EE 118: Introduction to Optical Engineering **Spring 2019**

Lecture 1: Introduction

Lecturer: Laura Waller *January 23* *Aditya Sengupta*

Note: *LaTeX format adapted from template courtesy of UC Berkeley EECS dept.*

1.1 Logistics

This class will cover optical physics, imaging systems, and optical devices. Laura Waller Office Hours: Cory 514, Wed 10-11am; Gautam: Cory 504, Mon 5-6PM.

Problem sets to be submitted under the door (Cory 514) on the day (1/30 for the first one which is already out).

Quizzes (basically midterms) on March 4 and April 24. Final project in groups of 3-4 - do either simulations or experiments. Give a short presentation in class. Level of reproducing a journal paper result. Grad student or postdoc mentor will be assigned.

1.2 Optics Abstraction Layers and Basic Terminology

Paraxial ray optics $\subset$ geometrical optics (ray tracing) $\subset$ wave optics $\subset$ electromagnetic optics $\subset$ quantum optics. Use the approximation that fully describes what you want to do. Geometric optics is ray optics without the small angle approximation. Wave optics is for light interference. Electromagnetic optics is wave optics with "vector effects" such as polarization.

The speed of light is independent of the wavelength of light. It is a constant 299,792,000 m/s. Light exists at many wavelengths, of which only a narrow band is visible.

Wave motion: a wave travels along a line of propagation, and individual particles move perpendicular to the line (up and down). The shape of a wave is described by a function $f(x-vt)$. At any instant, the shape can be found by setting time constant: $\psi(x,t)|_{t=0} = f(x,0) = f(x)$.

A wave is defined by several parameters. The amplitude is the height of a crest or trough from the line of propagation. The wavelength is the length along the line of propagation that covers one crest and one trough. Waves are described in time by their frequency f, their period T, and their velocity $v = \lambda f$. Waves do not have to be sinusoidal or periodic; a valid wave only has to be of the form $f(x-vt)$.

1.3 Deriving the 1D Wave Equation

Consider a wave of the form

$$\psi(x,t) = f(x-vt) \tag{1.1}$$

and let $x' = x - vt$. Hold time constant and differentiate with respect to space.

$$\frac{\partial \psi}{\partial x} = \frac{\partial f}{\partial x'}\frac{\partial x'}{\partial x} = \frac{\partial f}{\partial x'} \tag{1.2}$$

where the second term goes to 1 because of $\frac{\partial x'}{\partial x} = \frac{\partial (x-vt)}{\partial x} = 1$ for constant v and t. Now, take a derivative with respect to time:

$$\frac{\partial \psi}{\partial t} = \frac{\partial f}{\partial x'}\frac{\partial x'}{\partial t} = -v\frac{\partial f}{\partial x'} \tag{1.3}$$

because $\frac{\partial x'}{\partial t} = -v$. Take a second derivative in space (we'll discuss why afterwards):

$$\frac{\partial^2 \psi}{\partial x^2} = \frac{\partial}{\partial x}\left(\frac{\partial f}{\partial x'}\right) = \frac{\partial^2 f}{\partial x'^2} \tag{1.4}$$

and in time:

$$\frac{\partial^2 \psi}{\partial t^2} = \frac{\partial}{\partial t}\left(-v\frac{\partial f}{\partial x'}\right) = -v\frac{\partial}{\partial t}\left(\frac{\partial f}{\partial x'}\right) \tag{1.5}$$

$$\frac{\partial^2 \psi}{\partial t^2} = -v\frac{\partial}{\partial x'}\left(\frac{\partial f}{\partial t}\right) \tag{1.6}$$

Note that $\frac{\partial f}{\partial t} = \frac{\partial f}{\partial x'}\frac{\partial x'}{\partial t} = -v\frac{\partial f'}{\partial x'}$, and therefore we can get

$$\frac{\partial^2 \psi}{\partial t^2} = v^2\frac{\partial^2 f}{\partial x'^2} \tag{1.7}$$

Combining these two second derivatives, we get the wave equation:

$$\frac{\partial^2 \psi}{\partial x^2} = \frac{1}{v^2}\frac{\partial^2 \psi}{\partial t^2} \tag{1.8}$$

If a function ψ satisfies this, it is a valid wave. We took the second derivative so that there was no ambiguity in direction, as in the end we get a v^2 factor which is necessarily positive.

If we had speed as a function of time $v(t)$, there would be more partial derivative terms.

1.4 Harmonic waves

Most waves are sinusoidal, i.e. $\psi(x,t)|_{t=0} = A\sin(kx)$. k is the propagation number, with units of radians per spatial unit (such as rad/m). k is inversely related to λ. Now, we include time:

$$\psi(x,t) = A\sin(k(x-vt)) = f(x-vt) \tag{1.9}$$

We can explicitly show a relationship between k and λ by adding a phase factor to x. Shifting x by a wavelength should not affect the equation:

$$\sin(k(x-vt)) = \sin\left(k(x\pm\lambda) - vt\right) = \sin\left(k(x-vt)\pm 2\pi\right) \tag{1.10}$$

which shows us that

$$|k\lambda| = 2\pi \implies k = \frac{2\pi}{\lambda} \tag{1.11}$$

We can derive the wave equation for the specific case of this harmonic wave by taking partial derivatives, as with the more general case:

$$\psi = A\sin(k(x-vt)) \implies \frac{\partial^2\psi}{\partial x^2} = k\frac{\partial}{\partial x}\left(A\cos(kx-kvt)\right) = -k^2 A\sin(kx-kvt) \tag{1.12}$$

$$\frac{\partial^2\psi}{\partial t^2} = \frac{\partial}{\partial t}\left(A\cos(kx-kvt)\right)(-kv) = -k^2v^2 A\sin(kx-kvt) \tag{1.13}$$

Therefore the two second derivatives are off by a factor of v^2, as desired.

We can define the temporal frequency $\nu = 1/\tau$, where τ is the temporal period of the wave. It is related to the velocity by $v = \nu\lambda$. Angular temporal frequency $\omega = 2\pi\nu$ is also sometimes used.

1.5 Phases

The phase is the argument of the sinusoid. We can write the sinusoid as

$$\psi(x,t) = A\sin(kx-\omega t) \tag{1.14}$$

We see that the phase is a function of x and t. The velocity of a constant phase is called the phase velocity, $\frac{\omega}{k}$, but that will not be used much.

We can use the wave equation to check whether a wave is valid as well as to find its velocity. Examples:

$$\psi(z,t) = (az-bt)^2 = a^2z^2 - 2azbt + b^2t^2 \tag{1.15}$$

This satisfies the wave equation; we see that $2a^2$ is the second spatial derivative, and $2b^2$ is the second time derivative. Therefore

$$2a^2 = \frac{1}{v^2}2b^2 \implies v = \pm\frac{b}{a} \tag{1.16}$$

Since we see that the wave is travelling in the $+z$ direction, the velocity is $\frac{b}{a}$.

Similarly, shifting this by a constant still satisfies the wave equation: $(ax+bt+c)^2$ also satisfies the wave equation. Waves have to have both spatial and time dependence, so $\frac{1}{ax^2+b}$ is not a valid wave because its second time derivative is zero which in general does not satisfy the wave equation.

1.6 Transverse and Longitudinal Waves

Particles in a wave can either move perpendicular to the wave direction, or parallel to it - transverse and longitudinal. In the second case, instead of crests and troughs, the wave will consist of compressions and rarefactions, e.g. a sound wave.

Light is a transverse electromagnetic wave, with two fields transverse to the direction of motion: electric and magnetic. These fields are in phase (they have the same time dependence of phase). The wave travels at the speed of light in a vacuum.

In general, we describe only the E field:

$$E = E_0\cos(\omega t - kz) = \operatorname{Re}\left(E_0 e^{i(\omega t - kz)}\right) \tag{1.17}$$

It is necessary to know how to work with phasors (exponential representations of complex numbers), how to add them, etc. The two representations of complex numbers are connected by Euler's identity $e^{i\theta} = \cos\theta + i\sin\theta$.

