) = 0 \) provided \( \text{Re}(a) > 0, \text{Re}(b) > 0 \).

March 8, 2016 M. Saffman
Figure 4.12: BoB profiles for $w_0 = 82.76 \, \mu m$, $w_3 = 3 \, \mu m$, $b = 69 \, \mu m$ and $s = 1$ (left). These parameters correspond to $\lambda = 0.78 \, \mu m$ and $f = 1 \, mm$. The plotted intensity is normalized to the peak of the input Gaussian with waist $w_0$. The inset shows that the on-axis intensity does not vanish for any $b$. On the right the transverse profiles are shown at axial displacements up to $50 \, \mu m$ from the focal plane.

with

$$u = -2b^2 w_3^2 - w_0^2 (s - 1) + iw_0 w_3 s,$$

$$v = \frac{2b\rho_3}{w_0 w_3 s}.$$

This completes the calculation of the field of the bottle beam. To evaluate the field and intensity we must numerically evaluate the Lommel functions. This can be done either by numerical integration of (4.31) or by evaluation of the functions expressed as infinite sums of Bessel functions (Watson p. 537)

$$U_1(u, v) = \sum_{m=0}^{\infty} (-1)^m \left(\frac{u}{v}\right)^{1+2m} J_{1+2m}(v),$$

$$U_2(u, v) = \sum_{m=0}^{\infty} (-1)^m \left(\frac{u}{v}\right)^{2+2m} J_{2+2m}(v).$$

Let us calculate the intensity of the bottle beam as a function of $\rho_3$ at the focal plane $z_3 = f$ or $s = 1$. We find

$$I(\rho_3, s = 1) = \frac{\varepsilon_0 c}{2} \left| -i A_0 \frac{w_0}{w_3} e^{-\rho_3^2/w_3^2} + A_3 b \right|^2.$$

The factor of $-i$ in front of the Gaussian is a propagation phase due to the Gouy term in the expression for a Gaussian beam. The intensity is plotted in Fig. 4.12 for some representative parameters. Note that the on-axis intensity does not go to zero for any value of $b$ due to the small diffractive phase shift which prevents perfect field cancellation. This could be corrected for by using a phase shift that is slightly different from $\pi$ (analysis to be added). As the axial plane is moved away from the focus the on-axis intensity grows. We see that the minimum of the potential is about 2/3 of the radial peak in the focal plane. Figure 4.13 shows that the intensity profile is not symmetric about the focal plane as $z$ is varied.
The on-axis intensity at $\rho_3 = 0$ should be as small as possible for a bottle beam. On axis

$$A_{3b}(0) = 4i \frac{A_0 b^2 e^{-u/2}}{w_3 s (w_0 + iw_3) u} [U_1(u, 0) + iU_2(u, 0)]$$

$$= 4 \frac{A_0 b^2}{w_3 s (w_0 + iw_3) u} (1 - e^{-u/2}),$$

where we have used $U_1(u, 0) = \sin(u/2), U_2(u, 0) = 2\sin^2(u/4)$. The intensity is thus zero when

$$e^{-u/2} = 1 + i \frac{w_3^2 - w_0^2 (s - 1) + iw_0 w_3 s}{2w_3 (w_0 - iw_3)}.$$ (4.33)

Solving for $u$ and then $b$ we verify that there are no solutions for $s = 1$ and real $b$. We can get an approximate solution for $b$ in the limit of $w_3 \ll w_0$ and $s = 1$ which gives for the example in Fig. 4.12

$$\text{Re}(b) \approx 71.3 \ \mu m + O((w_3/w_0)^2)$$

provided we choose the appropriate branch of the solution to (4.33). This is close to the numerically found optimum of $b = 69. \ \mu m$.

### 4.1.12 More general transformations

We can generalize Fresnel diffraction theory to the propagation of a paraxial field through an arbitrary optical system described by an ABCD ray matrix. The output field can be written in terms of a Green function as

$$A_2(\rho_2) = \int \int d\rho_1 \ G(\rho_2; \rho_1) A_1(\rho_1),$$ (4.34)
where the Green function is given by $^{11,12}$

$$G(\rho_2; \rho_1) = \frac{-ik}{2\pi B} \exp \left[ i \frac{k}{2B} \left( A\rho_1^2 + D\rho_2^2 - 2\rho_1 \cdot \rho_2 \right) \right]$$  \hspace{1cm} (4.35)

and the ray matrix describing propagation from planes $1 \rightarrow 2$ is

$$M = \begin{pmatrix} A & B \\ C & D \end{pmatrix}.$$  \hspace{1cm} (4.36)

In imaging systems the matrix element $B$ is zero in which case Eq. (4.35) cannot be used directly. To proceed start by assuming $B \neq 0$, and using $AD - BC = 1$ rewrite (4.35) as

$$G(\rho_2; \rho_1) = \frac{-ik}{2\pi B} e^{i\frac{kC}{2A}\rho_2^2} \exp \left[ i \frac{kC}{2A} \left( x_1 - \frac{x_2}{A} \right)^2 + \left( y_1 - \frac{y_2}{A} \right)^2 \right].$$  \hspace{1cm} (4.37)

Taking the limit as $B \rightarrow 0$ and using $\delta(x) = \lim_{\epsilon \rightarrow 0} \frac{1}{\sqrt{\pi \epsilon}} e^{\frac{x^2}{\epsilon}}$ we find

$$G(\rho_2; \rho_1) = D e^{i\frac{kCD}{2}\rho_2^2} \delta(x_1 - x_2/A) \delta(y_1 - y_2/A).$$  \hspace{1cm} (4.38)

$^{11}$P. Baues, Huygens’ principle in inhomogeneous isotropic media and a general integral equation applicable to optical resonators, Opto-Electr. 1, 37 (1969).

4.1.13 Summary of diffraction formulae

In this section we list the most important formulae for diffraction calculations together with the equation numbers where they were defined.

Fresnel diffraction using transfer function, 2D Cartesian coordinates:

\[
A_2(\rho_2) = F^{-1}[F[A_1(\rho_1)]H(k_\perp, z)]
= \frac{1}{4\pi^2} \int_{-\infty}^{\infty} d\mathbf{k}_\perp e^{i\mathbf{k}_\perp \cdot \mathbf{\rho}_2} e^{-i\frac{k^2}{2z}(z_2 - z_1)} \int_{-\infty}^{\infty} d\rho_1 A_1(\rho_1) e^{-i\mathbf{k}_\perp \cdot \mathbf{\rho}_1} \tag{4.5}
\]

\[
H(k_\perp, z) = e^{-i\frac{k^2}{2z}}. \tag{4.6}
\]

Fresnel diffraction using impulse response, 2D Cartesian coordinates:

\[
A_2(\rho_2) = \frac{1}{2\pi} A_1 * h = \frac{-ik}{2\pi z} \int_{-\infty}^{\infty} d\rho_1 A_1(\rho_1)e^{i(k/2z)|\mathbf{\rho}_2 - \mathbf{\rho}_1|^2} \tag{4.7}
= \frac{-ik}{z} e^{i(k/2z)\rho_2^2} F \left[ A_1(\rho_1) e^{i(k/2z)\rho_1^2} \right]_{k_\perp,1=(k/z)\rho_2} \tag{4.9}
\]

\[
h(\rho) = F^{-1}[H] = \frac{-ik}{z} e^{i\frac{k}{2}\rho^2}. \tag{4.8}
\]

Fresnel diffraction using impulse response, radial symmetry:

\[
A_2(\rho_2) = \frac{-ik}{z} e^{i\frac{k}{2}\rho_2^2} \int_{0}^{\infty} d\rho_1 A_1(\rho_1)\rho_1 e^{i\frac{k}{2z}\rho_1^2} J_0 \left( \frac{k\rho_1\rho_2}{z} \right). \tag{4.23}
\]

Fraunhofer diffraction, 2D Cartesian coordinates:

\[
A_2(\rho_2) = \frac{-ik}{2\pi z} e^{i\frac{k}{2z}\rho_2^2} \int_{-\infty}^{\infty} d\rho_1 A_1(\rho_1) e^{-i(k/z)\rho_1 \cdot \rho_2} \tag{4.12}
= \frac{-ik}{z} e^{i\frac{k}{2z}\rho_2^2} F [A_1(\rho_1)]_{k_\perp,1=(k/z)\rho_2}.
\]

Fraunhofer diffraction, 1D:

\[
A_2(x_2) = \left( \frac{-ik}{2\pi z} \right)^{1/2} e^{i\frac{k}{2z}x_2^2} \int_{-\infty}^{\infty} dx_1 A_1(x_1) e^{-i(k/z)x_1x_2} \tag{4.15}
\]

Fourier transformation by lens:

\[
A_2(\rho_2) = -\frac{ik}{f} F [A_1(\rho_1)]_{k_\perp=-\frac{k}{f}}. \tag{4.18}
\]
4.2 Fourier optical image processing

Sophisticated image processing tasks can be performed rapidly using optical transformations. A few examples are given in this section.

4.2.1 Repeated objects

The Fourier transformation properties of lenses can be used for image processing. Let's start by considering the Fourier transform of a repeated object. An important example of a repeated object is a diffraction grating which is useful for determining the amount of energy as a function of frequency or wavelength in an optical beam.

Consider an object \( A(\rho) \) which is some arbitrary optical field that depends on the two-dimensional coordinate \( \rho = x\hat{x} + y\hat{y} \). A displaced version of the object centered at \( \rho_j \) can be written as

\[
A(\rho - \rho_j) = A(\rho) * \delta(\rho - \rho_j) = \int d\rho' A(\rho') \delta(\rho' - (\rho - \rho_j)) = A(\rho - \rho_j).
\]

Convolving any well behaved function with a delta function centered at \( \rho_j \) gives a copy of the function displaced by \( \rho_j \).

Suppose that we have \( N \) such displaced copies of the object centered at coordinates \( \rho_j, \ j = 1, N \). Using the convolution theorem for Fourier transforms the transform of the object field is

\[
\mathcal{F} \left[ \sum_j A(\rho - \rho_j) \right] = 2\pi \sum_j \mathcal{F}[A(\rho) * \delta(\rho - \rho_j)]
\]

\[
= 2\pi \mathcal{F}[A(\rho)] \sum_j \mathcal{F}[\delta(\rho - \rho_j)]
\]

\[
= 2\pi \tilde{A}(k_\perp) \sum_j \frac{1}{2\pi} e^{-i\rho_j k_\perp}
\]

\[
= \tilde{A}(k_\perp) \sum_j e^{-i\rho_j k_\perp}.
\]

If the copies of the object are uniformly spaced on a regular grid separated by a displacement vector \( h \) we can put \( \rho_j = jh \) and the sum becomes

\[
\sum_{j=1}^{N} e^{-i\rho_j k_\perp} = \sum_{j=1}^{N} e^{-ijh k_\perp}
\]

\[
= e^{-ih k_\perp N} \frac{e^{ih k_\perp N} - 1}{e^{ih k_\perp} - 1}.
\]
The intensity of the transformed object will thus be

\[ I = |\tilde{A}(k_\perp)|^2 \left| \frac{e^{ih\cdotk_\perp N} - 1}{e^{ih\cdotk_\perp} - 1} \right|^2 \]

\[ = |\tilde{A}(k_\perp)|^2 \frac{\sin^2(h \cdot k_\perp N/2)}{\sin^2(h \cdot k_\perp/2)}. \] (4.39)

Equation (4.39) can be understood as follows. The Fourier transform of \( N \) copies of an object, uniformly spaced by multiples of \( h \) is the Fourier transform of the object multiplied by a function \( \sin^2(\alpha N)/\sin^2(\alpha) \) with \( \alpha = h \cdot k_\perp/2 \). If \( \tilde{A}(k_\perp) \) has a maximum at the origin then \( N \) copies of the object will give a maximum intensity that is increased by a factor of \( N^2 \). This is to be expected since we have \( N \) times the field added together and squaring to get the intensity gives a factor \( N^2 \).

The width of the transform will be reduced by the \( \sin^2(\alpha N) \) factor which has a zero at \( \alpha_0 N = \pi \) or

\[ \alpha_0 = h \cdot k_\perp/2 = \pi/N. \]

Thus the width of the central maximum will be reduced by a factor of \( N \). In addition there will be secondary maxima at zeroes of the denominator when \( h \cdot k_\perp = 2m\pi \) or

\[ k_{\perp,m} = \frac{2\pi}{h} \]

when \( k_\perp \) is parallel to \( h \).

These features of an intensity increase by \( N^2 \), a reduction in width of \( N \), and periodic maxima of the transform are true for generic objects \( A(\rho) \). They are particularly useful for achieving enhanced spectral sensitivity.

### 4.2.2 Periodic apertures

Let’s check an explicit example of a repeated 1D aperture. Let the aperture function \( A(x) \) be a periodic array of \( N \) slits separated by \( h \) and each of width \( \delta x \). The Fraunhofer diffraction pattern in the focal plane of a lens of focal length \( f \) is

\[ A(x_2) = \left( -\frac{ik}{2\pi f} \right)^{1/2} \int_{-\infty}^{\infty} dx A(x)e^{-ikx_2}. \]

The aperture function is

\[ A(x) = \sum_{j=1}^{N} \text{rect} \left( \frac{x - (j - (N+1)/2) h}{\delta x/2} \right) \]

\[ = \sum_{j=1}^{N} \text{rect} \left( \frac{x}{\delta x/2} - \frac{x_j}{\delta x/2} \right) \]

with \( x_j = [j - (N+1)/2]h \) and \( \text{rect} \) the unit rectangular function, \( \text{rect}(x) = 1 \) for \( |x| \leq 1 \). The Fourier transform then takes the form

\[ A(x_2) = \left( -\frac{ik}{2\pi f} \right)^{1/2} \sum_{j=1}^{N} e^{ikx_2x_j/\delta x/2} \int_{-\infty}^{\infty} dx \text{rect}[x/(\delta x/2)] e^{-ikx_2/\delta x/2}. \]
The integral no longer depends on $j$ and we get

$$A(x_2) = \left(\frac{-ik(\delta x)^2}{2\pi f}\right)^{1/2} \text{sinc} \left[\frac{k\delta x x_2}{2f}\right] \sum_{j=1}^{N} e^{\frac{ikx_2}{f}}.$$ 

The geometric sum is evaluated using $\sum_{j=1}^{N} e^{-\imath a j} = \frac{e^{-\imath a N} - 1}{1 - e^{-\imath a}}$ so

$$A(x_2) = \left(\frac{-ik(\delta x)^2}{2\pi f}\right)^{1/2} \text{sinc} \left[\frac{k\delta x x_2}{2f}\right] \frac{\sin \left(\frac{Nkhx_2}{2f}\right)}{\sin \left(\frac{khx_2}{2f}\right)}.$$

This formula will reappear when we discuss the spectral resolution of a diffraction grating. The sinc function is just the Fourier transform of one unit cell of the grating, while the factor of $\frac{\sin \left(\frac{Nkhx_2}{2f}\right)}{\sin \left(\frac{khx_2}{2f}\right)}$ gives $N$ times narrower lobes which are repeated every $\Delta x_2 = 2\pi f/(kh)$. This analysis implies that the spectroscopic resolution of a diffraction grating increases with $N$, the number of grating periods.

### 4.2.3 Fourier filtering

Spectral filtering of images can be performed using the $4-f$ optical processor shown in Fig. 4.14. An input image $A_1(\rho_1)$ is Fourier transformed by a lens with focal length $f$ which gives the Fourier transform $A_2(\rho_2) \sim \mathcal{F}[A_1(\rho_1)]_{k_{\perp 1}=k\rho_2/f}$ in the back focal plane of the first lens. This is then transformed again with the second lens. When $L_1 = L_2 = f$ and $L = 2f$ the field in the back focal plane of the second lens $A_3(\rho_3)$ is an inverted image of the input field. We can modify this image by inserting amplitude or phase filters $t(\rho_2)$ into the Fourier plane. Using the Fourier convolution theorem we find

$$A_3(\rho_3) = -\frac{ik}{f} \mathcal{F}[t(\rho_2)A_2(\rho_2)]_{k_{\perp 2}=k\rho_3/f} \quad (4.40a)$$

$$= -\frac{ik}{f} \frac{1}{2\pi} \mathcal{F}[t(\rho_2)]_{k_{\perp 2}=k\rho_3/f} * \mathcal{F}[A_2(\rho_2)]_{k_{\perp 2}=k\rho_3/f}. \quad (4.40b)$$

Figure 4.14: Fourier filtering setup. The filter function $t(\rho_2)$ is positioned in the Fourier plane between the lenses.
We see that the Fourier filtering operation can be expressed in two equivalent ways. As a filtering operation described by Eq. (4.40a) where we multiply the transform of the input field by the Fourier plane filter $t(\mathbf{\rho}_2)$ and then transform again to get the output. Alternatively we convolve the transform of the input field with the transform of the filter function. This is the content of Eq. (4.40b).