We can combine solutions using complex numbers by $E_1 + iE_2$, which we can convert to polar form. Let $E_1 = E_0\cos(kz-\omega t)$ and let $E_2 = E_0\sin(kz-\omega t)$. Then

$$E_1 + iE_2 = E_0 e^{ikz}e^{i\omega t} \tag{1.18}$$

Only the e^{ikz} part of this is significant for linear media (everything we will deal with).

1.7 Plane waves

Plane waves are one-dimensional waves travelling in 2D or 3D. They are observed as wavefronts (lines) on a plane or in free space, and use trigonometry to characterize them. For instance, if a wavefront is propagating with wavelength λ at an angle θ from an optical axis (say the z-axis), its wavelength with respect to the optical axis is $\lambda_z = \frac{\lambda}{\cos\theta}$, and with respect to the y-axis (assuming the y-z plane) is $\lambda_y = \frac{\lambda}{\sin\theta}$.

We represent plane waves either with real or phasor notation. Real notation is of the form

$$\psi(\vec{r},t) = A\cos\left(\frac{2\pi}{\lambda}(z\cos\theta + y\sin\theta - \omega t)\right) \tag{1.19}$$

and phasor notation is

$$\psi(\vec{r}) = A\text{exp}\left(i\left(\frac{2\pi}{\lambda}(z\cos\theta + y\sin\theta - \omega t)\right)\right) \tag{1.20}$$

These are essentially the $kx - \omega t$ factor in multiple dimensions, with components of different directions adding together.

EE 118: Introduction to Optical Engineering **Spring 2019**

Lecture 2: Rays and Refraction

Lecturer: Laura Waller *January 28* *Aditya Sengupta*

2.1 Recap

Last time, we saw that light is a harmonic electromagnetic wave, described by the equation

$$E = E_0 \cos(\omega t - kz) = \mathrm{Re}\left(E_0 e^{i(\omega t - kz)}\right) \tag{2.1}$$

and we derived the wave equation:

$$\frac{\partial^2 E}{\partial x^2} = \frac{1}{v^2}\frac{\partial^2 E}{\partial t^2} \tag{2.2}$$

2.2 Spherical Waves

A spherical wave is created by waves coming out spherically from a point source, e.g. a small oscillating electric field. To describe this in an equation, we denote a surface of constant phase as having a constant kr, where r is the radial distance. The wave expression becomes

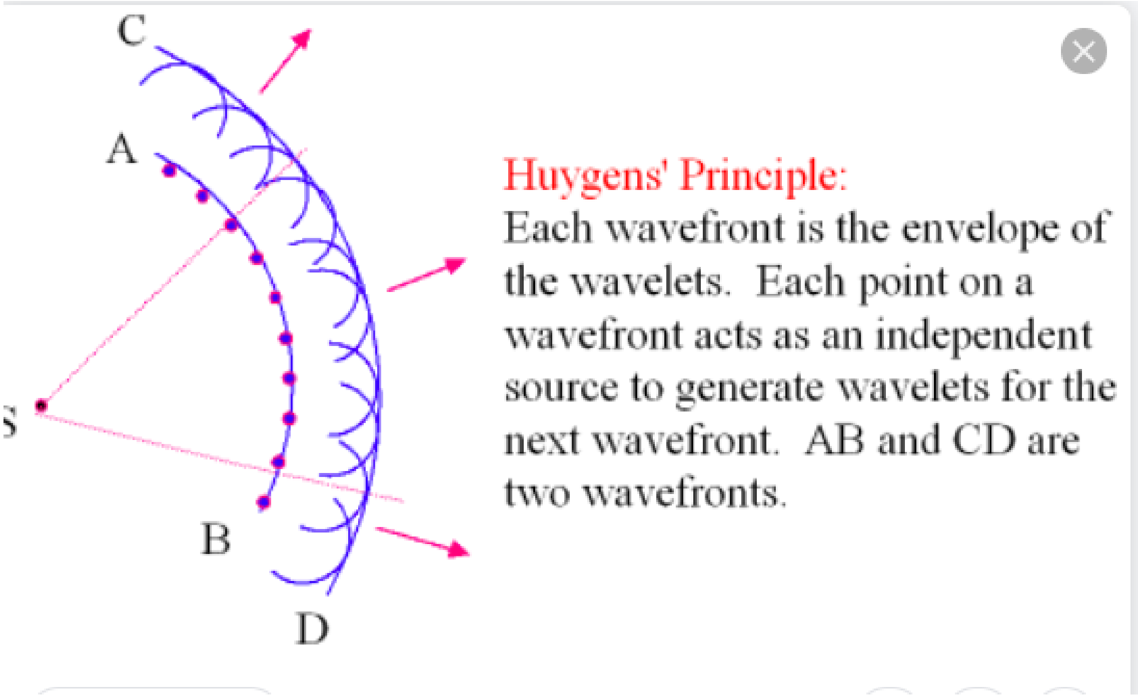
$$\vec{E}(\vec{r}, t) = \frac{E_0}{r}\cos(kr - \omega t) \tag{2.3}$$

This field drops off as r even though spheres usually drop off as r^2, because the energy (which is the squared magnitude of the field) drops off as r^2.

Wavefronts either radiate away from or travel toward a point - these are *diverging* or *converging* waves.

In the limit as $z \to \infty$ (if the point source is on a fixed small point on the z axis and spherical waves diverge from it) spherical waves become plane waves.

Huygens' principle states that every point on a wavefront serves as a source of spherical secondary wavelets. This means that every single point on a spherical wavefront would have its own spherical wave representation centered there, and their electric fields can be added. These secondary wavelets whose fields are added up constructively combine to make the next primary spherical wavefront.



Huygens' Principle:
Each wavefront is the envelope of the wavelets. Each point on a wavefront acts as an independent source to generate wavelets for the next wavefront. AB and CD are two wavefronts.

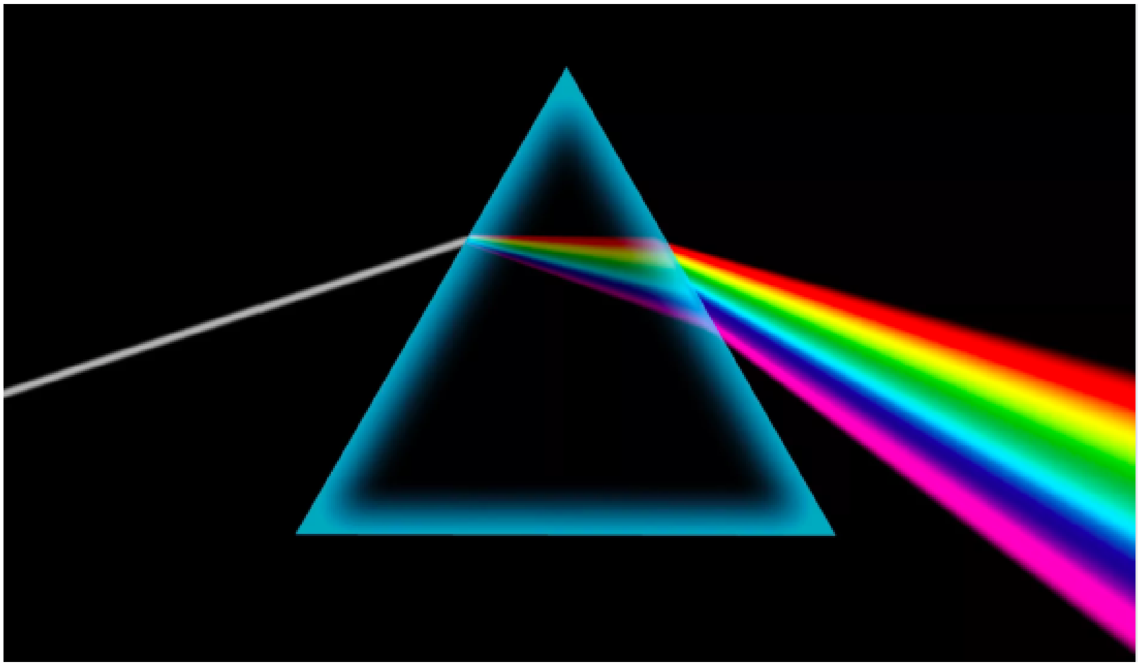
2.3 Refractive Index

The refractive index is how we describe how optically dense a material is. It is the ratio of the speed of light in a vacuum c to that in a medium v.

$$n = \frac{c}{v} \tag{2.4}$$

Colour is determined by the wavelength in free space, not in the medium. The frequency remains the same (because $E = hf$ and energy remains invariant), so we can think of colour being determined by frequency, but the wavelength changes because the speed of the wave changes ($v = \lambda f$).