Although the convolution theorem provides a compact expression for the output field it is easy to run into trouble in the evaluation of the convolution. Let us evaluate the output field explicitly using

$$A_3(\mathbf{\rho}_3) = -\frac{ik}{f} \frac{1}{2\pi} \iint d\mathbf{\rho}_2 t(\mathbf{\rho}_2) A_2(\mathbf{\rho}_2) e^{-i\frac{k\rho_2 \cdot \rho_3}{f}}.$$ 

Then use the Fourier representations

$$t(\mathbf{\rho}_2) = \frac{1}{2\pi} \iint d\mathbf{\rho}' \tilde{t}(\mathbf{\rho}') e^{i\rho_2 \cdot \mathbf{\rho}'}$$

$$A_2(\mathbf{\rho}_2) = \frac{1}{2\pi} \iint d\mathbf{\rho}' A_2(\mathbf{\rho}'') e^{i\rho_2 \cdot \mathbf{\rho}''}$$

to get

$$A_3(\mathbf{\rho}_3) = -\frac{ik}{f} \frac{1}{2\pi} \iint d\mathbf{\rho}' \tilde{A}_2(\mathbf{\rho}'') \iint d\mathbf{\rho}' \tilde{t}(\mathbf{\rho}')$$

$$\times \frac{1}{4\pi^2} \iint d\mathbf{\rho}_2 e^{-i\frac{k\rho_2 \cdot \rho_3}{f}} e^{i\rho_2 \cdot \mathbf{\rho}'} e^{i\rho_2 \cdot \mathbf{\rho}''}$$

$$= -\frac{ik}{f} \frac{1}{2\pi} \iint d\mathbf{\rho}' \tilde{A}_2(\mathbf{\rho}'') \iint d\mathbf{\rho}' \tilde{t}(\mathbf{\rho}') \delta(\mathbf{\rho}' + \mathbf{\rho}'' - \frac{k}{f} \mathbf{\rho}_3)$$

$$= -\frac{ik}{f} \frac{1}{2\pi} \iint d\mathbf{\rho}' \tilde{t}(\mathbf{\rho}') \tilde{A}_2(\mathbf{\rho}'') \delta(\frac{k}{f} \mathbf{\rho}_3 - \mathbf{\rho}')$$

$$= -\frac{ik}{f} \frac{1}{2\pi} \iint d\mathbf{\rho}' \tilde{t}(\mathbf{\rho}') \tilde{A}_2(\mathbf{\rho}') \delta(\frac{k}{f} \mathbf{\rho}_3 - \mathbf{\rho}'),$$

which makes explicit (4.40b).

As a check let $t(\mathbf{\rho}_2) = 1$, then $\tilde{t}(\mathbf{\rho}') = 2\pi \delta(\mathbf{\rho}')$ and

$$A_3(\mathbf{\rho}_3) = -\frac{ik}{f} \iint d\mathbf{\rho}' \tilde{A}_2(\mathbf{\rho}') \delta(\frac{k}{f} \mathbf{\rho}_3 - \mathbf{\rho}')$$

$$= -\frac{ik}{f} \tilde{A}_2(\frac{k}{f} \mathbf{\rho}_3)$$

$$= -\frac{ik}{f} \frac{1}{2\pi} \iint d\mathbf{\rho}_2 A_2(\mathbf{\rho}_2) e^{-i\frac{k\rho_2 \cdot \rho_3}{f}}$$

$$= \left(-\frac{ik}{f}\right)^2 \iint d\mathbf{\rho}_1 A_1(\mathbf{\rho}_1) \frac{1}{4\pi^2} \iint d\mathbf{\rho}_2 e^{-i\frac{k\rho_2 \cdot \rho_3}{f}} e^{-i\frac{k\rho_1 \cdot \rho_2}{f}}$$

$$= \left(-\frac{ik}{f}\right)^2 \iint d\mathbf{\rho}_1 A_1(\mathbf{\rho}_1) \delta\left(\frac{k}{f} (\mathbf{\rho}_3 + \mathbf{\rho}_1)\right)$$

$$= -A_1(-\mathbf{\rho}_3).$$

(4.42)
We obtain an inverted image as expected with a phase factor of $-1 = (-i)^2$ which results from diffraction. With other filter functions we can implement low pass, high pass, or bandpass operations. We can also implement directional filtering, e.g. to suppress or enhance edges parallel to a desired transverse direction.

This type of imaging arrangement can be used to establish the diffraction limited resolution of an optical imaging system. Suppose the input field is a sinusoidal pattern with period $\Lambda$. We can think of this as being created by two plane waves propagating at angles $\pm\theta/2$ from the optical axis where $\Lambda = \lambda/(2 \sin \theta/2)$. These waves have transverse wavevector components $\pm k_\perp$ where $\sin \theta/2 = k_\perp/k$. Upon Fourier transformation the field $A_2$ will have strong components at $\rho_2 = f k_\perp/k$. Due to the finite diameter $D$ of the first lens the maximum value of $\rho_2$ that passes through the optical system is $\rho_{2,\text{max}} = D/2$. The maximum transverse wavenumber is thus $k_{\perp,\text{max}} = k_f \rho_{2,\text{max}} = kD/2f$. This corresponds to $\sin(\theta_{\text{max}}/2) = D/2f$ and

$$\Lambda_{\text{min}} = \frac{\lambda}{2 \sin(\theta_{\text{max}}/2)} = \frac{\lambda f}{D}.$$  

We see that the minimum spatial scale that propagates through the optical system is proportional to $\lambda$. Recalling that the numerical aperture of the lens is defined as $NA = \sin(\theta/2)$ we can write this simply as

$$\Lambda_{\text{min}} = \frac{\lambda}{2(NA)}$$

which shows that the spatial resolution is inversely proportional to the numerical aperture.

### 4.3 Diffraction gratings

See handwritten class notes on web page.
Chapter 5

Gaussian beams

5.1 Lowest order solution

The amplitude of a Gaussian laser beam in a plane where the wavefront is flat can be written as

\[ A = A_0 e^{-\frac{r^2}{w^2}}, \quad (5.1) \]

where \( w \) is the waist. The diameter of the beam between points where the intensity has fallen to \( 1/e^2 \) of the maximum is the “Gaussian” diameter and is given by \( d = 2w \). It is sometimes useful to work with the full width at half-maximum (FWHM) of the intensity. This is given by \( d_{\text{FWHM}} = \sqrt{2 \ln 2} w \approx 1.177w \).

The power in the beam (5.1) is

\[ P = \left( \frac{\epsilon_0^2 n c}{2} \right) 2\pi \int_0^\infty dr |A|^2 = \left( \frac{\epsilon_0^2 n c}{2} \right) \frac{\pi w^2}{2} |A_0|^2. \quad (5.2) \]

A useful expression for the spatial variation of the intensity is

\[ I(r) = \left( \frac{2P}{\pi w^2} \right) e^{-2r^2/w^2}. \]

As we saw in the previous section propagation from the front to the back focal plane of a lens results in Fourier transformation of the spatial structure of the field. In general a Fourier transformed field will have a different structure than the original field. For example a fringe pattern \( A_1 \sim \cos(qx_1) \) transforms to \( A_2 \sim \delta(x_2 - \lambda f q) + \delta(x_2 + \lambda f q) \) and a “top-hat” beam \( A_1(x_1) \sim \text{rect}(x_1 - a) \) transforms to a sinc function \( A_2 \sim \sin(x_2a/\lambda f)(x_2a/\lambda f) \).

A natural question arises as to whether or not there exist fields that retain the same spatial structure apart from an overall scale change under Fourier transformation, or under the more general transformation given by Eqs. (4.34,4.35). Indeed there are families of functions that are Fourier self-similar. For a comprehensive discussion of this topic see the book on Fourier transforms by Titchmarsh. Perhaps the simplest example of a Fourier self-similar field is the Gaussian, \( A_1 \sim e^{-x_1^2/w^2} \) which transforms into \( A_2 \sim e^{-x_2^2 w^2/\lambda^2 f^2} \). The Gaussian function together with higher order generalizations in terms of Hermite-Gauss functions play a very important role in optics as they are the natural modes in terms of which we can decompose localized beams. The lowest order field emitted by a well-aligned laser is also a Gaussian.

It is therefore natural to look for Gaussian type solutions of Eq. (4.2). Let’s find the lowest order solution, the fundamental Gaussian beam. To do so we guess a solution of the form

\[ A = A_0 e^{-i(P(z) - \frac{\lambda f q}{2\sigma(z)} r^2)}, \quad (5.3) \]
assume cylindrical symmetry so that $\nabla^2 = \partial_{rr} + (1/r)\partial_r$ and plug in to Eq. (4.2) to get
\[
\left(\frac{k}{q(z)}\right)^2 \left(1 - \frac{dq}{dz}\right) r^2 - i2k \left(\frac{1}{q(z)} - i\frac{dP}{dz}\right) = 0.
\] (5.4)

This equation must hold for all $r$ so we get
\[
\frac{dq}{dz} = 1 \quad \text{and} \quad \frac{dP}{dz} = -\frac{i}{q}.
\] (5.5)

The solution for $q$ is
\[
q(z) = q(0) + z
\] (5.6)
from which we find
\[
P(z) = -i \ln \left(1 + \frac{z}{q(0)}\right).
\] (5.7)

We have put the integration constant to zero since it merely gives an overall phase shift.

At this point it is convenient to introduce two real beam parameters: the radius of curvature $R$ and the waist radius $w$. Put
\[
\frac{1}{q(z)} = \frac{1}{R(z)} + i\frac{\lambda}{\pi nw^2(z)}.
\] (5.8)

Now define the plane $z = 0$ to be where $q$ is purely imaginary, this will lead to localized solutions with an energy that decays exponentially as $r \to \infty$. Then
\[
\frac{1}{q(0)} = i\frac{\lambda}{\pi nw_0^2}
\] (5.9)
where $w_0 = w(0)$. After some algebra we find that (5.3) can be written as
\[
A = A_0 \frac{w_0}{w(z)} e^{-\eta(z)} e^{\frac{k}{2r_0^2}} e^{-\frac{w^2(z)}{w_0^2}},
\] (5.10)
where
\begin{align*}
w^2(z) &= w_0^2 \left(1 + \frac{z^2}{L_R^2}\right), \quad \text{(5.11a)} \\
R(z) &= z \left(1 + \frac{L_R^2}{z^2}\right), \quad \text{(5.11b)} \\
\eta(z) &= \tan^{-1} \left(\frac{z}{L_R}\right), \quad \text{(5.11c)} \\
L_R &= \frac{\pi nw_0^2}{\lambda} = \frac{kw_0^2}{2}. \quad \text{(5.11d)}
\end{align*}

A fundamental beam profile is shown in Fig. 5.1. The parameters are the beam waist $w$, the radius of curvature $R$, the Gouy phase $\eta$ and the Rayleigh length $L_R$. The Rayleigh length is the characteristic distance over which a Gaussian type beam remains collimated.
When \( z = L_R \), \( w \) increases from \( w_0 \) to \( \sqrt{2}w_0 \). As an example for \( \lambda = 0.63 \ \mu \text{m (HeNe laser)} \) and \( w_0 = 1 \ \text{mm} \) we get \( L_R = 5.0 \ \text{m} \). Some authors also define a confocal length \( b = 2L_R \).

The Gouy phase adds a phase shift of \( \pi \) when a Gaussian beam goes through a focus compared to the phase of a plane wave\(^1\). The physical origin of the Gouy phase is related to the transverse structure of the field\(^2\). Higher order beams have larger Gouy phase shifts.

We motivated the Gaussian beam solution by seeking a field that is an eigensolution of paraxial propagation. If we start with a Gaussian beam and send it through any paraxial optical system characterized by an ABCD matrix then the beam coming out will also be Gaussian. To calculate the output beam we can use Eq. (4.34). Given a Gaussian beam characterized by \( q_1 = q(z_1) \) at plane \( z = z_1 \) it will be transformed into a new Gaussian beam with \( q_2 = q(z_2) \) at plane \( z = z_2 \). Evaluation of Eq. (4.34) shows that

\[
q_2 = \frac{Aq_1 + B}{Cq_1 + D} \tag{5.12}
\]

where \( ABCD \) are the elements of the ray matrix describing the propagation from planes 1 to 2. Note that the transformation law is particularly simple for propagation in a homogeneous medium of length \( L \): \( q_2 = q_1 + L \).

This perhaps curious looking result can be thought of as a generalization of the behavior of rays to beams that diffract. Suppose we let \( w \to \infty \) then Eq. (5.8) becomes \( q = R \), where \( R \) is the radius of curvature. A ray with parameters \((x, \theta)\) crossed the axis at \( z = -x/\theta \). Therefore the ray can be thought of as being part of a beam with radius of curvature \( R = x/\theta \). We can then ask how the radius of curvature changes under propagation. We have

\[
R = \frac{x}{\theta} \to R' = \frac{x'}{\theta'} = \frac{Ax + B\theta}{Cx + D\theta} = \frac{Ax/\theta + B}{Cx/\theta + D} = \frac{AR + B}{CR + D}.
\]

If we label the input and output radii of curvature as \( R_1 \) and \( R_2 \) we have

\[
R_2 = \frac{AR_1 + B}{CR_1 + D}.
\]

---


March 8, 2016 M. Saffman
which is the same transformation as Eq. (5.12) for the \( q \) parameter. Thus we have the remarkable result that the complex \( q \) parameter transforms in exactly the same way as the real radius of curvature of a ray.

As a first example using Eq. (5.12) let’s calculate the width of a Gaussian beam that has an initial waist \( w_0 \) in a medium with index \( n_1 \). Initially the \( q \) parameter is \( q_1 = -i\pi n_1 w_0^2 / \lambda = -iL_R \). The new \( q \) parameter is

\[
q_2 = \frac{Aq_1 + B}{Cq_1 + D} = \frac{-iL_R A + B}{-iL_R C + D} = \frac{B^2 + A^2 L_R^2}{BD + ACL_R^2 + i(AD - BC)L_R}
\]

so

\[
\text{Im} \left[ \frac{1}{q_2} \right] = \frac{(AD - BC)L_R}{B^2 + A^2 L_R^2} = \frac{\lambda}{\pi n_2 w_2^2}.
\]

The output width is therefore

\[
w_2^2(z) = \frac{n_1}{n_2} w_0^2 \frac{A^2 + B^2}{L_R^2} \frac{1}{AD - BC}.
\]

The determinant is \( AD - BC = n_1 / n_2 \) so that

\[
w_2(z) = w_0 \sqrt{A^2 + \frac{B^2}{L_R^2}},
\]

which reduces to Eq. (5.11a) for a straight segment of length \( z \). Note that \( w_2(z) \) is in general not a beam waist but is the width of the beam at plane \( z_2 \). We will usually denote the beam waist by an extra subscript of 0 as in \( w_{20} \). In the particular case where the waist occurs at plane \( z_2 \) the width is equal to the waist. To find the position of the new waist we require \( R_2 = \infty \) or \( \text{Re} \left[ \frac{1}{q_2} \right] = 0 \) which gives the condition

\[
BD + ACL_R^2 = 0.
\]

As we will discuss below this condition, which can be thought of as the requirement for imaging of a beam waist to a beam waist, does not in general coincide with the condition for imaging in a ray optics approximation which is simply \( B = 0 \). The question arises as to whether or not the geometrical imaging condition \( B = 0 \) can also coincide with imaging of the beam waist. If \( B = 0 \) the condition for imaging the waist is \( AC = 0 \). Since the ray matrices satisfy \( AD - BC = n_1 / n_2 \), if \( B = 0 \) we must have \( AD = n_1 / n_2 \) which implies \( A \neq 0 \). The requirement for simultaneous geometric imaging together with reimaging of the beam waist is therefore \( B = 0 \) and \( C = 0 \). For a single lens \( C = -1/f \) so this is not possible, but it may be satisfied in more complicated multi-lens situations as discussed in Sec. 5.1.7. Of particular interest are zoom lenses which provide geometrical and beam waist imaging between fixed planes while allowing the magnification to be varied.

### 5.1.1 Higher order Gaussian beams

The circularly symmetric solution (5.3) is only the lowest order mode among an infinite set of successively higher order beams. To find the higher order solutions we start with the separable ansatz
5.1 Lowest order solution

\[ A_{mn}(x, y) = A_0 \frac{w_0}{w(z)} e^{-m(z)} e^{\frac{k}{2m(z)} r^2} e^{-\frac{r^2}{w^2(z)}} f_m(x, z) f_n(y, z), \]

with \( f_m, f_n \) unknown functions to be determined. This form of solution is substituted into the paraxial wave equation (4.2), we separate variables, and note that the solution of

\[ \frac{d^2 u}{dx^2} - 2x \frac{du}{dx} + 2nu = 0 \]

is the \( n^{th} \) order Hermite polynomial \( u = H_n(x) \), to find

\[ A_{mn}(x, y) = A_0 \frac{w_0}{w(z)} H_m \left( \sqrt{2} \frac{x}{w(z)} \right) H_n \left( \sqrt{2} \frac{y}{w(z)} \right) e^{-\eta_{m,n}(z)} e^{\frac{k}{2m(z)} r^2} e^{-\frac{x^2+y^2}{w^2(z)}}, \]

where \( \eta_{m,n}(z) = (1 + m + n) \eta(z) \) and \( \eta, R, w \) are given by Eqs. (5.11). The normalization integral for Hermite polynomials is

\[ \int_{-\infty}^{\infty} d\xi \ e^{-\xi^2} H_m(\xi) H_n(\xi) = \sqrt{\pi} \ 2^m m! \ \delta_{mn}. \]

March 8, 2016 M. Saffman
Figure 5.3: Intensity of Laguerre-Gaussian beams in the focal plane $z = 0$ with $\lambda = 0.5$, $w_0 = 1$ in regions of size $6 \times 6$.

We can therefore write the normalized transverse mode which has integral of the field squared set to unity as

$$A_{mn}(x, y, z) = \left[ \frac{2}{\pi^{2m+n}m!n!} \right]^{1/2} \frac{1}{w(z)} H_m \left( \frac{x}{w(z)} \right) H_n \left( \frac{y}{w(z)} \right) e^{-i\eta_{m,n}(z)} e^{\frac{k}{2w(z)} r^2} e^{-\frac{r^2 + \rho^2}{w(z)}}.$$  

These are called Hermite-Gaussian beams. The intensity of the first few beams is shown in Fig. 5.2.

### 5.1.2 Laguerre-Gaussian beams

Beyond the Hermite-Gaussian beams there exist several families of Gaussian beam solutions of the paraxial wave equation associated with orthonormal polynomials. The Laguerre-Gaussian beams are of particular interest in connection with orbital angular momentum. They are given by

$$A_{pl}(\rho, \phi, z) = \left[ \frac{2p!}{\pi (p + |l|)!} \right]^{1/2} \frac{1}{w(z)} \left[ \frac{\sqrt{2}\rho}{w(z)} \right]^{|l|} L_p^{|l|} \left( \frac{2\rho^2}{w(z)} \right) e^{-i\eta_{2p,|l|}(z)} e^{\frac{k}{2w(z)} r^2} e^{-\frac{r^2 + \rho^2}{w(z)}} e^{-i\phi}. $$
Figure 5.4: Gaussian beam transformation by a lens of focal length $f$. The incident waist $w_{10}$ is positioned a distance $z_1$ in front of the lens. The output waist $w_{20}$ occurs a distance $z_2$ after the lens.

The intensity of a Laguerre-Gaussian beam is proportional to

$$|A_{pl}|^2 = \frac{2p!}{\pi(p + |l|)!} \frac{1}{w^2(z)} \left[ \frac{\sqrt{2\rho}}{w(z)} \right]^{2|l|} \left[ L_p^{|l|} \left( \frac{2\rho^2}{w^2(z)} \right) \right]^2 e^{-\frac{2\rho^2}{w^2(z)}}.$$

The intensity of the first few Laguerre-Gaussian beams is shown in Fig. 5.3. Beams $p, l$ and $p, -l$ have the same intensity distribution but the opposite sign of azimuthal phase.

When $p = 0$, the Laguerre polynomial simplifies to $L_0^{|l|} = 1$ and the field amplitude near the origin is $A_{0l} \sim \rho^0$. Any Laguerre-Gaussian mode can be written in terms of Hermite-Gaussian functions and vice versa.

5.1.3 Gaussian beam transformation by a lens

As a second example consider transformation of a Gaussian beam by a thin lens as shown in Fig. 5.4. For the geometry of Fig. 5.4 the transformation matrix is

$$M = \begin{pmatrix} 1 - \frac{z_2}{f} & z_1 + z_2 - \frac{z_1 z_2}{f} \\ -\frac{1}{f} & 1 - \frac{z_1}{f} \end{pmatrix}$$

and Eqs. (5.8,5.12) lead to

$$R_2 = \frac{(L_1^2 + z_1^2)z_2^2 - 2fz_2L_1^2 - 2fz_1(z_1 + z_2) + f^2L_1^2 + f^2(z_1 + z_2)^2}{(L_1^2 + z_1^2)z_2 + (z_1 + z_2)f^2 - fL_1^2 - fz_1(z_1 + z_2)},$$

$$w_2^2 = \frac{\lambda}{\pi L_1 f} \left[ (L_1^2 + z_1^2)z_2^2 - 2fz_2(L_1^2 + z_1^2 + z_1 z_2) + f^2(L_1^2 + (z_1 + z_2)^2) \right].$$

Here we have introduced the Rayleigh length of the input beam $L_1 = \pi w_{10}^2/\lambda$. The output waist occurs at a value of $z_2$ such that $R_2(z_2) = \infty$. To find $z_2$ we note that the input waist has $q$ parameter $q_1 = -iL_1$ and at the output waist $q_2$ is pure imaginary. This gives the

---

condition \( BD + ACL_1^2 = 0 \) where \( A, B, C, D \) are the elements of the ray matrix connecting planes 1 and 2. Solving for this condition gives

\[
z_2 = f \frac{z_1(z_1 - f) + L_1^2}{(z_1 - f)^2 + L_1^2}
\]

and

\[
w_{20}^2 = w_{10}^2 \frac{f^2}{(z_1 - f)^2 + L_1^2}.
\]

The simplest situation occurs when \( z_1 = f \), i.e. the incident waist is in the front focal plane of the lens. In this case we find the output waist is at \( z_2 = f \) and \( w_{20}^2 = \frac{\lambda f^2}{\pi L_1} \).

When we have a positive lens with \( f > 0 \) this can be written as

\[
w_{10}w_{20} = \frac{\lambda f}{\pi}.
\]

This very useful result is exact for paraxial Gaussian beams. We will refer to this condition as a confocal arrangement since the beam waists lie in the front and back focal planes of the lens. Introducing the “confocal” waist \( w_c = \sqrt{\frac{\lambda f}{\pi}} \) which is invariant under the lens transformation and the dimensionless waist parameters \( \omega_1 = w_{10}/w_c, \omega_2 = w_{20}/w_c \) Eq. (5.19) takes on the simplified form

\[
\omega_1 \omega_2 = 1.
\]

Eq. (5.20)

When \( z_1 \neq f \) it is useful to describe the lens transformation in terms of dimensionless lengths \( \zeta_1 = z_1/f, \zeta_2 = z_2/f \). In terms of these new variables we find

\[
\zeta_2 = \frac{\zeta_1(\zeta_1 - 1) + \omega_1^4}{(\zeta_1 - 1)^2 + \omega_1^4}
\]

and

\[
\omega_1^2 \omega_2^2 = \frac{1}{1 + \frac{(\zeta_1 - 1)^2}{\omega_1^2}}.
\]

Equation (5.22) reduces to (5.20) when \( z_1 = f \) so \( \zeta_1 = 1 \). Figure 5.5 shows representative curves of \( \zeta_2 \) and \( \omega_2 \) for different input beam waists.

Some useful limits can be extracted from Eqs. (5.21,5.22). The maximum and minimum values of \( \zeta_2 \) occur when \( \zeta_1 = 1 \pm \omega_1^2 \) which give \( \zeta_2 = 1 \pm \frac{1}{2\omega_1^2} \). Thus for small \( \omega_1 \) the distance to the waist approaches \( \frac{1}{2\omega_1^2} \) or \( f \times \frac{w_{20}^2}{w_{10}^2} \). The corresponding output waist values satisfy \( \omega_1^2 \omega_2^2 = 1/2 \). This can also be expressed as \( \omega_2^2 = \frac{1}{2(\zeta_1 - 1)} \). The maximum possible value of \( \omega_2 \) occurs when \( \zeta_1 = 1 \) which gives \( \omega_1 \omega_2,\text{max} = 1 \). We see that for a given input beam waist the largest possible output waist occurs when the input waist is placed in the front focal plane.

Note that when \( \zeta_1 > 1 \) we always get \( \zeta_2 > 1 \). Stated in words: when the input waist is more than a focal length away from the lens, the output waist is also more than a focal length away. The slope of the \( \zeta_2/\zeta_1 \) mapping increases inversely with \( \omega_1 \). Thus when the input waist is small compared to the confocal waist \( w_c \), i.e. \( \omega_1 \ll 1 \) the slope is large and a small shift of the input waist position causes a large shift of the output waist position. Conversely when \( w_{10} \gg w_c \), i.e. \( \omega_1 \gg 1 \) the output waist position is pinned at \( \zeta_2 = 1 \), or
$z_2 = f$. Qualitatively this regime occurs when the input waist is large enough that $L_1 \gg f$ so the output waist appears at $z_2 = f$ after the lens.

When $\zeta_1 < 1$ we get $\zeta_2 < 1$. In other words when the input waist is closer than a focal length from the lens the output waist is also closer than a focal length, provided $\omega_1$ is not too small. When $\omega_1 \ll 1$ $\zeta_2$ can be negative which means the waist is in front of the lens, and no real waist occurs after the lens.

The somewhat counterintuitive behavior of the waist transformation can be contrasted with the requirement for imaging in a geometrical optics approximation $B = 0$, which gives the lens makers formula for a single lens $1/f = 1/z_1 + 1/z_2$. In geometrical imaging bringing the object closer to the lens pushes the image further away. The explanation of the lack of “waist imaging” when there is geometrical imaging is due to phase errors. The geometrical imaging condition implies that all rays leaving a single point in the object plane also meet at a single point in the image plane. However this condition says nothing about the path lengths of the rays or about the relative path lengths for different pairs of object and image plane points. Thus the geometrical imaging condition does not preclude quadratic phase errors which lead to non imaging of a Gaussian beam waist.

In many applications we wish to focus a Gaussian beam to a waist that is small compared to $w_c$. Thus $\omega_1 \gg 1$ and $\omega_2 \ll 1$. In this situation the relative motion of the output waist as the input waist position changes is most usefully quantified by normalizing the waist positions $z_1, z_2$ to the input and output Rayleigh lengths $L_1, L_2$. We thus introduce normalized deviations from the focal planes as

$$\zeta_1' = \frac{z_1 - f}{L_1}$$
$$\zeta_2' = \frac{z_2 - f}{L_2}.$$  

When $\zeta'_1 \ll 1$ it is not hard to show that $\zeta'_2 \simeq \zeta'_1$. In other words the relative position error normalized to the Rayleigh length is a constant under transformation between front and back focal planes.
Figure 5.6: Numerical aperture for a given focusing ratio at aperture parameter $p = 4$. The dashed line shows a large $r$ approximation.

### 5.1.4 Gaussian beam focusing

Another useful formula is the numerical aperture needed to focus a Gaussian beam to a desired spot size. Assume a waist of $w_0$ which expands to $w(z) = w_0 \sqrt{1 + z^2/L_R^2}$. An aperture with a diameter that is $p$ times larger than $w(z)$ fills a cone of half angle $\theta$. The numerical aperture of the aperture is

$$NA = \sin \theta = \frac{pw_0}{2} \sqrt{1 + z^2/L_R^2} \over \sqrt{z^2 + \frac{p^2w_0^2}{4} (1 + z^2/L_R^2)}$$

$$= \frac{pw_0}{2} \left( \frac{1}{2} + \frac{1}{L_R} \right)^{1/2} \sqrt{1 + \frac{p^2w_0^2}{4} \left( \frac{1}{2} + \frac{1}{L_R} \right)}.$$

At large distances $z \gg L_R$ this tends to

$$NA \approx \frac{pw_0}{2L_R} \frac{1}{\sqrt{1 + \frac{p^2w_0^2}{4L_R^2}}}.$$

The transmission of a Gaussian beam through an aperture that is $p$ times larger in diameter than the beam waist is

$$T = \frac{2}{\pi w^2} \int_0^{pw_0/2} dr 2\pi r e^{-r^2/w^2}$$

$$= 1 - e^{-p^2/2}.$$

An aperture diameter of four times the waist is a good rule of thumb since $p = 4$ gives $T = 0.9997$. Using $p = 4$ we find

$$NA \approx \frac{2w_0}{L_R} \frac{1}{\sqrt{1 + \frac{4w_0^2}{L_R^2}}} = \frac{1}{\sqrt{1 + \frac{\pi^2w_0^2}{4\lambda^2}}} \quad (5.23)$$
5.1 Lowest order solution

Figure 5.6 shows the required numerical aperture from Eq. (5.23) as a function of the ratio \( r = \frac{w_0}{\lambda} \). In the small aperture limit of \( r \gg 1 \) we find

\[ NA \simeq \frac{2}{\pi r} = \frac{2\lambda}{\pi w_0} \]

which is shown as a dashed line in the figure.

5.1.5 Fiber collimation

A specific case of the waist transformations occurs in the common optical task of collimating a beam from a single mode fiber. The fiber output will be taken to be a beam of waist \( w_1 \). The beam propagates a distance \( z_1 \), and is then focused by a lens of focal length \( f \) which forms a new waist \( z_2 \) after the lens, which is exactly the geometry of Fig. 5.4.

Rewriting Eqs. (5.21,5.22) the new waist occurs at

\[ z_2 = f \frac{\left( \frac{z_1}{f} \right) \left( \frac{z_1}{f} - 1 \right) + \left( \frac{w_1}{w_c} \right)^4}{\left( \frac{z_1}{f} - 1 \right)^2 + \left( \frac{w_1}{w_c} \right)^4} \]

and has the value

\[ w_2^2 = w_1^2 \frac{w_c^4}{w_1^4 + w_c^4 \left( \frac{z_1}{f} - 1 \right)^2} \]

Here \( w_c = \sqrt{\frac{\lambda f}{\pi}} \) is the confocal waist. Let us take \( \lambda = .78 \) \( \mu \)m, \( w_1 = 3.05 \) \( \mu \)m, and \( f = 6.24 \) mm which give \( w_c = 39.4 \) \( \mu \)m. The output waist, and waist position for values of \( z_1 \) close to \( f \) are shown in Fig. 5.7.