The refractive index varies slightly with the wavelength of the incident light. Conventionally, refractive indices are described at a specific middle wavelength, but in general they are a function of wavelength. The canonical example of dispersion, in which different colours are bent by different amounts because of the slightly different refractive index, is a prism.



If two strings (one thick and one thin) are connected and a wave crosses the connection point, then the propagation speed and wavelength change at the point, but not the frequency or period. The wavelength increases when the string goes from thick to thin.

2.4 Optical Path Length

The optical path length is the integral of the refractive index over the path that the light travels. The time required for a ray to traverse a path is the OPL/c.

$$OPL = \int_{\text{path}} n(r)ds \tag{2.5}$$

Suppose there is an irregularly-shaped object whose thickness is described by $s(x)$. The OPL is then given by $\Delta n \cdot s(x)$. It is related to phase delay (where the phase is the argument to the cosine of the wave expression) because k varies with different refractive index. Recall that $k = \frac{2\pi}{\lambda}$, and λ varies. Therefore the phase delay due to the different medium is given by

$$\phi(z) = \frac{2\pi}{\lambda} OPL \tag{2.6}$$

2.5 Fermat's Principle and Snell's Law

A ray of light is defined as any path for which the total OPL is stationary. A formal definition of stationary requires calculus of variations, which is beyond the scope of this class. Fermat's principle means that a ray must always follow a path for which no other path is optically shorter. Immediately, from this, we can conclude that in a homogeneous medium light travels in a straight line.

Rays of light bend when the index of refraction changes. The shortest path in both media individually is a straight line, where the angle varies at the interface between the two. The relationship between these two angles is given by Snell's law,

$$n_1 \sin\theta_1 = n_2 \sin\theta_2 \tag{2.7}$$

Imagine a lifeguard running to save someone in the water. On sand, she runs at speed v and in the water she travels at speed v/n. To minimize her travel time, what angle should she start out with?

Assume the lifeguard starts at $(0,0)$ and wants to get to (a,b), where the sand/water interface is parallel to the y axis at some x distance s. Suppose she crosses the interface at (s,c). The time taken is

$$t = \frac{\sqrt{s^2+c^2}}{v} + \frac{(a-s)^2+(b-c)^2}{v/n} \tag{2.8}$$

which can be minimized by taking a derivative with respect to c, the variable quantity (the crossing point), and setting it to zero:

$$\frac{dt}{dc} = \frac{c}{v\sqrt{s^2+c^2}} + n\frac{c-b}{v\sqrt{(a-s)^2+(b-c)^2}} = 0 \tag{2.9}$$

The angle between the origin and (s,c) is given by $\theta = \arctan\frac{s}{c}$, and that between (s,c) and (a,b) is $\theta' = \arctan\frac{c-b}{s-a}$. Converting to sines, we get

$$\sin\theta = \frac{c}{\sqrt{c^2+s^2}} \quad \sin\theta' = \frac{c-b}{\sqrt{(c-b)^2+(s-a)^2}} \tag{2.10}$$

which we can substitute into the derivative equation,

$$\frac{\sin\theta}{v} - \frac{\sin\theta'}{v/n} = 0 \tag{2.11}$$

or

$$\sin\theta = n\sin\theta' \tag{2.12}$$

which is Snell's law.

With Snell's law, we can ask questions like: *A laser beam impinges on the top surface of a 2cm thick parallel glass (*$n = 1.5$*) plate at an angle of 35 degrees. How long is the actual path through the glass?*

We first find the angle in the glass,

$$\sin 35 = 1.5 \sin \theta_g \implies \theta_g = 22.48^\circ \tag{2.13}$$

which allows us to find the length via trigonometry,

$$L = \frac{2\text{cm}}{\cos 22.48^\circ} = 2.164\text{cm} \tag{2.14}$$

In the wave optics view, a refracted beam is the sum of many spherical wavelets at the interface.

2.6 Total internal reflection, partial reflection, transmission

Total internal reflection (TIR) occurs when light does not refract at all, when the angle of incidence is above some critical angle, and light entirely bounces off the interface.

Ray optics is insufficient in explaining partial reflection and transmission. Rays bend toward the normal when entering a medium of higher optical density, but when the relative index is very high, say on the order of 1000, taking the inverse sine (as required by Snell's law) is not possible. At this point, ray optics cannot predict the behaviour of the light, but by wave optics, we see that light reflects off the interface. The critical angle is the angle of incidence θ_1 for which

$$n_1 \sin \theta_1 = n_2 \implies \theta_1 = \arcsin \frac{n_2}{n_1} \tag{2.15}$$

The internet runs on total internal reflection, by a system of "planar" waveguides, a high-index dielectric material sandwiched between lower index dielectrics. Repeated cases of total internal reflection let the light bounce up and down the high-index material and in this way the light propagates. There is a tradeoff in the selection of the refractive index for optical fibers; for high-speed propagation, n should be low, but for TIR, n should be high.

Prisms can be structured so that TIR occurs at multiple interfaces to cause a net angle change. A retroreflector, for example, is a prism designed to return the light to the source, and a pentaprism causes incoming light to go off at a right angle.

2.7 Fermat's Principle in inhomogeneous media

In materials where n is not a constant, i.e. $n = n(r)$, the path in general is given by an integral rather than a linear relationship,

$$L = \int_s n(r) ds \tag{2.16}$$

EE 118: Introduction to Optical Engineering **Spring 2019**

Lecture 3: ABCD Matrices

Lecturer: Laura Waller *January 30* *Aditya Sengupta*

Last time, we covered refraction. Note that refraction from lower to higher n bends towards the normal, and higher to lower bends away from the normal. This is a quick way to figure out the relative indices of two materials.

3.1 Spatial frequency

Consider a ray of light with wavefronts transverse to it, at an angle θ from an optical axis (chosen as z). The spatial frequency of this wave is described by part of the electric field equation,

$$E(x) = \sin\left(\frac{2\pi}{\lambda_x}x\right) \tag{3.1}$$

The spacing between two consecutive wavefronts corresponds to a phase shift of 2π. The spatial period is given by $\lambda_x = \frac{1}{k_x}$. The angle of the ray (with the optical axis) relates the wavelength to the spatial frequency in the x direction,

$$\sin\theta = \lambda k_x \tag{3.2}$$

At small angles, we get

$$\theta \approx \lambda k_x \tag{3.3}$$

In words, the direction is proportional to the spatial frequency. This is true for monochromatic light at small angles. Note that this definition allows for a negative spatial frequency (with a negative angle); physically, this represents the direction being reversed.

3.2 ABCD matrices

In homogeneous media, rays move in straight lines. To define a ray, we therefore need an initial point and a direction; a ray is parameterized by an initial y and its angle of propagation, $\begin{bmatrix} y_0 \\ \theta \end{bmatrix}$. Conventionally, we represent the y point at which $x = 0$.

ABCD matrices give us a systems approach to modelling imaging systems. A 2D matrix takes in and returns a (y, θ) ray matrix:

$$g_{out} = Hg_{in} \tag{3.4}$$

$$\begin{bmatrix} y_{out} \\ \theta_{out} \end{bmatrix} = \begin{bmatrix} A & B \\ C & D \end{bmatrix} \begin{bmatrix} y_{in} \\ \theta_{in} \end{bmatrix} \tag{3.5}$$

Since this is a linear operator, it is easy to deal with, but not that widely applicable. It involves the paraxial approximation ($\sin\theta \approx \theta$, etc.)