A commonly made mistake is to place the lens further than \( f \) away from the fiber end face so that the waist is smaller than the maximum. This occurs if we try to collimate the fiber output by minimizing the spot size on a distant screen a distance \( z_s \) from the lens. We see from the figure that for the parameters used \( z_2 < 0.5 \) m. If \( z_s \) is larger than this it is not possible to adjust the lens so the waist is on the screen. In this situation the beam radius on the screen is given by

\[ w_s(z_1) = w_2(z_1) \sqrt{1 + \frac{(z_s - z_2(z_1))^2}{L_2^2}} \]

---

March 8, 2016 M. Saffman
Figure 5.8: Width on a screen at distance $z_s$ vs. lens position $z_1$. The curves are labeled with $z_s = 1, 2, 5, 10$ m.

Figure 5.9: Beam width as $z_1$ is varied (left) and as $z_2$ is varied (right). The flat line in the left hand graph is an exact result for all $z_1$. The flat line in the right hand graph is an approximate result for $z_2$ similar to $f$.

where $L_2 = \pi w_2(z_1)^2/\lambda$.

Figure 5.8 shows the shift of the apparent best collimation position $z_1$ as a function of $z_s$. Comparison with Fig. 5.7 shows that to produce the maximum waist with an error less than 5% it is necessary to have the screen at least 5 m away. Such a large distance is inconvenient in most laboratory settings.

It might appear that a possible procedure is as follows. We minimize the waist on a screen that is say 1 m away. Then monitor the beam width a distance $f$ after the lens with a beam scanner and move the fiber a small distance towards the lens to maximize the measured width. However, the beam width a distance $z_2 = f$ after the lens is independent of $z_1$! It can easily be shown that the width at $z_2 = f$ is simply $w_2 = \lambda f/\pi w_1$ which is a constant. Thus moving the fiber will not change the measured beam width.

The behavior of the beam width for different values of $z_1, z_2$ is shown in Fig. 5.9. The correct position $z_1$ can be set by measuring the beam width as $z_2$ is changed, and adjusting $z_1$ so that the beam width stays almost constant, despite changing $z_2$. Setting $z_1 = f$ we
5.1 Lowest order solution

find

\[ w_2 = w_1 \left( 1 + \frac{z_2^2}{f^2} - 2 \frac{z_2}{f} + \frac{f^2}{L_1^2} \right)^{1/2}. \]

For \( z_2 \approx f \) we have \( w_2 \approx w_1 f / L_1 = \lambda f / \pi w_1 \). Large values of \( z_2 \) result in the usual quadratic behavior \( w_2 \approx w_1 z_2 / f \), but for \( z_2 \) similar to \( f \) the beam width will be closest to constant when \( z_1 = f \) as shown in the right hand side of Fig. 5.9.

5.1.6 Refocusing with a mirror

Another example arises in the context of optical frequency conversion in a nonlinear crystal. As shown in Fig. 5.10 we wish to refocus a beam waist in a second pass through a nonlinear crystal. The ABCD matrix from the beam waist plane to the mirror and back is

\[
\begin{bmatrix}
1 & L \\
0 & 1
\end{bmatrix}
\begin{bmatrix}
1 & 0 \\
0 & 1/n
\end{bmatrix}
\begin{bmatrix}
1 & L_a \\
0 & 1
\end{bmatrix}
\begin{bmatrix}
1 & 0 \\
0 & 1
\end{bmatrix}
\begin{bmatrix}
1 & L_a \\
0 & 1
\end{bmatrix}
\begin{bmatrix}
1 & 0 \\
0 & n
\end{bmatrix}
\begin{bmatrix}
1 & L \\
0 & 1
\end{bmatrix}
\]

\[
= \begin{bmatrix}
1 & \frac{2}{R}(L_a + L/n) \\
0 & \frac{2}{nR}(L_a + L/n)(L_a - R + L/n)
\end{bmatrix}
\]

\[
1 - \frac{2}{R}(L_a + L/n). \tag{5.24}
\]

Setting \( BD + ACL_1^2 = 0 \) gives the condition \( R = 2(L/n + L_a) \) for reimaging the waist. This result could also have been deduced from the behavior of a thin lens without multiplying all the ABCD matrices together by recalling that the effective diffraction length inside a medium of index \( n \) is \( L/n \) so the effective distance to the mirror is \( L/n + L_a \) and the focal length is \( f = R/2 \). Thus the waist is in the front focal plane when \( R = 2f = 2(L/n + L_a) \).

5.1.7 Zoom lenses for Gaussian beams

The question arises as to whether or not the geometrical imaging condition \( B = 0 \) can also coincide with reimaging of the beam waist which requires \( BD + ACL_1^2 = 0 \) as discussed above. We see that if \( B = 0 \) the condition for a new waist is \( AC = 0 \). Since the ray matrices are unimodular \( AD - BC = 1 \), if \( B = 0 \) we must have \( AD = 1 \) which implies \( A \neq 0 \). The requirement for simultaneous geometric imaging together with reimaging of the beam waist is therefore \( B = 0 \) and \( C = 0 \). For a single lens \( C = -1/f \) so this is not possible, but it may be satisfied in more complicated multi-lens situations as discussed in this section.

![Figure 5.10: Refocusing of light in a crystal with a curved mirror](image-url)
Of particular interest are zoom lenses which provide geometrical and beam waist imaging between fixed planes while allowing the magnification to be varied.

### 5.1.8 Two-lens zoom

A two-lens configuration does not function as a true zoom lens, but it can provide variable magnification with relatively small axial displacement of the beam waist\(^4\). This is called a “varifocal” zoom. The analysis proceeds as follows. The ray matrix for the two-lens zoom shown in Fig. 5.11 is

\[
M = \frac{1}{f_1 f_2} \times \left( f_1 (f_2 - z_3) + z_2 z_3 - f_2 (z_2 + z_3) \right) \frac{f_1 ((z_1 + z_2 + z_3) f_2 - (z_1 + z_2) z_3) - z_1 ((z_2 + z_3) f_2 - z_2 z_3)}{(z_2 - f_1 - f_2)}
\]

The geometrical imaging condition is \(B = 0\) or

\[
z_2 = f_1 + \frac{f_1^2}{z_1 - f_1} + \frac{f_2 z_3}{z_3 - f_2}.
\]

For \(B = 0\) the waist imaging condition is \(C = 0\) or

\[
z_2 = f_1 + f_2.
\]

Simultaneous solution of (5.25,5.26) gives

\[
z_3 = \frac{f_2}{f_1} (f_1 + f_2) - \frac{f_2^2}{f_1^2} z_1.
\]

The geometrical magnification as well as the ratio of output to input waists is then given by

\[
A = \frac{f_1 (f_2 - z_3)}{f_2 (f_1 - z_1)} = -\frac{f_2}{f_1}.
\]

Although the two-lens configuration can simultaneously provide geometrical imaging and waist imaging it cannot do so while providing zoom functionality since the magnification is fixed at $-f_2/f_1$.

If we keep $B = 0$ but drop the requirement that the waist appears at the geometrical image plane, i.e. $z_2 \neq f_1 + f_2$ then the geometrical image appears at

$$z_{3g} = \frac{f_2 z_1 z_2 - f_1 f_2 (z_1 + z_2)}{f_1 (f_2 - z_1 - z_2) + z_1 (z_2 - f_2)} \quad (5.27)$$

and the magnification is

$$A = \frac{f_1 f_2}{f_1 (f_2 - z_1 - z_2) + z_1 (z_2 - f_2)}.$$

A parfocal zoom lens provides variable magnification without axial image displacement, i.e. $z_1 + z_2 + z_{3g} = Z = \text{constant}$. Inserting into (5.27) we can write the object to image distance as

$$Z = z_1 + z_2 + \frac{f_2 z_1 z_2 - f_1 f_2 (z_1 + z_2)}{f_1 (f_2 - z_1 - z_2) + z_1 (z_2 - f_2)} \quad (5.28)$$

It is not hard to show that there are no choices of $f_1, f_2$ that allow the magnification to be varied while keeping $Z$ fixed. We conclude that a parfocal or true zoom lens is not possible with only two lenses.

However variable magnification accompanied by some motion of the image plane, or varifocal zoom behavior, is possible. The relation between magnification and image shift is found by rewriting (5.28) as

$$Z = z_1 + z_2 + A \left( \frac{z_1 z_2}{f_1} - z_1 - z_2 \right) \quad (5.29)$$

If we put $z_2 = z_1 f_1 / (z_1 - f_1)$ the term in parentheses vanishes and $Z = \frac{z_1^2}{z_1 - f_1}$ while $A = \frac{f_1}{f_1 - z_1}$ so $Z = -Az_1^2 / f_1$. This is not a useful solution since the smallest image shift is at $z_1 \approx 0$ which means $z_2 \approx 0$ which is not physical, as the lenses would be on top of each other. Let us instead choose $z_1 = f_1$ which gives $\blacklozenge$ (to be added, but this is not very interesting).

5.1.9 Three-lens zoom

To provide geometrical and waist imaging with variable magnification we turn to a three lens configuration as shown in Fig. 5.12. This can provide variable magnification between fixed input and output planes and is referred to as a “parfocal” zoom. The distance from input plane to zoom output plane is $Z = z_1 + z_2 + z_3 + z_4 = z_{10} + z_{20} + z_{30} + z_{40}$. If we start with nominal positions $z_{10} = f_1$, $z_{20} = f_1 + f_2$, $z_{30} = f_2 + f_3$, $z_{40} = f_3$ then we have geometrical and Gaussian waist imaging to the reimaged output plane with static magnification $M_0 = f_2 f_3 / (f_1 f_3)$. The lens values can be chosen as desired, also either positive or negative, to give a desired static magnification. Negative lenses (concave) lead to a shorter overall zoom lens and may be preferred for comapactness.

March 8, 2016 M. Saffman
Figure 5.12: Layout of a three lens zoom. The three lens zoom is afocal so we include a fixed \( f_4 \) to transform the output. The nominal distances are \( z_{10}, z_{20}, z_{30}, z_{40} \). The lenses move by amounts \( \delta_1, \delta_2, \delta_3 \) during zoom operation giving actual distances \( z_1 = z_{10} + \delta_1, z_2 = z_{20} + \delta_2 - \delta_1, z_3 = z_{30} + \delta_3 - \delta_2, \) and \( z_4 = z_{40} - \delta_3. \)

By moving the lenses, i.e. changing \( \delta_1, \delta_2, \delta_3 \) away from zero we get a new magnification \( M = rM_0 \). ABCD analysis shows that this can be done while keeping \( B = C = 0 \) provided

\[
\begin{align*}
\delta_1 &= \sqrt{r - 1} \frac{f_1^2((f_1^2 - rf_2^2)(f_3^2 - rf_2^2))^{1/2}}{r^2f_2^3 - rf_1^2f_2} \\
\delta_2 &= \sqrt{r - 1} \frac{rf_2((f_1^2 - rf_2^2)(f_3^2 - rf_2^2))^{1/2}}{(r^2f_2^4 - f_1^2f_2^2)} \\
\delta_3 &= \sqrt{r - 1} \frac{rf_2((f_1^2 - rf_2^2)(f_3^2 - rf_2^2))^{1/2}}{(r^2f_2^4 - f_1^2f_3^2)}.
\end{align*}
\]

These solutions are written in a form valid for \( r > 1 \) and \( (f_1^2 - rf_2^2)(f_3^2 - rf_2^2) > 1 \). If we wish to vary the zoom magnification both smaller and larger than a nominal value we can simply start with \( M_0 \) at the small end and use these solutions for \( r > 1 \). An example is shown in Fig. 5.13. The initial magnification with the lenses confocally spaced is \( M_0 = -0.9 \) so varying \( r \) between 1 and 1.25 changes the magnification to \( M = -1.125 \). The total length of the zoom is \( Z = 15 \).
5.2 Apertured Gaussian Beams

The power transmission of a Gaussian beam with waist $w$ through a circular aperture of diameter $D$ is

$$T = \frac{2}{\pi w^2} \int_0^{D/2} dr \frac{2\pi r e^{-2r^2/w^2}}{w^2} = 1 - e^{-D^2/2w^2}. \quad (5.30)$$

Choosing $D = 2w, 3w, 4w$ gives $T = 0.8647, 0.98889, 0.99966$.

Aperturing the beam is useful for removing the intensity tails and for smoothing the spatial profile. This is typically done in a Fourier filtering configuration shown in Fig. 5.14. Say we have a Gaussian beam with waist $w_0$ but the beam has some additional spatial structure we wish to suppress. We Fourier transform to a new waist $w_1$, pass through an aperture of diameter $D$ and transform back to the original waist $w_0$. If the aperture were not present the output beam would be the same as the input beam. With an aperture spatial structure gets blocked in the Fourier plane. If the aperture is too big unwanted spatial frequencies giving beam structure will still be transmitted. If the aperture is too small, we lose power, and we add “ringing” to the output beam due to the discontinuity in intensity at the aperture edges.

It is not obvious what the optimal aperture diameter should be. We can estimate the ringing effect as follows. The discontinuity in the beam amplitude at the aperture edge is $\delta A = e^{-D^2/4w_1^2}$ for a beam with unit peak amplitude. If we think about the fringe due to this discontinuity on the background of a unit amplitude Gaussian we get a peak intensity of $\delta I = (1 + \delta A)^2 - 1 \approx 2\delta A$. Taking $D = 4w_1$ to get more than 0.999 transmission gives
Figure 5.15: Intensity profiles in back focal plane for $w_1 = w_2 = 0.126 \text{ mm}$, $\lambda = 0.5 \mu\text{m}$, $f = 100. \text{mm}$ and $D/w_1 = 1, 2, 3, 4, 5$ (left). The profile for $D/w_1 = 5$ is indistinguishable from that with $D = \infty$ shown in purple. The right hand plot shows the weak oscillating tails for $D/w_1 = 4, 4.25, 4.5$.

$\delta I = 0.037$, or 3.7%. This is a rough estimate, but it highlights the problem that very small intensity changes can lead to relatively large interference.

We can calculate more precisely with a Fourier model. Let’s first consider the ideal case of a centered circular aperture with diameter $D$ and an on axis Gaussian beam of waist $w_1$, $A_1(\rho_1) = A_{10}e^{-\rho_1^2/w_1^2}$. The field in the back focal plane has radial symmetry and can be expressed as

$$A_2(\rho_2) = -i\frac{k}{f}A_{10} \int_0^{D/2} d\rho_1 \rho_1 e^{-\rho_1^2/w_1^2} J_0\left(\frac{k}{f}\rho_1\rho_2\right). \quad (5.31)$$

Put $t = \frac{2}{D}\rho_1$, $v = \frac{kD}{2\rho_1}\rho_2$, $u = -iD^2/2w_1^2$, to get

$$A_2(\rho_2) = \left(\frac{kw_1^2}{2f}A_{10}\right) u \int_0^1 dt t e^{-i\frac{u}{2}t^2} J_0(ut).$$

The integral can be expressed in terms of the Lommel functions (4.31)

$$U_1(u, v) = u \int_0^1 dt t J_0(ut) \cos\left[\frac{u}{2}(1 - t^2)\right] \quad (5.32a)$$

$$U_2(u, v) = u \int_0^1 dt t J_0(ut) \sin\left[\frac{u}{2}(1 - t^2)\right] \quad (5.32b)$$

whence

$$A_2(\rho_2) = \left(\frac{kw_1^2}{2f}A_{10}\right) e^{-iu/2}(U_1 + iU_2). \quad (5.33)$$

Let’s check the amplitude at the origin $\rho_2 = 0$. We find

$$A_2(0) = \left(\frac{kw_1^2}{2f}A_{10}\right) e^{-iu/2}(U_1(u, 0) + iU_2(u, 0)).$$

From the function definitions we see that

$$U_1(u, 0) = \sin(u/2), \quad U_2(u, 0) = 2\sin^2(u/4)$$
Figure 5.16: Airy ring structure of the intensity profiles in the back focal plane for $D/w_1 = 0.5, 1, 2$, and all other parameters the same as in Fig. 5.15. The red curves are for a Gaussian beam with no aperture, the blue curves are the apertured Gaussian and the green curves are from the Airy formula (4.22).

so

$$A_2(0) = -i \left( \frac{k w_1^2}{2 f} A_{10} \right) \left[ 1 - e^{-D^2/4w_1^2} \right]. \quad (5.34)$$

When there is no aperture the output field is a Gaussian with waist $w_2 = 2f/kw_1$ and amplitude

$$A_2(\rho_2) = -i A_{10} \frac{w_1}{w_2} e^{-\rho_2^2/w_2^2} = -i A_{10} \frac{k w_1^2}{2 f} e^{-\rho_2^2/4f^2},$$

which agrees with the first term of (5.34). The peak intensity falls off faster than the power transmission given by Eq. (5.30) which implies that the beam effectively is broadened.

The intensity profiles for several values of aperture diameter are shown in Fig. 5.15. As the aperture diameter is decreased the transmitted power is reduced and the beam becomes broader but there is no obvious modulation imparted to the main beam. There is weak modulation in the far wings for $D/w_1 \sim 4$, at $< 10^{-6}$ in relative intensity. For $D \sim w_1$ the output is essentially the transform of a uniform circular disk and Airy rings can be seen, as shown in Fig. 5.16. When $D < w_1$ the apertured intensity is close to uniform and the intensity profile matches closely that of the Airy expression (4.22). As $D/w_1$ increases the intensity inside the aperture becomes more and more nonuniform and the antinodes of the diffraction are in approximately the same positions as predicted by the Airy formula, but are substantially suppressed.

We conclude from the above plots that aperturing at $D \sim 4w_1$ does not impart significant extra modulation to the central lobe of the beam. This calculation leaves open the question of why the estimate based on the edge discontinuity overestimates the beam modulation. This may be due to the fact that a Gaussian beam picks up a diffractive phase of $\pi/2$ when propagating from Fourier to image planes. By Babinet’s principle we can think of the interference as arising from superposition of the transformed whole Gaussian minus the transform of the excised wings of the Gaussian. If the wings do not acquire the same $\pi/2$ phase they interfere with the main beam in quadrature which suppresses the modulation amplitude.
Figure 5.17: Intensity profiles for $D/w_1 = 4$, and all other parameters the same as in Fig. 5.15 for $z = (0, 0.1, 0.5, 1, 2)L_R$ with $L_R = \pi w_2^2/\lambda = 200$ mm.

### 5.2.1 Apertured beam in intermediate plane

The question arises as to the structure of the apertured beam if we look in a plane a distance $z$ away from the Fourier plane. We can calculate this using Fresnel diffraction as

$$A_3(\rho_3) = -i \frac{k}{z} e^{\frac{ik}{z} \rho_3^2} \int_0^\infty d\rho_2 A_2(\rho_2) \rho_2 e^{\frac{ik}{z} \rho_2^2} J_0(k \rho_3 \rho_2 / z)$$

$$= -A_{10} \frac{k}{f} e^{\frac{ik}{f} \rho_3^2} \int_0^\infty d\rho_2 \rho_2 e^{\frac{ik}{f} \rho_2^2} J_0(k \rho_3 \rho_2) \int_0^{D/2} d\rho_1 \rho_1 e^{-\rho_1^2/w_1^2} J_0(k \rho_1 \rho_2)$$

$$= -A_{10} \frac{k}{f} e^{\frac{ik}{f} \rho_3^2} \int_0^{D/2} d\rho_1 \rho_1 e^{-\rho_1^2/w_1^2} \int_0^\infty d\rho_2 \rho_2 e^{i \frac{ik}{f} \rho_2^2} J_0(k \rho_3 \rho_2) J_0(k \rho_1 \rho_2).$$

Gradshteyn and Ryzhik 6.633 give

$$\int_0^\infty dx x J_0(ax) J_0(bx) e^{-c^2 x^2} = \frac{1}{2c^2} e^{-\frac{a^2}{4c^2} + \frac{b^2}{4c^2}} I_0\left(\frac{ab}{2c^2}\right) \quad a > 0, \quad b > 0, \quad |\arg c| < \pi/4.$$

We have $a = k \rho_3 / z, b = k \rho_1 / f$ which are both positive but $c^2 = -ik/2z$ so $\arg c = -\pi/4$. We can add a small convergence factor $\epsilon > 0$ to the integral so $c^2 = \epsilon - ik/2z$ and use

$$\int_0^\infty d\rho_2 \rho_2 e^{(i \frac{k}{f} - \epsilon) \rho_3^2} J_0(k \rho_3 \rho_2) J_0(k \rho_1 \rho_2) = \frac{1}{2\epsilon - ik / z} e^{-\frac{k^2 (\frac{\epsilon}{2\epsilon - ik / z})^2}{4\epsilon - 2ik / z}} I_0\left(\frac{k^2}{2\epsilon - ik / z} \rho_1 \rho_3\right).$$

Letting $\epsilon \to 0$ we get a well behaved result which we use to express the output field as

$$A_3(\rho_3) = -i A_{10} \frac{k}{f} \int_0^{D/2} d\rho_1 \rho_1 e^{-\rho_1^2/w_1^2} e^{-i \frac{4k^2}{2\epsilon - 2ik / z}} J_0(k \rho_1 \rho_3).$$

This last expression can again be written in terms of Lommel functions as

$$A_3(\rho_3) = -i \frac{A_{10}}{z/f - i2f/kw_1^2} e^{-uw/2} [U_1(u, v) + iU_2(u, v)]$$

(5.35)
with

\[ u = \frac{kzD^2}{4f^2} - \frac{i D^2}{2w_1^2}, \quad v = \frac{kD}{2f} \rho_3. \]

We can check that letting \( z \to 0 \) and \( \rho_3 \to \rho_2 \) we recover (5.33) as expected. Intensity profiles for several values of \( z \) are shown in Fig. 5.17. We see the expected smooth diffractive spreading of the beam with only weak intensity modulation in the wings.
Finally we can ask what the beam looks like in the near field a distance $z$ from the aperture plane. Using Fresnel diffraction formulae we have

$$A_3(\rho_3) = -\frac{k}{z} e^{\frac{i}{2\pi} \rho_3^2} \int_0^{D/2} d\rho_1 e^{-\rho_1^2/w_1^2} \rho_1 e^{\frac{i}{2\pi} \rho_1^2} J_0(k\rho_1/\rho_3).$$

This is the same as (5.31) with some changes of variables so the solution is

$$A_3(\rho_3) = i \frac{A_{10}}{1 + i2z/kw_1^2} e^{\frac{i}{2\pi} \rho_3^2} e^{-iu/2}(U_1 + iU_2) \quad (5.36)$$

with

$$u = -i \frac{D^2}{2} \left( \frac{1}{w_1^2} - i \frac{k}{2z} \right), \quad v = \frac{kD}{2z} \rho_3.$$
Chapter 6
Optical resonators

An important application of ray matrices arises in the analysis of resonant optical cavities. In Sec. 2.2 the Fabry-Perot interferometer was analyzed assuming excitation with infinitely wide plane waves. As pointed out in Eq. (2.19) wide input beams are needed to achieve high finesse in devices with planar mirrors. This limit can be removed by using spherical mirror cavities that support a spectrum of stable eigenmodes. We can analyze the stability and resonance properties of these cavities using the formalism of ray matrices. Implicit in this approach is the assumption that each optical element has a transverse size that is large compared to the width of the resonating eigenmode. When this assumption does not hold a more sophisticated analysis is required that accounts for scattering at the edges of physical elements (see Gordon, Li, Siegman).

6.1 Resonator eigenmodes

An eigenmode satisfies \( q_1 = q_2 = q \) which gives

\[
\frac{1}{q} = \frac{D - A}{2B} \pm \frac{1}{2B} \sqrt{(A - D)^2 + 4BC}
\]

where \( A, B, C, D \) are the elements of the ray matrix describing one resonator round trip. Since the ray matrices are unimodular we can write this as

\[
\frac{1}{q} = \frac{D - A}{2B} \pm \frac{i}{2B} \sqrt{4 - (A + D)^2}.
\]

The \( q \) parameter is related to the waist and radius of curvature of the cavity mode by

\[
\frac{1}{q} = \frac{1}{R} + \frac{\lambda}{\pi nw^2}.
\]

The condition for a confined mode is that the square of the waist radius is positive, which requires \(|A + D| \leq 2\). The mode parameters are then

\[
R = \frac{2B}{D - A}, \quad (6.1a)
\]

\[
w^2 = \pm \frac{2\lambda}{\pi n} \frac{B}{[4 - (A + D)^2]^{1/2}}. \quad (6.1b)
\]
where the ± sign should be used for $B$ positive/negative. A waist occurs when $R = \infty$ which requires $A = D$ so that
\[
 w_0^2 = \pm \frac{\lambda B}{\pi n (1 - A^2)^{1/2}}.
\] (6.2)

Note that the elements of the round trip ray matrix, and therefore the solutions for $R$ and $w$ depend on the choice of reference plane. Different choices will lead to different solutions. Alternatively we can choose a desired reference plane and then use the ray matrix for propagation to find the mode parameters at a different resonator plane. It is often convenient to choose a reference plane with some obvious symmetry where e.g. a waist occurs.

### 6.2 Linear cavity

An important example is provided by the two-mirror linear cavity shown in Fig. 6.1. Two mirrors with radii of curvature $R_1$ and $R_2$ are separated by a distance $L$. We start with a reference plane at the center of the cavity. The beam at this plane has an initial $q$ parameter given by $q(z = 0) = q_1$. After one round trip through the cavity we have
\[
 q_2 = \frac{A q_1 + B}{C q_1 + D}
\]
where $A, B, C, D$ are the elements of the round trip ray matrix which is given by
\[
 M = M_l(L/2)M_r(R_2)M_l(L)M_r(R_1)M_l(L/2).
\] (6.3)

Here $M_l, M_r$ are the ray matrices for linear propagation and mirror reflections from Fig. 3.4. We find
\[
 A = 1 + \frac{2L^2 - L(R_1 + 3R_2)}{R_1 R_2} \\
 B = 2L + \frac{2L^3 - 3L^2(R_1 + R_2)}{2R_1 R_2} \\
 C = \frac{4L - R_1 - R_2}{R_1 R_2} \\
 D = 1 + \frac{2L^2 - L(3R_1 + R_2)}{R_1 R_2}
\]
Figure 6.2: Stability diagram for a linear resonator. The shaded regions are stable, and everywhere else is unstable. The points marked with dots are symmetric plane-plane ($g_1 = g_2 = 1, R_1 = R_2 = \infty$), confocal ($g_1 = g_2 = 0, L = R_1 = R_2$), and concentric ($g_1 = g_2 = -1, L = 2R_1 = 2R_2$). The figure also shows asymmetric hemispherical and concave-convex resonators.

We see that when $R_1 = R_2$ then $A = D$ and the waist is at the center of the cavity. The sign convention for $R_1, R_2$ is the same as for the focal length $f = R/2$ of the equivalent lens in an unfolded version of the resonator. Thus, for the cavity in Fig. 6.1, $R_{1,2}$ are both positive.

The stability condition $|A + D| < 2$ takes the form

$$-1 < \frac{L(L - R_1 - R_2)}{R_1 R_2} < 0.$$ 

It is customary to define parameters $g_1 = 1 - L/R_1$ and $g_2 = 1 - L/R_2$. The stability condition can then be written as

$$0 < g_1 g_2 < 1.$$ 

The stability diagram in the $g_1, g_2$ plane is shown in Fig. 6.2.

To calculate the waist size it is convenient to first determine the position of the waist, which could be inside or outside the cavity. The matrix of Eq. (6.3) cannot be used when $R_1 \neq R_2$ since the plane where the waist occurs is not in the center of the cavity. To find the waist position assume a starting reference plane that is offset by a distance $\delta L$ from the cavity center. Calculating the round trip matrix and setting $A = D$ gives

$$\delta L = \frac{L(R_2 - R_1)}{2(2L - R_1 - R_2)}.$$
We have defined the position as $\delta L > 0$ when the waist is to the left of the cavity center, i.e. closer to mirror 1. We can check that when $R_1 = \infty$ then $\delta L = L/2$ and the waist is at mirror 1. Although the radii of curvature of the self-consistent modes at the mirrors matches the mirror radii of curvature in the linear cavity this is not generally true. For example in the ring cavity of Sec. 6.6 the radius of curvature of the stable mode does not match the cavity mirror radii.

Calculating with a reference plane offset by $\delta L$ we find the round trip matrix and then using Eq. (6.2) we find

$$w_0^2 = \frac{\lambda L}{\pi} \sqrt{\frac{g_1 g_2 (1 - g_1 g_2)}{g_1 + g_2 - 2 g_1 g_2}}.$$ 

For a symmetric resonator $R_1 = R_2 = R$ so $g_1 = g_2 = g$ we get

$$w_0 = \left( \frac{\lambda L}{2\pi} \right)^{1/2} \left[ \frac{1 + g}{1 - g} \right]^{1/4} = \left( \frac{\lambda}{2\pi} \right)^{1/2} \left[ L(2R - L) \right]^{1/4}.$$ 

Let’s evaluate this for the symmetric resonators at the stability boundaries. For the confocal resonator $L = R$ and $g = 0$ so $w_0 = (\lambda L / 2\pi)^{1/2}$. Note that this result could have been easily derived using the expression for transformation of a Gaussian beam by a lens $w_1 w_2 = \lambda f / \pi$ with $w - 1 = w_2 = w_0$ and $f = R/2 = L/2$. The Rayleigh length of the beam is $L_R = L/2$ so the beam width on the mirrors is $\sqrt{2} w_0$. The confocal resonator has the special property that the mode is intrinsically stable against small misalignments, even though it is marginally stable (it sits on the stability boundary in Fig. 6.2). The mode stability derives from the fact that $dw_0/dL = 0$ so small changes in the resonator length have only a small effect on the mode size.

For the concentric resonator $L = 2R$ and $g = -1$ so $w_0 = 0$. The ABCD analysis breaks down in this case. The actual mode size is determined by the size of the end mirrors which are filled by the cavity mode.

For the planar resonator $g = 1$ so $w_0 = \infty$. Again the ABCD analysis breaks down. The actual mode size is determined by diffraction from the edges of the cavity mirrors and requires a more sophisticated analysis.

### 6.3 Linear cavity with internal optical element

Another important example is provided by the two-mirror linear cavity with an internal optical element shown in Fig. 6.1. Two mirrors with radii of curvature $R_1$ and $R_2$ are separated by a distance $L$ and there is an element of length $L_c$ with index $n_c$. This type of cavity is commonly used for second harmonic generation and optical parametric oscillation and as a laser, where the element represents the gain medium.

We could repeat the analysis of Sec. 6.2 to find the stability condition but this is not necessary. When there is no internal crystal the matrix for propagation between the mirrors is $M_t(L) = \begin{pmatrix} 1 & L \\ 0 & 1 \end{pmatrix}$. When the crystal is present this becomes $M_t(L_{\text{eff}})$ where the effective diffractive length is

$$L_{\text{eff}} = L - L_c + \frac{L_c}{n}.$$
6.4 Resonant frequencies

The resonant frequencies of plane wave fields inside a planar cavity were found in Sec. 2.2 from the condition $2Lkn = 2\pi q$ where $q$ is the longitudinal mode index. Resonant modes in spherical cavities have a transverse field profile which can correspond to any of the higher order modes detailed in Sec. 5.1.1. The resonant frequency is different for different transverse modes due to the Gouy phase factor $\eta_{mn}(z) = (1+m+n)\eta(z)$. Here $m, n$ label the transverse modes.
modes with $m = n = 0$ the lowest order Gaussian beam mode. Any of the transverse modes can be excited with proper mode matching in a stable cavity. Because the mode frequency depends on $q, m,$ and $n$ an optical resonator can be used for combined spectral and spatial filtering of an optical field.

The resonant frequencies are found by starting at the reference plane where the waist occurs and enforcing the condition

$$2Lkn(z) + (1 + m + n) \int_C \eta(z + L) - \eta(z - L) = 2\pi q.$$ 

It can be shown, (details to be added) after considerable algebra, that the resonant frequencies are

$$\nu_{qmn} = \nu_{\text{FSR}} \left[ q + \frac{1 + m + n}{\pi} \cos^{-1}(\pm \sqrt{g_1 g_2}) \right].$$

Here $\nu_{\text{FSR}}$ is the cavity free spectral range calculated for a plane wave and the sign of $\sqrt{g_1 g_2}$ is given by the sign of $g_1$ ($g_1$ and $g_2$ have the same sign for stable modes). The part proportional to $(m + n) \cos^{-1}(\pm \sqrt{g_1 g_2})$ arises from the higher order Gouy phase.