For example, consider free-space propagation in air. A ray starts at some x_{in} with some θ_{in} initial angle, and after propagation over distance d, it becomes (x_{out}, θ_{out}). We can write an ABCD matrix to represent this. The angle does not change, but the x position does:

$$\begin{bmatrix} x_o \\ \theta_o \end{bmatrix} = \begin{bmatrix} 1 & d \\ 0 & 1 \end{bmatrix} \begin{bmatrix} x_i \\ \theta_i \end{bmatrix} \tag{3.6}$$

Conventionally, we call counterclockwise angles positive, movement to the right along the optical axis $+z$, and we call curvature to the left positive.

We can construct the ABCD matrix for Snell's law. We linearize the equation, to make it $n_1\theta_1 = n_2\theta_2$. Here, the x does not change at the interface, so the top row of the matrix would be 10, and Snell's law gives us the relation on the angle:

$$\begin{bmatrix} x_o \\ \theta_o \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & \frac{n_1}{n_2} \end{bmatrix} \begin{bmatrix} x_i \\ \theta_i \end{bmatrix} \tag{3.7}$$

Multi-element systems can cascade ABCD matrices because they are all linear operators. Matrix multiplication is not commutative, so operations have to be applied in reverse order: if a ray goes through matrices O_1, O_2, O_3 in that order, we write

$$\begin{bmatrix} x_f \\ \theta_f \end{bmatrix} = O_3 O_2 O_1 \begin{bmatrix} x_i \\ \theta_i \end{bmatrix} \tag{3.8}$$

Lenses can be complicated. We have to be aware of the limitations of ray optics, for example, if we cascade together multiple magnifying systems to get a net magnification on the order of 10^8, we would not see atoms even though these are on the order of the magnification, because wave optics does not allow this.

If the position and angle variations are small, we can think of each ABCD matrix element as a derivative:

$$\begin{bmatrix} x_o \\ \theta_o \end{bmatrix} = \begin{bmatrix} \frac{\partial x_o}{\partial x_i} & \frac{\partial x_o}{\partial \theta_i} \\ \frac{\partial \theta_o}{\partial x_i} & \\ \frac{\partial \theta_o}{\partial \theta_i n} & \end{bmatrix} \begin{bmatrix} x_i \\ \theta_i \end{bmatrix} \tag{3.9}$$

A and D are, respectively, the magnification and angular magnification. B and C do not have intuitive meanings but we will see their effects later.

3.3 Law of reflection

A ray departing some point P at some angle θ is reflected symmetrically at the same angle θ. The symmetric path POP', where O is the point of incidence and P' is a point reached by the reflected ray, is a reflection of Fermat's principle; POP' has a stationary value of the path length.

3.4 Curved reflecting surfaces

A common type of optical device is a reflective dish, such as a radio dish. These take in radiation from a far-away source (at infinity) and reflect each ray at precise angles to land on a detector. (This was done accidentally by skyscrapers made of metal with curved sides.) We can focus light with parabolic reflectors better than spherical ones. We can find the optimal shape. Say this is given by $s(x)$, and say that we want all the rays to focus at a point F, on the z axis (on an x-z plane) at a distance f from the origin.

For all the paths to focus on the same point, their optical path lengths should all be the same. Consider the distance between F and the origin, which is $2f$. We want this to be the same as some other path; this is given by $f - s$ for the distance from the same vertical as the focus to the surface, and by $\sqrt{x^2 + (f-s)^2}$ for the path from the arbitrary point on the reflector to the focus (by Pythagoras' Theorem).

$$\begin{aligned} 2f &= f - s + \sqrt{x^2 + (f-s)^2} && (3.10) \\ f + s &= \sqrt{x^2 + (f-s)^2} && (3.11) \\ x^2 &= f^2 + 2fs + s^2 - f^2 + 2fs - s^2 && (3.12) \\ x^2 &= 4fs && (3.13) \\ s &= \frac{x^2}{4f} && (3.14) \end{aligned}$$

which is the definition of a parabola.

Often, we make spherical lenses rather than parabolic ones, because they are much easier to manufacture, and because sometimes the point source approximation is not quite accurate. This induces a spherical aberration due to the deviation from the ideal curve. In curved mirrors, $f = \frac{R}{2}$. These have an ABCD matrix as follows:

$$\begin{bmatrix} 1 & 0 \\ -\frac{2}{R} & 1 \end{bmatrix} \tag{3.15}$$

Here, the x does not change at the point of reflection, and the angle changes as a function of the x as well as the incident θ. We can derive this as follows:

$$(s + R)^2 + x^2 = R^2 \tag{3.16}$$

(derivation skipped for time; see lecture slides)

3.5 Lenses

Lenses are the refractive analogue to mirrors, where they are like "unfolded" mirrors. Light comes to a focus on the other side, but geometrically, they are similar. We can see this by finding the $s(x)$ for perfect focusing in a lens, going from air to glass. We set it up the same way, but the difference here is optical path length varies with n. We get the ellipsoidal refractor,

$$\left(s - \frac{n}{n+1}f\right)^2 + \frac{n^2}{n^2-1}x^2 = \left(\frac{n}{n+1}f\right)^2 \tag{3.17}$$

To get this, we start from the optical path length on the axis, which is nf. Off the axis, this is $s + n\sqrt{x^2 + (f-s)^2}$. So we set these equal:

$$nf = s + n\sqrt{x^2 + (f-s)^2} \tag{3.18}$$

$$(n^2 - 1)s^2 - 2n(n-1)fs + n^2x^2 = 0 \tag{3.19}$$

This gives us the above equation.

We usually make spherical lenses instead because they are easier to make and in the paraxial approximation they are correct for all angles.

Spherical lenses are defined by their radius of curvature. All rays passing through a spherical lens come to a focal point F. We can find an ABCD matrix to represent this.

$$\begin{bmatrix} x_o \\ \theta_o \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ \frac{n_1-n_2}{Rn_2} & \frac{n_1}{n_2} \end{bmatrix} \begin{bmatrix} x_i \\ \theta_i \end{bmatrix} \tag{3.20}$$

There is no change in the x, as we expect. The $\frac{n_1}{n_2}$ term is similar to Snell's law, as we expect. If $R \to \infty$, this just becomes Snell's law. (Complete derivation in the slides).

3.6 Optical power

The paraxial refraction equation is

$$\theta_o = \frac{n_1}{n_2}\theta_i + \frac{n_1 - n_2}{Rn_2}h \tag{3.21}$$

When $\theta_i = 0$, where is the focus? The difference in the refractive index will determine whether it is before or after the center of curvature. We can find the focal length expression, $f = -h/\theta_o = \frac{n_2 R}{\Delta n}$. The optical power is defined based on this as $\phi \equiv \frac{n_2 - n_1}{R}$. This has units of inverse meters.

EE 118: Introduction to Optical Engineering **Spring 2019**

Lecture 4: ABCD matrices and imaging

Lecturer: Laura Waller *February 4* *Aditya Sengupta*

4.1 Administrative

Note: I was going over the previous lectures to do the homework and noticed I've got a few more errors than I thought. Sorry about this! I'll start proofreading before I compile and upload these, and if I get the chance I'll go back and fix the previous lectures.

Hecht uses a different ABCD matrix convention than the normal one; instead of $\begin{bmatrix} x \\ \theta \end{bmatrix}$, they use $\begin{bmatrix} \alpha \\ x \end{bmatrix}$, so it is necessary to be aware of which components are being operated on by which input variables.