Let’s calculate the mode frequencies for the three basic linear resonators, planar, confocal, and concentric. For the planar resonator $g_1 = g_2 = 1$ and

$$\nu_{qmn} = q\nu_{\text{FSR}}.$$ 

All transverse modes have the same frequency since there are no focusing elements, and no Gouy phase shift.

For the confocal resonator $g_1 = g_2 = 0$ and

$$\nu_{qmn} = \left[ \left( q + \frac{1}{2} \right) + \left( \frac{m + n}{2} \right) \right] \nu_{\text{FSR}}.$$ 

When $m + n$ is an even integer transverse modes associated with longitudinal index $q$ are degenerate with longitudinal mode $q' = q + (m + n)/2$. In other words all modes with $m + n$ even, including $m + n = 0$ will be simultaneously resonant. When $m + n$ is an odd integer the transverse modes are spaced halfway in between axial modes. By tuning the resonator length we can select simultaneous resonance for all even or all odd transverse modes.

For the concentric resonator $g_1 = g_2 = -1$ and

$$\nu_{qmn} = (q + 1 + m + n) \nu_{\text{FSR}}.$$ 

Different transverse modes are all simultaneously degenerate with axial modes at multiples of the free spectral range.

We see that the planar and concentric geometries allow simultaneous resonance of all transverse modes. Nevertheless these geometries are not preferable in practice since they are very sensitive to small cavity misalignments. The confocal geometry is relatively insensitive to alignment errors while allowing simultaneous resonance of all even or odd transverse modes.
Figure 6.5: Mode matching to a linear resonator. Without the input coupler (top) the focusing lens produces a waist $w_0$ at position $z_0$. With the input coupler in place (bottom) the waist is transformed to $w_2$ at position $z_2$.

### 6.5 Mode matching

An important practical aspect when working with optical cavities is the necessity of mode matching an input beam to the cavity mode. We can define an overlap function between an incident beam $A_1$ and a cavity mode amplitude $A_2$ by the integral

$$O = \int dx dy \ A_1^* A_2.$$  \hspace{1cm} (6.4)

The profiles $A_1$ and $A_2$ are functions of $z$, however it is straightforward to show that provided $A_1$ and $A_2$ are solutions of the paraxial equation (4.2) the overlap $O$ is independent of $z$. We can therefore simplify the analysis by assuming the cavity mode has a waist at $z = 0$ and calculating the overlap in this plane. Parameterizing the ratio of the waists of the amplitudes by $s = w_{01}/w_{02}$ and the axial displacement of the incident beam waist by $\zeta = z_1/z_{R2}$ we find that the power overlap is given by

$$|O|^2 = \frac{4s^2}{(1 + s^2)^2 + \zeta^2}.$$  \hspace{1cm} (6.5)

A plot of the overlap is shown in Fig. 6.4. In order to exceed 90% power overlap the beam waist ratio must not deviate by more than $-20\% - +30\%$ and the waist position must not be displaced by more than about $1/2$ of a Rayleigh length. These requirements also apply to the problem of maximizing coupling efficiency into a single mode optical fiber. It should be apparent that mode matching to higher order transverse modes requires preparing the input beam to also have a higher order mode structure.

In order to achieve good mode matching the input beam parameters must be carefully controlled. Consider the geometry shown in Fig. 6.5. A lens with focal length $f$ is positioned
to create a beam waist $w_0$ at $z = z_0$. When the beam is transmitted through the front mirror (thickness $t$ and index $n_g$) of the linear cavity (the input coupler) it acquires a radius of curvature at $z = 0$ given by

$$R_1 = R \frac{L_R^2 + z'^2}{\tilde{n}(L_R^2 + z'^2) - Rz'}$$

where $L_R = \pi w_0^2 / \lambda$, $\tilde{n} = n_g - 1$, and $z' = z_0 + \frac{n_g - 1}{n_g} t$. The input coupler is concave to the right so $R > 0$. As long as $R_1 < 0$ a new waist will be formed to the right of the mirror. If $R_1 > 0$ the new waist is to the left of the mirror, i.e. the beam is diverging in the resonator and will certainly be poorly mode matched. The transition occurs at a cavity mirror radius of

$$R_m = \tilde{n} \frac{L_R^2 + z'^2}{z'}.$$

For $R < R_m$ the new waist will be to the left of the front mirror. Assuming $R > R_m$ the waist is formed at

$$z_2 = R \frac{(R - \tilde{n}z')z' - \tilde{n}L_R^2}{(R - \tilde{n}z')^2 + \tilde{n}^2 L_R^2}.$$  \hspace{1cm} (6.6)

The value of the new waist is

$$w_2^2 = w_0^2 \frac{R^2}{\tilde{n}^2 L_R^2 + (R - \tilde{n}z')^2}.$$  \hspace{1cm} (6.7)

Equations (6.6,6.7) determine $z_2, w_2$ as functions of $z_0, w_0$. For practical use they must be inverted to find the corresponding values of $z_0, w_0$. As they are nonlinear equations it is convenient to solve them graphically.

Let’s look at a representative case of modematching to a confocal cavity with $R = 0.1$ m mirrors and a wavelength of $\lambda = 1$ $\mu$m. The confocal cavity waist is $w_2 = \sqrt{\lambda R/2 \pi} = 126$ $\mu$m at $z_2 = R/2 = 5$ cm. To modematch to this waist we should generate a beam with

$$w_0 = w_2 \sqrt{\tilde{n}^2 L_R^2 + (z'')^2}. \hspace{1cm}$$

Solving (6.6) for $z_0$ we find

$$z_0 = \frac{R^2 + 2\tilde{n}R(z_2 - \frac{\hat{w}}{n_g} t) - 2z_2 \frac{\tilde{n}^3}{n_g} t \pm \sqrt{R^4 - 4L_R^2 \tilde{n}^2 (R + \tilde{n}z_2)^2}}{2 \tilde{n}(R + \tilde{n}z_2)}.$$

### 6.6 Ring cavity

Ring cavities are used frequently for optically pumped lasers such as the Ti:Sa and for nonlinear interactions. They have several distinguishing features relative to linear cavities that are advantageous. First they support unidirectional oscillation which prevents spatial hole burning, and second the round trip losses for parametric interactions are reduced relative to linear cavities without requiring that particular phase conditions be satisfied by the cavity mirrors.
A prototypical ring cavity in a bow-tie geometry containing a crystal used at normal incidence is shown in Fig. 6.6. The geometrical round trip cavity length is $L$ while the curved mirrors with radius of curvature $R$ are separated by a distance $L_2$. The optical crystal has length $L_c$ and index of refraction $n_c$. The distance from the crystal end surface to a curved mirror is then $(L_2 - L_c)/2$ and the distance between the curved mirrors for a ray traversing the long side of the cavity is $L - L_2$. With these definitions we find that the round trip ray matrix, starting from the center of the crystal as the reference plane, has elements

$$A = \frac{2(L - L_2)(L_2 - L'_c) - 2R(L - L'_c) + R^2}{R^2}$$  \hspace{1cm} (6.8)$$

$$B = \frac{[L_2 - R - L'_c] [(L - L_2)(L_2 - L'_c) - R(L - L'_c)]}{R^2}$$  \hspace{1cm} (6.9)$$

$$C = \frac{4(L - R - L_2)}{n_c R^2}$$  \hspace{1cm} (6.10)$$

$$D = A$$  \hspace{1cm} (6.11)$$

where $L'_c = L_c \frac{n_c - 1}{n_o}$.

Since $A = D$ the waist is located at the reference plane in the center of the crystal as expected. To see how the cavity mode waist depends on parameters we have plotted $w$ found from Eq. (6.1) in Fig. 6.7 as a function of the mirror separation $L_2$ for several values of the
cavity length. The condition for a confined mode ($|A + D| < 2$) can be used to find limits on the possible values of $L_2$. A little algebra shows that a stable mode exists for

$$\frac{L + L'}{2} - \frac{1}{2} \sqrt{(L - L')^2 - 4R(L - L')} + 4R^2 < L_2 < \frac{L + L'}{2} - \frac{1}{2} \sqrt{(L - L')^2 - 4R(L - L')}.$$ 

This analysis is incomplete due to astigmatism which results in different stability conditions in the tangential and sagittal planes.

### 6.6.1 Ring cavity with angled crystal

Ring cavities are often used for optically pumped lasers such as dye, or Ti:Sapphire lasers. To avoid reflection losses at the crystal surface it is common to orient the crystal at Brewster’s angle as shown in Fig. 6.8. In this case the astigmatism of the angled crystal can be used to compensate the mirror astigmatism, as was first shown by Kogelnik, et al.\(^1\)

Accurate modeling of the cavity requires the optical propagation lengths between the curved mirrors to be accounted for carefully. We assume the cavity beams propagate in a medium with index $n_1$ and the crystal has index $n_c$. Brewster’s angle is $\theta_B = \tan^{-1}(n_r)$ where $n_r = n_c/n_1$. The refracted angle inside the crystal is $\theta'_B = \tan^{-1}(1/n_r) = \cos^{-1}(n_r/\sqrt{1 + n_r^2}).$

Figure 6.9: Calculated waists at the center of the Ti:Sa crystal (left) and between the flat mirrors (right). Parameters: $\lambda = 0.685$ $\mu$m, $L_c = 1$ cm, $n_c = 1.77$, $R = 2.5$ cm, $\theta = 19$ deg., $L_T = 15$ cm, $h = 2.3$ cm.

The refracted beam has a geometrical path length inside the crystal of $L_c$. Formulae for the $x$ and $y$ displacements of the beam in the crystal are given in Fig. 6.8. The total round trip length is $L_T = L_1 + L_2 + L_{d1} + L_{d2}$. The distance $L_1$ is the optical path length between the curved mirrors (the actual $x$ distance between the mirrors is slightly smaller). One way to analyze the mode structure is to fix $L_T$ and look at what happens as $L_1$ is varied. We assume $L_c, n_c, \theta$ are fixed constants. The remaining cavity distances are then found from

$$L_2 = (L_{d1} + L_{d2}) \cos(2\theta) - L_1 + (L_c - t_x)$$

$$L_{d1} = h/\sin(2\theta)$$

$$L_{d2} = (h + t_y)/\sin(2\theta)$$

and

$$L_T - L_1 = L_{d1} + L_{d2} + L_2 = \frac{1 + \cos(2\theta)}{\sin(2\theta)} (t_y + 2h) - L_1 + L_c - t_x.$$ 

The last equation can be solved for $h = h(L_T, L_1)$ which then determines the other lengths $L_2, L_{d1}, L_{d2}$.

Using the above definitions the round trip ray matrix, starting from the center of the crystal as the reference plane, is given by

$$M_t = M_t(L_c/2)M_{i,t}(n_1, n_c, \theta_B)M_t((L_1 - L_c)/2)M_{i,t}(R, \theta)M_t(L_T - L_1)$$

$$\times M_{r,t}(R, \theta)M_t((L_1 - L_c)/2)M_{i,t}(n_c, n_1, \theta_B')M_t(L_c/2)$$

where the matrices are defined in Fig. 3.4. Subscript $t$ denotes tangential matrices and an analogous expression holds for matrices in the sagittal plane.

A sample calculation for a compact Ti:Sa ring laser is shown in Fig. 6.9. Although the astigmatism is well compensated, the tangential and sagittal waists inside the cavity differ by a factor of about 2 which complicates mode matching.

The free spectral range in frequency units is

$$\nu_{FSR} = \frac{c}{(L_T - L_c)n_1 + L_c n_c}.$$ 

March 8, 2016 M. Saffman
For the parameters of Fig. 6.9 we find $\nu_{FSR} = 1.90$ GHz.

For the purposes of checking the mechanical layout it is useful to find explicit values for the coordinates of the cavity beam vertices. These are given in Table 6.1.

<table>
<thead>
<tr>
<th>Position</th>
<th>x</th>
<th>y</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$t_x/2$</td>
<td>$t_y/2$</td>
</tr>
<tr>
<td>2</td>
<td>$x_1 + (L_1 - L_c)/2$</td>
<td>$y_1$</td>
</tr>
<tr>
<td>3</td>
<td>$x_2 - L_{d1} \cos(2\theta)$</td>
<td>$y_1 + h$</td>
</tr>
<tr>
<td>4</td>
<td>$x_5 + L_{d2} \cos(2\theta)$</td>
<td>$y_1 + h$</td>
</tr>
<tr>
<td>5</td>
<td>$-x_2$</td>
<td>$-y_1$</td>
</tr>
<tr>
<td>6</td>
<td>$-x_1$</td>
<td>$-y_1$</td>
</tr>
</tbody>
</table>

Table 6.1: Coordinates of beam vertices 1 – 6 in the Ti:Sa ring cavity.

Figure 6.10: Layout of the Ti:Sa cavity for the parameters of Fig. 6.9.
Chapter 7
Electro-optics

In this chapter we cover a potpourri of topics in the areas of modulators, detectors, and stabilization schemes.

7.1 Optical Modulators

Several different devices can be used to imprint amplitude or phase modulation on an optical beam. These include mirrors mounted on piezoelectric transducers, liquid crystal devices, acousto-optic modulators, electro-optic modulators, and spatial light modulators.

7.1.1 Acousto-optic deflector

Consider a device with acoustic wave velocity $v$ at an applied frequency $f_a$ as shown in Fig. 7.1. Light of wavelength $\lambda$ is diffracted by the wave through an angle $\theta_d = 2\theta_B$. At the Bragg angle $\theta_B$ the light reflected from multiple planes interferes constructively giving a strong diffracted intensity. Referring to the figure the condition for this to occur is $2hkn = m2\pi$ where $n$ is the refractive index inside the Bragg cell and the integer $m$ specifies the diffraction

![Figure 7.1: Bragg diffraction in an acousto-optic deflector with acoustic wavelength $\Lambda$.](image)

March 8, 2016 M. Saffman
order. Substituting $h = \Lambda \sin(\theta_B)$ we get the condition

$$\sin(\theta_B) = m \frac{\pi}{nkd} = m \frac{\lambda/n}{2d} = m \frac{\lambda/n f_a}{2v}.$$ 

The deflection angle is

$$\theta_d = 2\theta_B = 2 \sin^{-1} \left( m \frac{\lambda/n f_a}{2v} \right) \simeq m(\lambda/n)f_a.$$ 

The deflector can be used as a beam scanner giving parallel displacements by placing the device in the front focal plane of a lens.

The deflected light is frequency shifted by an amount which can be calculated from energy conservation. An incident photon has energy $E = h\omega$ and momentum $K = v_a/\Omega_a$. Energy and momentum conservation dictate that

$$\omega_d = \omega + m\Omega_a, \quad k_d = k + mK.$$ 

When the incident angle deviates from the Bragg angle the diffracted intensity falls off. The dependence off the diffracted power on the Bragg angle mismatch can be found from a coupled wave analysis...

### 7.1.2 Time bandwidth product

If the device has length $L$ along the direction of propagation of the acoustic wave the time to fill the acoustic medium with a new frequency is $T = L/v$. The time bandwidth product is defined as $TBW = T\Delta f_a$ where $\Delta f_a$ is the allowable range of acoustic wave frequencies. This range is set by some combination of material parameters, the piezo transducer, and associated electronic circuit elements.