4.2 Describing lenses with ABCD matrices

Consider a thick lens with thickness d (a block of glass between two curved surfaces), radii for the two curved surfaces R_1, R_2, and refractive index going from n_1 to n_2. We can create an ABCD matrix for the thick lens by multiplying together three ABCD matrices for, in the order left to right, the second curved surface, the propagation through the block, and the first curved surface:

$$\begin{bmatrix} x_{out} \\ \theta_{out} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ \frac{n_2 - n_1}{R_2 n_1} & \frac{n_2}{n_1} \end{bmatrix} \begin{bmatrix} 1 & d \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ \frac{n_1 - n_2}{R_1 n_2} & \frac{n_1}{n_2} \end{bmatrix} \begin{bmatrix} x_{in} \\ \theta_{in} \end{bmatrix} \tag{4.1}$$

This can be simplified by approximating d as going to zero. This is called the *thin-lens approximation.* We find the net ABCD matrix by multiplying together the two matrices, and we get

$$\begin{bmatrix} 1 & 0 \\ \frac{n_2 - n_1}{R_2 n_1} + \frac{n_2}{n_1}\left(\frac{n_1 - n_2}{R_1 n_2}\right) & \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ \frac{n_2 - n_1}{n_1}\left(\frac{1}{R_2} - \frac{1}{R_1}\right) & \end{bmatrix} \tag{4.2}$$

Intuitively, we see that the bigger the difference in refractive index, the greater the bending (the change in θ), which makes sense.

The lens maker's formula is a slightly different statement of this, which gives us the focal length of a lens:

$$\frac{1}{f} = \frac{n_2 - n_1}{n_1}\left(\frac{1}{R_1} - \frac{1}{R_2}\right) \tag{4.3}$$

This lets us simplify the ABCD matrix to

$$\begin{bmatrix} x_{out} \\ \theta_{out} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ -\frac{1}{f} & 1 \end{bmatrix} \begin{bmatrix} x_{in} \\ \theta_{in} \end{bmatrix} \tag{4.4}$$

A typical exam question might be to find the focal length of an optical system by cascading several ABCD matrices, and looking at the C component of the resulting matrix.

4.3 Converging and diverging lenses

A converging lens is one in which the focus is in front of the lens, and the lens brings rays of light that come in parallel to a focus. A diverging lens has the focus behind the lens, and the lens takes rays of light that come in parallel so that they are diverging to infinity, as if they come from a focal point.

Converging or positive lenses have $R_1 > 0$ and $R_2 < 0$ (biconvex). R_1 can be taken to infinity, which corresponds to one side being flat; these are *plano-convex* lenses. Plano-concave and biconcave lenses have the opposite signs on radii. An interesting case is a meniscus lens, which has $R_1 > 0, R_2 > 0$ and their relative magnitudes ddetermine whether it is converging or diverging.

4.4 Consecutive Lenses

Consider a thin biconvex lens next to a thin plano-convex lens. The ABCD matrix is

$$\begin{bmatrix} 1 & 0 \\ -1/f_2 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ -1/f_1 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ -1/f_1 - 1/f_2 & 1 \end{bmatrix} \tag{4.5}$$

This tells us that the inverse focal length is additive. The optical power (inverse focal length) of a thin lens is the sum of the optical powers of the two surfaces.

4.5 Determinants of ABCD Matrices

The determinant of an ABCD matrix has some physical meaning. To find out what this is, we can write the ABCD matrix for Snell's law and take its determinant, as an example:

$$\begin{bmatrix} 1 & 0 \\ 0 & \frac{n_1}{n_2} \end{bmatrix} \implies \begin{vmatrix} 1 & 0 \\ 0 & \frac{n_1}{n_2} \end{vmatrix} = \frac{n_1}{n_2} \tag{4.6}$$

For a lens, this is

$$\begin{bmatrix} 1 & 0 \\ -\frac{1}{f} & 1 \end{bmatrix} \implies \begin{vmatrix} 1 & 0 \\ -\frac{1}{f} & 1 \end{vmatrix} = 1 \tag{4.7}$$

This is also the refractive index change, $\frac{n_1}{n_1}$. Therefore the determinant is the ratio of initial to final n.

When $A = 0$, we can write the relationship in terms of linear relationships:

$$x_{out} = B\theta_{in} \tag{4.8}$$

$$\theta_{out} = Cx_{in} + D\theta_{in} \tag{4.9}$$

Physically, this means that only the input angle determines the output position. This makes physical sense for a converging lens, where all the rays meet at a point. The focusing is only a function of the input angle.

When $D = 0$, we get

$$x_{out} = Ax_{in} + B\theta_{in} \tag{4.10}$$

$$\theta_{out} = Cx_{in} \tag{4.11}$$

Physically, this means the input plane is the input focal plane. Both $A = 0$ and $D = 0$ can happen at the same time.

4.6 $2f$ and $4f$ systems

A lens can carry out a Fourier transform. Consider a biconvex lens between planes P_1 and P_2, with a distance f between the lens and the plane on either side. We write an ABCD matrix by combining the propagations through f with the lens matrix,

$$\begin{bmatrix} x_o \\ \theta_o \end{bmatrix} = \begin{bmatrix} 1 & f \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ -\frac{1}{f} & 1 \end{bmatrix} \begin{bmatrix} 1 & f \\ 0 & 1 \end{bmatrix} \begin{bmatrix} x_i \\ \theta_i \end{bmatrix} \tag{4.12}$$

$$\begin{bmatrix} x_o \\ \theta_o \end{bmatrix} = \begin{bmatrix} 0 & f \\ -\frac{1}{f} & 0 \end{bmatrix} \begin{bmatrix} x_i \\ \theta_i \end{bmatrix} \tag{4.13}$$

Here, we see that $x_o = f\theta_i$ and $\theta_o = -\frac{1}{f}x_{in}$. Angle and position switch places. This is related to a Fourier transform because the angle is related to a spatial frequency, $\theta_x \approx \lambda f_x$. Based on this, real space was switched with frequency space, which is what a Fourier transform does.

Consider a $4f$ system, which is similar to the $2f$ system but with two lenses with focal lengths f_1 and f_2. We multiply together the two $2f$ system matrices to get

$$\begin{bmatrix} 0 & f_2 \\ -\frac{1}{f_2} & 0 \end{bmatrix} \begin{bmatrix} 0 & f_1 \\ -\frac{1}{f_1} & 0 \end{bmatrix} = \begin{bmatrix} -\frac{f_2}{f_1} & 0 \\ 0 & -\frac{f_1}{f_2} \end{bmatrix} \tag{4.14}$$

If $f_1 = f_2$, we get the negative of the identity matrix. That is, we get back the original image, but inverted. If $f_1 \neq f_2$, we get a magnified image (with magnification given by the A parameter). Microscopes have a low f_1 and a high f_2 so that their magnification is high; telescopes have a high f_1 and a low f_2 so that they can scale down the night sky to the scale of a lens. The D parameter gives angular magnification.

Consider a $4f$ system with the distances between the planes and the closer lenses changed; instead of f_1, there is a distance s_o to the first lens from the input plane, and instead of f_2, there is a distance s_i from the second lens to the output plane. The changed ABCD matrix can be written as the combination of three propagations between two lenses. We get

$$\begin{bmatrix} -\frac{f_2}{f_1} & f_1 + f_2 - \frac{f_1 s_i}{f_2} - \frac{f_2 s_o}{f_1} \\ 0 & -\frac{f_1}{f_2} \end{bmatrix} \tag{4.15}$$

4.7 Interpreting ABCD Matrices, contd.

Consider the case where $C = 0$. Here, the rays must be collimated and parallel, because the angle does not vary with x. If $B = 0$, then $x_o = Ax_i$. This represents a measure of how in focus an image is.

4.8 Ideal imaging requirements

In an ideal case, every point object is represneted as a point source of light that goes into an imaging system, and that imaging system yields a point image. This is rarely the case, which is why software such as Zemax exists. Anything that deviates from this ideal is called an *aberration.*

Consider a biconvex lens with lengths s_o and s_i separating the lens from the two planes. The ABCD matrix is

$$\begin{bmatrix} 1 & s_i \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ -\frac{1}{f_1} & 1 \end{bmatrix} \begin{bmatrix} 1 & s_o \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} 1 - \frac{1}{f} & s_o - \frac{s_i s_o}{f} - s_i \\ -\frac{1}{f} & -\frac{s_o}{f} + 1 \end{bmatrix} \tag{4.16}$$

For what choices of s_o and s_i is this lens in focus, i.e. when is $B = 0$?

$$s_o + s_i = \frac{s_i s_o}{f} \implies \frac{1}{s_i} + \frac{1}{s_o} = \frac{1}{f} \tag{4.17}$$

This gives us a condition that allows us to change the focal distance. If an object is far away from a lens, the image will come into focus close to the lens; as the object is moved closer and closer, the image is formed farther and farther away.

With this constraint, we can find the A component which gives us the magnification:

$$M_{transverse} = -\frac{s_i}{s_o} \tag{4.18}$$

At $s_i = s_o = 2f$, we have a magnification of 1.

A pinhole camera always has all of its light in focus, but they block out most of the light while doing so. Lenses can be thought of as a series of pinholes that shifts an image due to refraction.

4.9 Ray tracing for thin lenses

There are three rules to be followed for tracing rays through a thin lens: a ray through the center of a lens is not diverted, a ray through the focus emerges parallel to the optical axis, and a ray parallel to the optical axis passes through the focus.

EE 118: Introduction to Optical Engineering **Spring 2019**

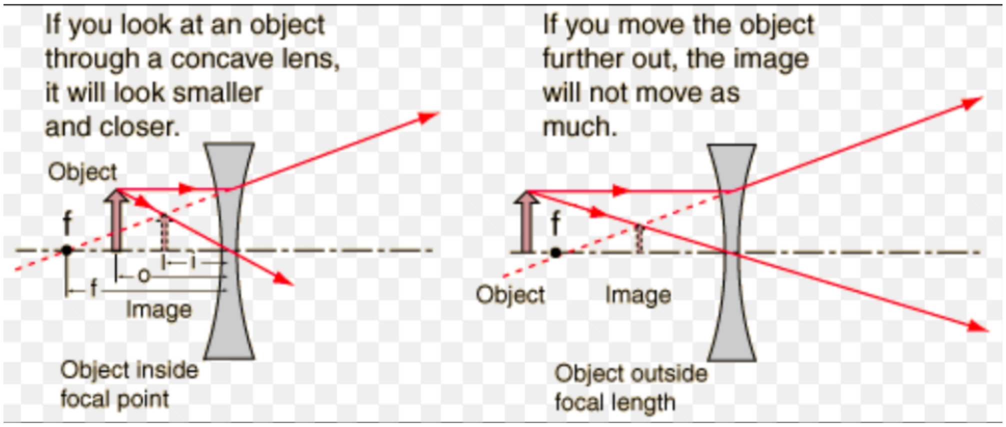
Lecture 5: Imaging Systems

Lecturer: Laura Waller *6 February* *Aditya Sengupta*

Last time, we saw the condition for successful imaging, $\frac{1}{s_o} + \frac{1}{s_i} = \frac{1}{f}$. This can also be stated as $x_i x_o = f^2$, where x_i and x_o are the distances from the object or the image to the focal plane.

When $s_i < 0$, the rays are diverging and an image is formed behind the lens. This is a virtual image; if a screen were placed at the position we calculate, behind the lens, we would not see anything. That occurs when $s_o < f$ and $f > 0$, or when $s_o > f$ and $f < 0$. An example of a virtual image is the image we see in a mirror.

There are four possible cases in which we would see real or virtual images. If the lens is positive, then if the object is outside the focal length, we get a real and inverted image, and if the object is inside the focal length, we get a virtual and erect image. If the lens is negative, then if the object is outside the focal length, (moved on too fast but here's a picture explaining it)


If you look at an object through a concave lens, it will look smaller and closer.
Object
f
o
i
f
Image
Object inside focal point
If you move the object further out, the image will not move as much.
f
Object
Image
Object outside focal length

5.1 Multi-element imaging systems

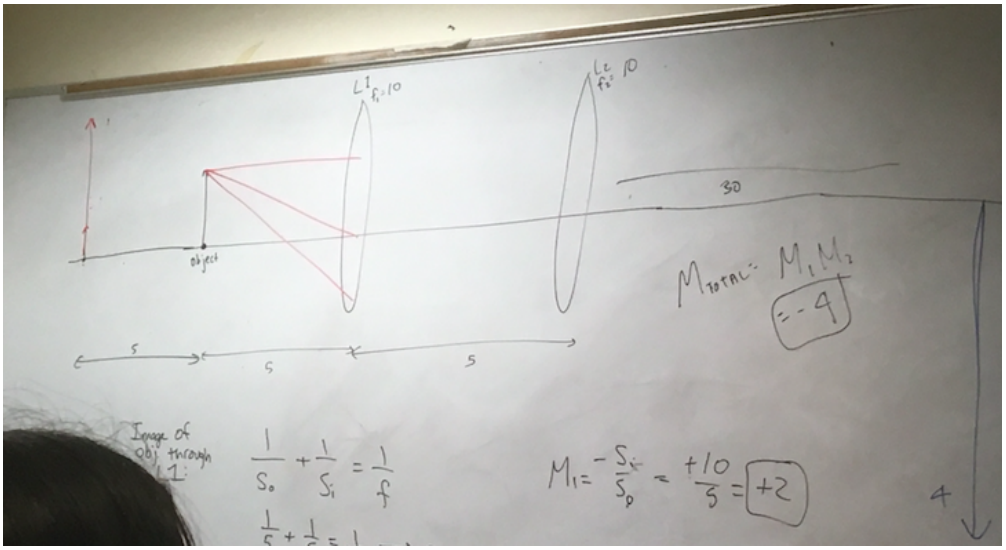
Sometimes, it helps to trace rays through a system rather than find out the effect of the system through an ABCD matrix, which is a long process. Consider two convex lenses with $f = 10$ spaced a distance 5 apart, and consider an object a distance 5 to their left. For the first propagation, we get an image formed at s_i:

$$\frac{1}{s_i} + \frac{1}{s_o} = \frac{1}{f} \implies \frac{1}{s_i} = \frac{1}{10} - \frac{1}{5} \implies s_i = -10 \tag{5.1}$$

which we see from the rays as well. The magnification is +2, confirming that the image is magnified and erect. The virtual image is then the object for the second lens, and it is beyond the focus of the second lens (it is at a distance of 15), so we get a real inverted image.

$$\frac{1}{15} + \frac{1}{s_i} = \frac{1}{10} \implies s_i = 30 \tag{5.2}$$

The magnification is $\frac{-30}{15} = -2$. In total, $M_{total} = M_1 M_2 = -4$.



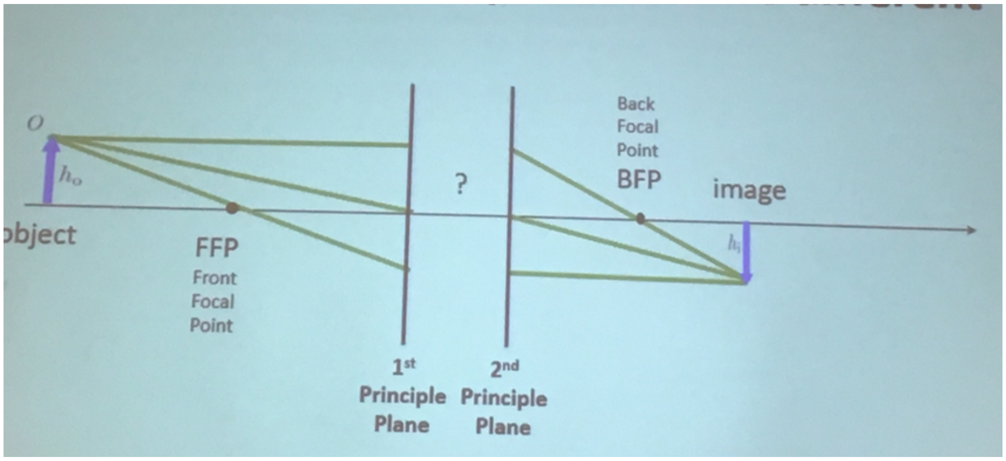
5.2 Front and back focal planes

The back focal plane is the z position where a ray entering parallel to the optical axis passes through it, and the front focal plane is the z position where a ray that exits parallel to the optical axis initially crosses the axis.