Suppose the incident optical beam has a Gaussian intensity profile with waist $w_0 = L/q_1$ with $q_1$ some constant. The output is Fourier transformed by a lens of focal length $f$ to give a waist in the back focal plane of $w = (\lambda f/\pi w_0)$. The change in position of the center of the output beam as the frequency is changed by $\Delta f_a$ is $\Delta x = f\Delta \theta_B = (f\lambda/v)\Delta f_a$. The number of output spots separated by $q_2 w$ is therefore

$$N = \frac{\Delta x}{q_2 w} = \frac{(f\lambda/v)\Delta f_a}{q_2(\lambda f/\pi w_0)} = \frac{\pi w_0 \Delta f_a}{q_2 v} = \frac{\pi}{q_1 q_2} (T\Delta f_a) = \frac{\pi}{q_1 q_2} TBW.$$ 

Let us assume $q_1 = 4$. A smaller $q_1$ would give more resolvable spots but would lead to clipping of the intensity at the aperture of the Bragg cell since $w_0 = L/q_1$. To get good separation of the output beams we will take $q_2 = 2$ which corresponds to an intensity crosstalk between output beam centroids of $e^{-2(q_2 w)^2/w^2} = e^{-2q_2^2} = 3.3 \times 10^{-4}$. With these choices the number of resolvable spots is $N = \frac{\pi}{8} TBW$. 


7.1.3 Electro-optic modulator

Another widely used device is the electro-optic modulator (EOM). A laser beam propagates through a crystal which has a nonzero Pockels coefficient. Applying an external voltage to the modulator changes the refractive index and if this is done periodically in time frequency sidebands are imprinted on the laser beam.

7.2 Pound Drever Hall locking

7.3 Tilt-Locking

Let’s say the optical field in one dimension consists of two modes as

\[ A(x,t) = A_{00}(x,t) + A_{01}(x,t) \]

where \( A_{00}(x,t) = e^{-x^2}e^{-i\omega t}e^{i\phi_0} \) and \( A_{01}(x) = xe^{-x^2}e^{-i\omega t} \). A large detector will give the signal

\[
V \sim \int dx |A|^2 \\
= \int dx \left[ |A_{00}|^2 + |A_{01}|^2 + A_{00}^* A_{01} + A_{00} A_{01}^* \right] \\
= V_{00} + V_{01} + V_{\text{beat}}.
\]

Since the 00 and 01 modes are orthogonal, as can be easily checked, \( V_{\text{beat}} = 0 \), so \( V = V_{00} + V_{01} \).

Suppose now we replace the large detector with a split detector that produces a signal \( V_+ \) that is the integrated intensity for \( x > 0 \) and \( V_- \) that is the integrated intensity for \( x < 0 \). Then

\[
V_+ = \frac{V_{00}}{2} + \frac{V_{01}}{2} + \int_0^\infty dx (A_{00}^* A_{01} + A_{00} A_{01}^*) \\
V_- = \frac{V_{00}}{2} + \frac{V_{01}}{2} + \int_{-\infty}^0 dx (A_{00}^* A_{01} + A_{00} A_{01}^*).
\]
The difference signal obtained by subtracting the photodiode outputs is thus

\[ V_d = V_+ - V_- \]

\[ = \int_0^\infty dx \left[ A_{00} A_{01}^* + A_{00}^* A_{01} \right] - \int_0^\infty dx \left[ A_{00} A_{01}^* + A_{00}^* A_{01} \right] \]

\[ = \int_0^\infty dx \left[ A_{00}(x) A_{01}(x) + A_{00}^*(x) A_{01}(x) - A_{00}(-x) A_{01}^*(-x) - A_{00}^*(-x) A_{01}(-x) \right]. \]

Then use \( A_{00}(x) = A_{00}(-x) \) and \( A_{01}(x) = -A_{01}(-x) \) to get

\[ V_d = 2 \int_0^\infty dx \left[ A_{00}(x) A_{01}^*(x) + A_{00}^*(x) A_{01}(x) \right] \]

\[ = 2 \int_0^\infty dx \left[ A_{00}(x) A_{01}^*(x) + A_{00}^*(x) A_{01}(x) \right] \]

\[ = 4 \cos(\phi_0) \int_0^\infty dx \; x \; e^{-2x^2}. \]

Note that for \( \phi_0 = \pi/2 \) the difference signal is zero (this is the case sketched in Fig. 1 of the paper by Shaddock, et al.), whereas for other relative phase shifts the difference will be positive or negative. The value of \( \phi_0 \) will, for fixed cavity length, depend on the tilt angle. Increasing the tilt angle will increase the magnitude of \( \phi_0 \). If we now tune the cavity then \( \phi_0 \) will change to \( \phi = \phi_0 + \theta_0 \) where \( \theta_0 \) is the phase shift of the reflected 00 mode. Thus the difference signal will be

\[ V_d = 4 \cos(\phi_0 + \theta_0) \int_0^\infty dx \; x \; e^{-2x^2}. \]

If we choose the working point at \( \phi_0 = \pi/2 \) and the cavity has a high finesse so that \( \theta_0 \simeq 2 \tan^{-1}(\delta) \) where \( \delta = (\omega - \omega_c)/\gamma \) is the normalized cavity detuning. Then

\[ V_d \sim \sin \left[ 2 \tan^{-1}(\delta) \right] = 2 \frac{\delta}{1 + \delta^2}. \]

Figure 7.2 shows \( V_d \) which has the desired dispersive shape. Quantitative comparisons with experiment could be obtained by using the full formulas for the Fabry-Perot phase shift as a function of cavity tuning.

### 7.3.1 Littrow laser scanning

A convenient approach to building a tunable laser diode is to use feedback from a diffraction grating in the so-called Littrow configuration where the first order diffracted order propagates back towards the laser as shown in Fig. 7.3. Tilting the grating changes the angle \( \theta \) as well as the cavity length \( L_c \). When the geometry is chosen correctly the optical frequency determined by the diffraction grating angle stays in resonance with the changing cavity length so that the laser can be tuned continuously without mode hops.

The conditions for synchronous scanning are found as follows. The grating equation for the Littrow configuration of diffraction into the -1 order is \( \sin(\theta) = \lambda_d/(2\Lambda) \) where \( \lambda_d \) is the
optical wavelength, $\Lambda$ is the diffraction grating period, and $\theta$ is the angle the beam makes with the grating normal. The roundtrip cavity phase for a beam of wavelength $\lambda$ that starts at the laser, propagates to the grating, and returns to the laser, is

$$\phi = \phi_0 + 2\frac{2\pi}{\lambda}L_c - \frac{2\pi}{\Lambda}x$$

where $L_c = l + h\tan(\theta) - s/\cos(\theta)$, $x = h/\cos(\theta) - s\tan(\theta)$ is the coordinate in the grating plane where the beam meets the grating, and $\phi_0$ is a fixed phase offset. We assume that at wavelength $\lambda_d$ the horizontal offset $h$ has been chosen such that the total round trip phase is $\phi = 2\pi n$, with $n$ an arbitrary integer. The cavity is then on resonance.

When the grating is rotated the wavelength $\lambda_d = 2\Lambda\sin(\theta)$ changes. In order to keep the cavity on resonance we require that $d\phi/d\theta = 0$. This condition can be written as

$$\frac{d\phi}{d\theta} = \frac{4\pi}{\lambda} \frac{dL_c}{d\theta} - \frac{4\pi}{\lambda^2} \frac{d\lambda}{d\theta} - \frac{2\pi}{\Lambda} \frac{dx}{d\theta}$$

where $L_c = h\tan(\theta) - l\tan(\theta)/\cos(\theta)$, $x = (h/\cos(\theta) - l)\tan(\theta)$. The cavity phase variation is zero for $L_c = h\tan(\theta)$ or $l = s/\cos(\theta)$.

Lack of precision in constructing the laser will nonetheless result in a lack of synchronism during a wavelength scan. To evaluate how much error in the offset distance $l$ can be tolerated we need to know the width of the resonances due to the diffraction grating and cavity length. Typical diffraction efficiencies for gratings used in the Littrow configuration are 10-50 % resulting in a relatively low finesse Fabry Perot longitudinal mode structure. The width of the diffraction grating reflectivity as a function of laser frequency can be estimated as $\Delta\theta = \theta/p$ where the number of lines of the diffraction grating that are illuminated by a
beam of width $d$ is given by $p = d/(\Lambda \cos(\theta))$. Thus $\Delta \theta = \theta \cos(\theta) \Lambda/d$ and the number of longitudinal cavity modes inside the diffraction peak is roughly

$$m = \frac{\Delta \theta}{\theta} \left| \frac{\delta \nu_c}{\delta \theta} \right| \frac{1}{\nu_{\text{FSR}}} = \frac{\theta N}{p} \quad \quad (7.2)$$

The longitudinal mode index is roughly $N = 2L_c/\lambda$ which gives

$$m = \frac{\theta 2L_c \Lambda \cos(\theta)}{\lambda d}.$$  

Using $\lambda = .78 \mu m$, $\Lambda = 1/1200$ mm, and $d = 4$ mm we get $\theta = 28^\circ$ and $m \simeq 6.9$. Thus there are always a handful of longitudinal modes under the peak of the diffraction response which implies that the cavity will oscillate at any value of $L_c$.

Returning to the question of synchronicity a naive estimate is that the laser will mode hop when $|\delta \nu_d - \delta \nu_c| \sim \nu_{\text{FSR}}$. Assuming $l = l_0 + y$ we can show that the scan distance $\delta \nu$ before the synchronicity error is $\nu_{\text{FSR}}$ is given by

$$\delta \nu = \nu_{\text{FSR}} \frac{L_c}{y}.$$  

Using the above parameters and $y = 1$ mm gives $\delta \nu = 150$ GHz. Although this estimate is quite crude we expect that building the cavity with the error in $l$ not more than $\sim 1$ mm should allow scans of 10’s of GHz.

### 7.3.2 Frequency comb stability

The optical frequency of one tooth of an optical frequency comb is given by

$$f = N f_{\text{rep}} + f_{\text{off}}$$

where $N$ is the index of the comb tooth, $f_{\text{rep}}$ is the cavity repetition rate, and $f_{\text{off}}$ is the offset frequency from the nearest tooth which comes from the $f - 2f$ lock. The uncertainty in $f$ due to the uncertainty of the repetition rate and offset lock circuits is

$$\delta f = N \delta f_{\text{rep}} + \delta f_{\text{off}}.$$  

Consider a Menlo Systems Yb fiber comb with $f_{\text{rep}} = 250$ MHz. Typical numbers for the comb lock uncertainties are $\delta f_{\text{rep}} = 3$ mHz and $\delta f_{\text{off}} = 5$ Hz. We use this to look at a 920 nm light beam with frequency 326 THz, so $N = 1.3 \times 10^6$. The lock uncertainties therefore imply

$$\delta f = 3.9 \text{ kHz}.$$  

In addition each tooth of the comb is broadened to about $\delta f_{\text{tooth}} = 100$ kHz due to phase noise. The $\delta f$ set by the lock uncertainties can only be reached after averaging for a time $t$ such that $\delta f_{\text{tooth}}/\sqrt{\delta f_{\text{tooth}}^2 t} \sim \delta f$ or

$$t \sim \frac{\delta f_{\text{tooth}}^2}{(\delta f)^2} = 6.6 \text{ ms}.$$
This time is short enough that it is not of concern for our application of stabilizing a cw laser which is narrowed using a stable cavity.

In addition to this uncertainty there is the uncertainty due to the instability of the low frequency reference oscillator. The repetition rate is referenced to a to a Novus OXCO with $f_{\text{osc}} = 10 \text{ MHz}$. The oscillator is slaved to a GPS timing signal for long term stability. The repetition rate multiplier is $N_{\text{osc}} = f_{\text{rep}}/f_{\text{osc}} = 25$. The stability floor due to the external reference is

$$\delta f' = (NN_{\text{osc}})\delta f_{\text{osc}}.$$ 

The Novus has $\delta f_{\text{osc}}/f_{\text{osc}} = 4 \times 10^{-10}$ at 1 sec. giving

$$\delta f' = 130 \text{ kHz}.$$ 

At 10 sec. the Novus improves to $\delta f_{\text{osc}}/f_{\text{osc}} = 2.5 \times 10^{-11}$ giving

$$\delta f' = 8.1 \text{ kHz}.$$ 

Influence of the reference oscillator phase noise has not been considered.

A more stable reference oscillator would make a big difference at 1 sec. averaging time, but at 10 sec. we could only win a factor of about 2, given our current comb lock uncertainties.

### 7.4 Photodetection and signal to noise ratio

#### 7.4.1 Thermal noise

Let’s start by recalling some basics about useful systems of units and the background thermal noise level of electrical components. It is convenient to introduce a logarithmic scale for measuring electrical quantities. Thus a given power level can be described on a logarithmic scale as a number of decibels or dB, defined by the relation

$$dB \equiv 10 \log_{10} P,$$ 

where $P$ is the power. Equation (7.3) defines a relative scale so that the relation between two power levels $P_1$ and $P_2$ measured in dB is simply $10 \log_{10} P_1/P_2$.

An absolute power scale is often needed. Therefore the notation dBm has been introduced, where dBm are simply dB’s referenced to a power level of 1 mW. Thus a power $P$ measured in dBm is

$$P[\text{dBm}] \equiv 10 \log_{10} \frac{P[\text{W}]}{0.001},$$

(7.4) Unless explicitly written otherwise it is always assumed that dBm refer to a power and not a voltage or current level.

Let’s now specialize to 50 $\Omega$ systems. The root mean square or rms power load on a resistor $R$ is $P = RI^2 = V^2/R$, where $I$ and $V$ are the rms current and voltage respectively. Thus the power load on a 50 $\Omega$ resistor can be written in terms of the rms currents and voltages as

$$P_{50\Omega} = 47. + 10 \log_{10} \frac{I_{\text{rms}}^2}{V_{\text{rms}}^2} [\text{dBm}],$$

(7.5) 

$$= 13. + 10 \log_{10} \frac{V_{\text{rms}}^2}{I_{\text{rms}}^2} [\text{dBm}].$$

(7.6)
We can always express these relations in terms of peak to peak waveforms \((f(t) = (f_{pp}/2) \cos \omega t)\) using \(f_{pp} = 2\sqrt{2}f_{rms}\).

Electrical components at temperature \(T\) exhibit noise due to the random thermal excitation of charges. This is known as Johnson or Nyquist noise, or simply thermal noise. Unless other noise sources are larger the thermal noise sets a lower limit to the detection of weak signals. A detector of resistance \(R\) generates a flat noise spectrum (white noise) with a rms power level

\[
P_{th} = 4k_BTB,
\]

where \(k_B\) is Boltzmann’s constant and \(B\) is the measurement bandwidth. Equivalently we can write

\[
I^2_{th} = \frac{4k_BTB}{R}
\]

and

\[
V^2_{th} = 4k_BTRB.
\]

Assuming a room temperature of 300 K we note that \(k_BT = 0.026\) eV = \(4.16 \times 10^{-21}\) J and the thermal noise power is given by

\[
P_{th} = -167.8 + 10 \log_{10} B[\text{Hz}] \text{ dBm}.
\]

This is generally a small number: the thermal noise power in room temperature 50Ω systems in a 1 MHz bandwidth is about -114. dBm. Equivalently the thermal noise density is 0.91 nV/Hz\(^{1/2}\).

### 7.4.2 Detector characteristics

The CO\(_2\) laser scattering system uses photoconductive HgCdTe detectors of type PCI-L-3 from Vigo Systems Ltd. These are room temperature devices with an active area of 1 mm\(^2\). The damage threshold of these detectors is 100 W/cm\(^2\), so the maximum input power should not exceed 1 W. Relevant data from the test sheets delivered with the detectors are shown in Table 7.1.

The detector electronics are shown in Fig. 7.4. The detectors are biased using a load resistor of \(R_L = 950\) Ω. The voltage across the detector resistance \(R_d\) is amplified by a Perry Amplifier model 490. It is a 50 Ω device (input and output impedance) with 26 dB of gain.
### 7.4 Photodetection and signal to noise ratio

Table 7.1: Detector parameters. The derived results are calculated from the supplied data assuming $\lambda = 10.6 \, \mu m$ and a quantum efficiency $\eta = 0.6$. The noise figure is the excess detector noise relative to the ideal thermal noise of a resistor with the same resistance as the detector.