The front and back focal distances do not have to be the same. These define the first and second principal planes. They have new ray tracing rules. These are:

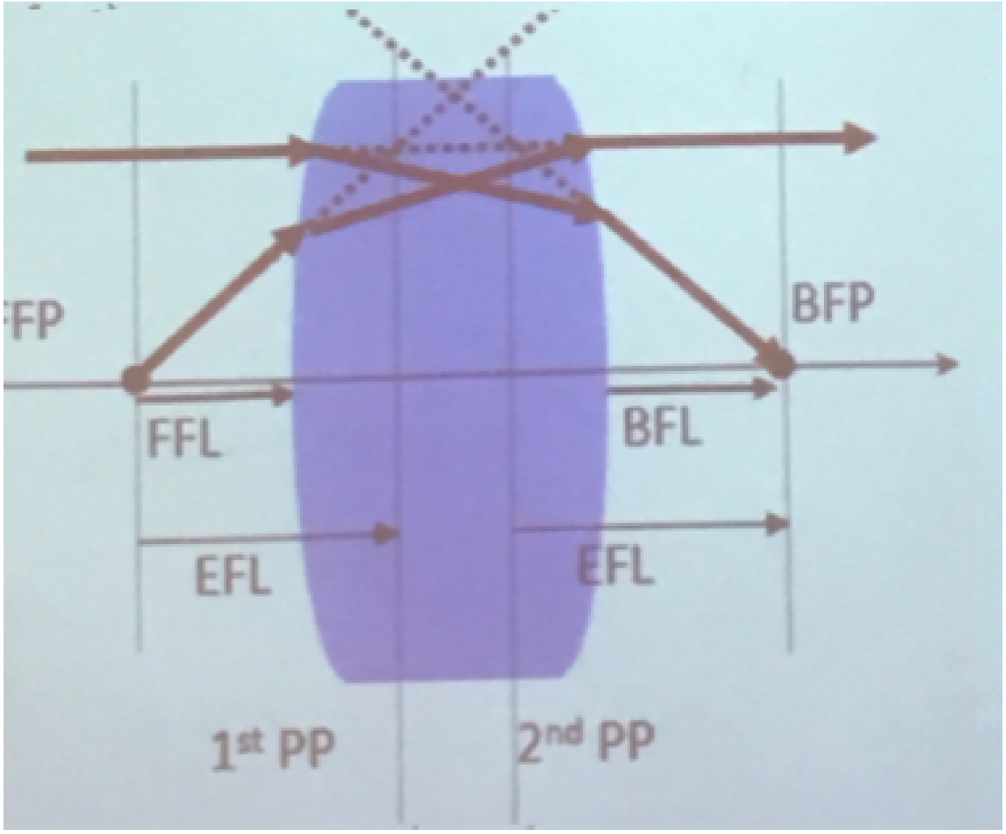
1. trace a ray from infinity through object to the second PP, then bend to go through BFP and meet the image tip.
2. trace a ray from the object tip through the FFP, then bend at the first PP to go to infinity.
3. the intersection of the traced rays is the image point.
4. (this is the weird part) the ray from the object through the intersection of the first PP with the optical ais should emerge at the intersection of the second PP and the optical axis, then go through the image point.

The principal planes tell you where to cut the system to treat it as if it were a thin lens.



Introducing the principal planes allows us to keep the imaging condition and magnification relations the same as for a thin lens.

We can combine focal lengths and principal planes. The distance between two principal planes is the thickness of the lens, in addition to the thin-lens approximation on either side.



This is an important diagram for exams.

5.3 Photography

5.3.1 Stops

A field stop limits the field of view. The sensor in a camera cuts out unnecessary parts of the field. An aperture stop limits the quantity of light that is collected. Both can be due to the physical limits of the lens or sensor, or can be controlled with a diaphragm. Stops can be used to kill strong light sources that are not desired, or to limit stray reflections (such as in a microscope).

Field stops affect the field of view, and is directly determined by the focal length. If we zoom in on a camera, the field of view is limited, and there is not much angular spread. It is convenient to talk about a field of view in terms of angles.

5.3.2 Field of view

For a fixed sensor size, decreasing the focal length increases the field of view. They are related by

$$FOV \propto \arctan \frac{1}{f} \tag{5.3}$$

(there was a more specific formula on the slides up for about 5 seconds)

Lenses can be constructed to provide wide angles, such as fisheye lenses and gigapixel cameras.

5.3.3 Pupils

The effective size of a stop may be larger or smaller due to refraction, so we define pupils in order to see how it will affect the object. The entrance pupil is defined as the image of the aperture stop as seen from an axial point on the object through the elements preceding the stop. The exit pupil is defined similarly on the other side.

The entrance and exit pupils determine the cone of light that enters or leaves the system. Some rays from an object hit the stop and are therefore blocked, meaning that the image can be partially impeded. In photography, the aperture stop controls exposure; a smaller aperture stop is less light throughput. Aperture stops also control the depth of field and resolution.

5.3.4 Chief and marginal rays

These represent the center and edge cases for the bundle of rays that pass through the system. The chief ray is the ray from the maximum object height that passes through the center of the stop, and the marginal ray is the most extreme ray angle allowed by the optical system coming from the object base (the highest angle that doesn't get clipped by the aperture stop).

To find the limiting aperture stop and entrance pupil, we image each component leftwards and find which is the most limiting one.

EE 118: Introduction to Optical Engineering **Spring 2019**

Lecture 6: Imaging Systems II

Lecturer: Laura Waller *11 February* *Aditya Sengupta*

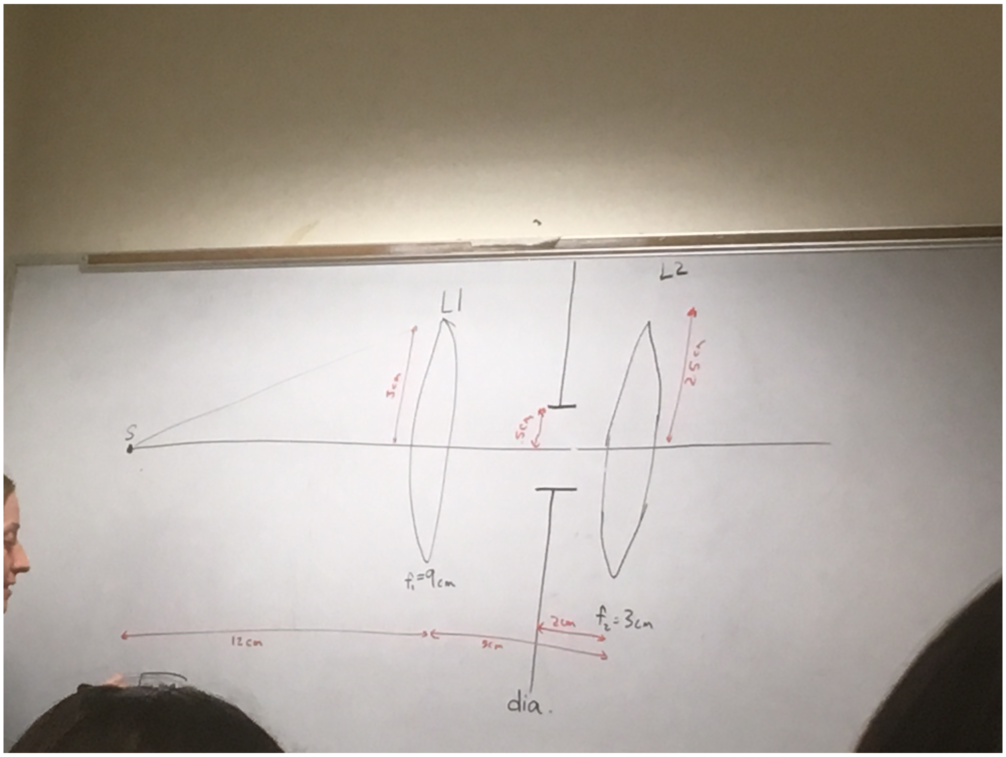
6.1 Recap

Quick reminders: suppose we have a multi-element imaging system. The aperture stop is the most limiting stop. To get the entrance pupil, we start from the aperture stop and image it to the left. To get the exit pupil, do the same to the right. It is possible to have many apertures, but only one aperture stop.