<table>
<thead>
<tr>
<th></th>
<th>Units</th>
<th>no. 5467 PCI-L-3</th>
<th>no. 5541 PCI-L-3</th>
<th>no. 5300 PCI-L</th>
</tr>
</thead>
<tbody>
<tr>
<td>type</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>test conditions</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>temperature</td>
<td>$T$ K</td>
<td>293</td>
<td>293</td>
<td>296</td>
</tr>
<tr>
<td>detector area</td>
<td>$A_d$ m²</td>
<td>$1 \times 10^{-6}$</td>
<td>$1 \times 10^{-6}$</td>
<td>$1 \times 10^{-6}$</td>
</tr>
<tr>
<td>bias current</td>
<td>$I_b$ mA</td>
<td>10</td>
<td>10</td>
<td>20</td>
</tr>
<tr>
<td>chopper frequency</td>
<td>Hz</td>
<td>397</td>
<td>397</td>
<td>400</td>
</tr>
<tr>
<td>optical power</td>
<td>$P$ W</td>
<td>0.002</td>
<td>0.002</td>
<td>0.0015</td>
</tr>
<tr>
<td>test results</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>signal voltage</td>
<td>$V$ $\mu V$</td>
<td>3000</td>
<td>2500</td>
<td>2600</td>
</tr>
<tr>
<td>vol. responsivity</td>
<td>$\mathcal{R}_v$ V/W</td>
<td>1.5</td>
<td>1.25</td>
<td>1.7</td>
</tr>
<tr>
<td>noise density</td>
<td>$nV / Hz^{1/2}$</td>
<td>1.1</td>
<td>1.25</td>
<td>1.55</td>
</tr>
<tr>
<td>detectivity</td>
<td>$D^*$ cm Hz$^{1/2}$/W</td>
<td>$1.3 \times 10^8$</td>
<td>$1 \times 10^8$</td>
<td>$1.1 \times 10^8$</td>
</tr>
<tr>
<td>detector resistance</td>
<td>$R_d$ $\Omega$</td>
<td>52</td>
<td>53</td>
<td>73</td>
</tr>
<tr>
<td>derived results</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>noise figure</td>
<td>dB</td>
<td>1.65</td>
<td>2.76</td>
<td>4.63</td>
</tr>
<tr>
<td>NEP (1 MHz BW)</td>
<td>NEP W</td>
<td>$7.7 \times 10^{-7}$</td>
<td>$1 \times 10^{-6}$</td>
<td>$9.1 \times 10^{-7}$</td>
</tr>
<tr>
<td>photoconductive gain</td>
<td>$G$</td>
<td>$5.5 \times 10^{-3}$</td>
<td>$4.5 \times 10^{-3}$</td>
<td>$4.5 \times 10^{-3}$</td>
</tr>
</tbody>
</table>
over a bandwidth of 1 kHz to 100 MHz, and a noise figure of 2.8 dB. The maximum output level is 3 V_{pp}.

Following the preamp there is a Minicircuits ??? MHz low pass filter, and then a Boston Electronics 493A/40 wide band amplifier with 40 dB gain in a bandwidth of 1 kHz to 500 MHz. The 493A/40 has a maximum output of 5 V_{pp} and an input noise of 50 µV rms, wideband. This translates into an input noise power in the full bandwidth of -73. dBm, or -83. dBm in a 50 MHz bandwidth. Since the first stage 26 dB preamp amplifies its input thermal noise (+ noise figure) to a level of -62. dBm in a 50 MHz bandwidth the noise at the input to the second preamp is completely negligible. Finally, following the 2nd preamp there is a Boston Electronics 491 video line driver with unity gain.

Given the above data the noise level seen at the output of the detector/preamp assemblies should be the thermal limit increased by the noise figures of the detector and the first stage preamp, and amplified by the total gain. This translates to a noise level using the Hameg HM5005 spectrum analyzer 250 kHz bandwidth setting of about -43. dBm. This level agrees to within 1-2 dB of baseline measurements done at W7-AS in November '98. There is no noticeable increase of the noise level when the detector bias current is increased from 0 to 10 mA.

### 7.4.3 Photodetection

The scattering diagnostic is based on a heterodyne detection scheme where a strong local oscillator is mixed with the weak signal due to scattering from the plasma. To understand why a simpler direct detection scheme is not practical for this experiment we need to account for the noise processes associated with photodetection.

The current responsivity $R_i$ of the detector is defined by

$$R_i = \frac{I}{P}, \quad (7.11)$$

where $I$ is the photogenerated current and $P$ is the optical power illuminating the detector$^1$. The photogenerated current is given by

$$I = e\eta \frac{P}{h\nu} G, \quad (7.12)$$

where $e$ is the electronic charge, $\eta$ is the quantum efficiency of the detector, $h$ is Planck’s constant, $\nu$ is the optical frequency, and $G$ is the photoconductive gain. The photoconductive gain is a dimensionless quantity given by $G = \tau / \tau_{tr}$ where $\tau$ is the lifetime of a photogenerated charge carrier, and $\tau_{tr}$ is the transit time across the device. Since the transit time depends on the applied electric field the gain increases with bias current. See, for example, Dereniak and Boreman$^?$ for a detailed discussion. Using Eqs. (7.11,7.12) we have

$$R_i = \frac{e\eta}{h\nu} G. \quad (7.13)$$

The voltage responsivity is then given by

$$R_v = R_{eq} R_i = \frac{e\eta}{h\nu} G R_{eq}, \quad (7.14)$$

$^1$We will generally use calligraphic type face for optical quantities.
where \( R_{eq} = \frac{R_d R_L}{R_d + R_L} \) is the equivalent resistance of the detector and load resistors in parallel. Note that the responsivity is proportional to the product \( \eta G \), and it would be necessary to measure additional characteristics of the carrier transport in the device in order to determine separate values of \( \eta \) and \( G \). We have assumed that \( \eta = 0.6 \) in order to calculate the photoconductive gain given in Table 7.1. While the photoconductive gain of cooled devices can be much larger than unity, operation at room temperature decreases the carrier lifetime, and hence the gain.

In a direct detection experiment the signal to noise ratio of the resulting electronic signal power is

\[
z = \frac{I^2}{I^2_{th} + I^2_{gr} + I^2_{pa} + I^2_{1/f}},
\]

where

\[
I^2_{th} = \frac{4k_B T B}{R_{eq}},
\]

\[
I^2_{gr} = 4e^2 G^2 B \left( \frac{P + P_{bg}}{h\nu} + g_{th} \right),
\]

\[
I^2_{pa} = \frac{4k_B T B}{R_{eff,pa}},
\]

are the thermal, generation recombination, and preamplifier noise powers respectively, and \( I^2_{1/f} \) is the \( 1/f \) noise. In the above \( P_{bg} \) is the optical power due to background radiation, \( g_{th} \) is the thermal (phonon assisted) generation rate of charge carriers, and \( R_{eff,pa} \) is the noise effective input impedance of the preamplifier\(^2\). Since we will be working at an intermediate frequency of 40 MHz the \( 1/f \) noise can be assumed negligible. Equation (7.15) can then be rewritten in the form

\[
z = \frac{1}{4B} \frac{e^2 G^2 g_{th}^2}{R_{eq} + \frac{k_B T}{R_{eff,pa}} + e^2 G^2 (g + g_{bg} + g_{th})},
\]

where \( R_{eff} = \frac{R_{eq} R_{eff,pa}}{(R_{eq} + R_{eff,pa})} \) is the combined noise effective impedance of the detector, the load resistor, and the preamplifier, \( g = \eta P / h\nu \) is the carrier generation rate due to the signal, and \( g_{bg} = \eta P_{bg} / h\nu \) is the carrier generation rate due to background radiation.

Depending on which term in the denominator of Eq.(7.19) dominates detection will be thermal or background limited. Consider first the case where thermal noise dominates. Neglecting the \( 2^{nd} \) term in the denominator and setting the signal to noise ratio to unity we find for the minimum detectable power,

\[
P_{min,th} = \sqrt{\frac{4k_B T B}{R_{eff}} \frac{h\nu}{e\eta G}}.
\]

Using the parameters given in Table 7.1 for detector 5467 and using a 1 MHz bandwidth we find \( P_{min,th} = 1.2 \times 10^{-6} \) W. This agrees well with the NEP calculated in Table 7.1 since we have here included the degradation due to the preamplifier noise. Thus at mW optical power levels detection is thermally limited.

---

\(^2\)The noise effective input impedance is \( R_{eff,pa} = R_{pa} 10^{-NF/10} \), where \( NF \) is the noise figure in dB.
It is instructive to calculate how large an optical power is required for detection to be background limited. The thermal carrier generation rate is unknown, but as a first approximation we may set it equal to the background rate \( g_{bg} \). The background rate is found by integrating the Planck distribution for thermal radiation times the spectral dependence of the photodetector sensitivity. As a rough estimate we may take the effective spectral bandwidth of the detector as 1 \( \mu \)m and using the curve given in \([?]\) p.69, estimate \( \mathcal{P}_{bg} \sim 3 \times 10^{-5} \) W for our 1 mm\(^2\) detectors. The background generation rate is then \( g_{bg} = 9.6 \times 10^{14} \) sec\(^{-1}\). Equating the terms in the denominator of Eq. (7.19) gives

\[
g^* = \frac{k_B T}{R_{eff} e^2 G^2} - 2g_{bg}
\]

for the carrier generation rate \( g^* \) at which the signal dependent noise is equal to the thermal noise. We find \( g^* = 3.2 \times 10^{20} \) sec\(^{-1}\) or, equivalently, for the corresponding optical power \( \mathcal{P}^* = 10 \) W. Since \( \mathcal{P}^* \) is considerably greater than the damage level of the detector (which is 1 W) it is not possible to reach the signal noise limit in direct detection with the room temperature detectors we are using.

### 7.4.4 Heterodyne detection

In order to improve the sensitivity of the scattering diagnostic we adopt a heterodyne detection scheme. This means that the weak signal to be measured is combined with a much stronger local oscillator at the surface of the detector. Equation (7.15) for the signal to noise ratio then takes the form

\[
z = \frac{\mathcal{I}_{hd}^2}{\mathcal{I}_{lo}^2 + \mathcal{I}_{th}^2 + \mathcal{I}_{pa}^2 + \mathcal{I}_{1/2}^2},
\]

with

\[
\mathcal{I}_{hd} = \frac{\eta e}{h \nu} \eta_{hd} \sqrt{\mathcal{P}_{lo} G},
\]

\[
\mathcal{I}_{lo}^2 = 4e^2 G^2 B \frac{\eta \mathcal{P}_{lo}}{h \nu},
\]

where \( \mathcal{P}_{lo} \) is the optical power of the local oscillator and \( \eta_{hd} \) is a factor less than one that accounts for the efficiency of the heterodyne mixing at the detector surface. \( \eta_{hd} \) requires some care to measure and we will assume that it has a numerical value of 1. Defining in the same fashion as above a carrier generation rate due to the local oscillator \( g_{lo} \), and again neglecting 1/f noise, we can rewrite Eq. (7.22) as

\[
z = \frac{1}{4B} \frac{\eta_{hd} e^2 G^2 g_{lo}}{\mathcal{P}_{lo}^2 + e^2 G^2 (g + g_{lo} + g_{bg} + g_{th})}.
\]

As we found for the case of direct detection thermal noise is always dominant with the room temperature detectors. We can therefore simply write

\[
z = \frac{R_{eff}}{4k_B T B} \eta_{hd} e^2 G^2 g_{lo},
\]

\[
= \frac{R_{eff}}{4k_B T B} \left( \frac{\eta \eta_{hd} e G}{h \nu} \right)^2 \mathcal{P}_{lo}
\]

(7.26)
7.4 Photodetection and signal to noise ratio

<table>
<thead>
<tr>
<th>detection scheme</th>
<th>$P_{\text{min,th}}$ [W]</th>
<th>0.1</th>
<th>1.0</th>
<th>10.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>direct</td>
<td>$3.8 \times 10^{-7}$</td>
<td>$1.2 \times 10^{-6}$</td>
<td>$3.8 \times 10^{-6}$</td>
<td></td>
</tr>
<tr>
<td>power limited heterodyne</td>
<td>$1.3 \times 10^{-12}$</td>
<td>$1.3 \times 10^{-11}$</td>
<td>$1.3 \times 10^{-10}$</td>
<td></td>
</tr>
<tr>
<td>quantum limited heterodyne</td>
<td>$1.2 \times 10^{-14}$</td>
<td>$1.2 \times 10^{-13}$</td>
<td>$1.2 \times 10^{-12}$</td>
<td></td>
</tr>
</tbody>
</table>

Table 7.2: Minimum detectable power in Watts of the detectors and electronics installed at W7-AS for different detection limits and bandwidths.

for the signal to noise ratio. We see that the power of the optical signal for which the signal to noise ratio is unity is inversely proportional to the local oscillator power and is given by

$$P_{\text{min,hd}} = \frac{4k_B T B}{R_{\text{eff}}} \left( \frac{h\nu}{\eta_{\text{hd}} e G} \right)^2 \frac{1}{P_{\text{lo}}}. \quad (7.27)$$

Setting $P_{\text{lo}}$ to 0.1 times the power at which detector damage occurs we find a minimum detectable optical power of $1.3 \times 10^{-11}$ W, assuming a 1 MHz detection bandwidth. This is about five orders of magnitude less than the minimum detectable power found above for direct detection which demonstrates the superiority of the heterodyne configuration.

Nonetheless with a cooled infrared detector it is possible to achieve signal, or quantum, noise limited operation. In this case the dominant noise source is that generated by the local oscillator. Hence, from Eq. (7.25), the quantum limited signal to noise ratio is

$$z_{\text{ql}} = \frac{1}{4B} \eta_{\text{ld}}^2 g. \quad (7.28)$$

Setting $z_{\text{ql}}$ to unity we find the minimum detectable power in a 1 MHz bandwidth to be

$$P_{\text{min,ql}} = \frac{4Bh\nu}{\eta_{\text{hd}}^2} = 1.2 \times 10^{-13} \text{ W.} \quad (7.29)$$

In Table 7.2 we compare the minimum detectable powers for several different bandwidths for the cases of thermally limited direct detection, local oscillator power limited heterodyne detection, and quantum limited detection. We can conclude that a cooled infrared detector would increase the sensitivity by up to two orders of magnitude. This advantage has to be evaluated against the additional inconvenience, or cost, of cooled detectors. At the current time (November 1998) the limiting factor in our measurements at small $k_{\text{scat}}$ is not the power level of the local oscillator, but rather amplification of the 40 MHz leakage signal due to parasitic scattering of the main beam into the local oscillator. Improvements to the beam scraper on the receiver table are intended to improve this situation.

Another issue is how much electronic amplification is needed to record a low-level signal $P_{\text{min,hd}}$? The analog to digital electronics we are using have an input full range of 2 V$_{\text{pp}}$, or +10 dBm. An 8 bit conversion is made so the dynamic range is about 48 dB, and the minimum recordable input level is -38 dBm. The maximum electronic gain is $66 + 37.5 = 103.5$ dB, thus the minimum recordable power level at the detector is about -142 dBm.

The photocurrent obtained for a minimum power $P_{\text{min,hd}}$ is

$$I_{\text{min,hd}} = \frac{\eta e}{h\nu} \eta_{\text{hd}} \sqrt{P_{\text{min,hd}} P_{\text{max,lo}} G}, \quad (7.30)$$

March 8, 2016 M. Saffman
and the corresponding electrical power is

\[ P_{\text{min,hd}} = \frac{V^2}{R_d} = \frac{(R_{\text{eq}} I_{\text{min,hd}})^2}{R_d} = \frac{R_{\text{eq}}^2}{R_d} \left( \frac{\eta_{\text{hd}} eG}{h\nu} \right)^2 P_{\text{min,hd}} P_{\text{max,lo}}. \] (7.31)

Putting in the numbers for the 1 MHz bandwidth example we find \( P_{\text{min,hd}} = -103 \) dBm. Thus at full electronic gain the minimum detectable signal will be amplified to about full range at the A/D converter.

A reasonable electronic amplification level in practice will be that for which the amplified thermal noise corresponds to say 4 bits after A/D conversion. This implies an electronic input gain setting of roughly 30 dB on the A/D card.

### 7.5 References


Index