The field stop limits the angular acceptance of chief rays, where the chief ray is the one that goes through the center of the aperture stop. If the object is made larger, then the ray just above the chief ray would be clipped by the field stop, and so it would not make it to the aperture stop. The field stop defines the field of view. Colloquially, this is sometimes described as the height of the tallest object that makes it through the system, but this is dependent on object position; a better description would be based on the angle determining the view. Formally, the field of view is the angle between the pair of chief rays that just make it through the edges of the field stop. Entrance and exit windows are analogous to entrance and exit pupils for the field stop instead of for the aperture stop. The pupils are images of the aperture stop, and the windows are images of the field stop.

6.2 Example

Consider the optical system shown below,



To find the stop, we image each object backwards to the observing point. For lens 1, we get

$$\tan\theta_1 = \frac{3cm}{12cm} \implies \theta_1 = 14^\circ \tag{6.1}$$

For the diaphragm, we image backwards using $x_0 x_i = f^2$. $f = 9$ and $x = -6$, so $x_i = -13.5cm$. This is 4.5cm to the right of L1. The magnification is $-13.5/-9 = 1.5$, so the diaphragm image is magnified and to the right of the aperture. The angle subtended by this image is given by its height above and below the optical axis (0.75cm) over the distance of 16.5cm from the observer. Therefore

$$\tan\theta_2 = \frac{0.75}{16.5} \implies \theta_2 = 2.6^\circ \tag{6.2}$$

This is the most limiting object so far. Now, we need to image L2 all the way back through the entrance. To image it back through, we ignore the aperture, because if the aperture is blocking rays, it will end up being the more limiting factor. Put another way, if the aperture were significant, then we would automatically know that $\theta_3 > \theta_2$. By imaging through L1 in the same way as for the diaphragm, we get $\theta_3 = 7.6^\circ$. Therefore the diaphragm is the aperture stop.

Now, we can find the entrance and exit pupils. The entrance pupil is the image of the aperture stop through L1, which we found in order to get the angle θ_2. The exit pupil is the image through L2. It turns out that the exit pupil is before the entrance pupil. This potentially makes optical systems more compact.

6.3 Definitions

6.3.1 Numerical aperture

The numerical aperture describes the range of angles subtended by the imaging system from an axial object. It is given by $NA = n \sin \theta$, where θ is the half-angle subtended, from the optical axis to the extent of the system, and n is the refractive index of the surrounding medium.

For a given lens diameter, NA is related to focal length. Dry objective lenses are limited to NA = 0.95. Higher NAs are accessible in different immersion media. The aperture stop in photography, which is determined by the f number which is inversely related to the NA, controls the size of the aperture.

6.3.2 Depth of focus

For a given object distance, the depth of focus is the range of image distances for which the image is in focus. This is the effect that blurs out anything unwanted

6.3.3 Depth of field

For a given image distance, the depth of field is the range of object distances for which the object is in focus. Our eyes use the depth of field as a clue as to the sizes of objects and their distances from the camera/point of view (the depth).

Depth of field is inversely proportional to numerical aperture squared.

(Activity)

EE 118: Introduction to Optical Engineering **Spring 2019**

Lecture 7: Pupils, DoF equations, vignetting

Lecturer: Laura Waller *13 February* *Aditya Sengupta*

(guest lecturer who didn't name himself)

7.1 Pupils

7.1.1 Recap

An aperture stop is an object that limits the cone angle of light accepted from some object. The chief ray passes through the center of the aperture stop. If we extend the object space version of the chief ray (without any effect from the lens), we get the entrance pupil. The aperture stop and entrance pupil are images of each other. To find the height, we find the marginal ray, which starts from the optical axis and passes through the extent of the aperture stop. The entrance pupil's highest point is located where the object-space version of the marginal ray would end up. Image formation is determined entirely by the chief and marginal rays.

The imaging equation $\frac{1}{t_A} + \frac{1}{t_E} = \frac{1}{f}$ can be used to locate the aperture stop. We use t_A, the distance from the lens to the aperture, as the image distance, so the convention for the image distance has to be used. For a lens to the left of the aperture stop, t_A is positive and t_E is negative.

7.1.2 Circle of confusion

The circle of confusion is the largest diameter circular spot we are willing to tolerate in an imaging system. If something is defocused by a distance d, i.e. the image is observed at $s_o + d$, then by similar triangles, we can find the radius of the circle of confusion.

7.1.3 Calculating depth of field

$$DoF \approx \frac{2D_s^2 NC}{f^2} \tag{7.1}$$

This can be derived using similar triangles; we consider a far and near point for the object distance and corresponding near and far points that determine the depth of focus.

The sensor size is also a significant factor in determining the depth of field. A DSLR with a larger sensor than an iPhone 6 has a 20x smaller depth of field (which helps to focus the subject of a picture), even with the same field of view and f number.

The circular shape of the circle of confusion comes from the aperture. The circle of confusion can be made sharp, and a blurry effect is sometimes desired. The shape and quality of an out-of-focus blur is referred to in photography as "bokeh". Changing the shape of the aperture changes the shape of confusion.

7.2 Vignetting

Vignetting refers to when a lens is small enough that it cuts off rays. It refers to a field-varying numerical aperture. We can derive the conditions on lens size to avoid vignetting; this is the same angular size as the aperture stop, as we can see by propagating the marginal ray through the system (assuming aperture stop to the left of the lens) and using similar triangles. Say this height is $|y_a|$. We can do the same with the chief ray to get a height $|y_b|$. Then, the condition on element height is $C = |y_a| + |y_b|$.

Vignetting causes space-varying bokeh. (See images from slides for examples)

A system is vignetted if any rays from off-axis object points that are aimed to pass through the entrance pupil instead get cut off by some surface in the system that is not the aperture stop. On-axis vignetting is not possible; if a lens were shrunk sufficiently to cause on-axis vignetting, it would become the new aperture stop.

7.3 Common imaging systems

7.3.1 Single-lens camera

An example of this is the human eye. The entrance pupil and the aperture stop are the same, and so it is easy to trace a ray of light through. The iris is the aperture stop, and the cornea does most of the focusing. In nature, chambered eyes are common (unsure of the definition). Reflection-based eyes are also possible. Compound eyes are made up of many little imaging systems.

7.3.2 Eye defects and their correction

There are two major types of defects of the eye: myopia (nearsightedness) and hypermetropia (farsightedness). Nick (I think that's his name) had an accident but it's all okay now.

7.4 Telecentricity

Telecentricity means that the entrance pupil is at infinity. If this is the case, the system is telecentric in object space. If the exit pupil is at infinity, the system is telecentric in image space. If both are at infinity, the system is doubly telecentric, which is a $4f$ system. This consists of (for example) two positive lenses of focal lengths f_1 and f_2. The magnification is $M = -\frac{f_2}{f_1}$.

$4f$ systems are afocal, meaning that they have no focal length. This means a collimated beam comes out collimated as well. It has images, without having a focal length (don't think about it too hard). The real answer is to try and locate the principal planes; the two rays that intersect to form the principal plane are parallel.

In telecentric systems, magnification does not depend on defocusing. The size of an object is invariant after imaging. This makes telecentric systems useful for measurement.

EE 118: Introduction to Optical Engineering **Spring 2019**

Lecture 8: Aberrations I

Lecturer: Laura Waller *20 February* *Aditya Sengupta*

8.1 What are aberrations?

Aberrations can affect all optical systems, including imaging systems. The goal of an imaging system is to perfectly map a plane to a plane; all points in a plane in object space map to points on a plane in image space. But, as showed by Maxwell, this is impossible. For example, a spherical lens cannot bring light to a perfect focus.

Aberration theory is the study of the systematic ways in which a system of lenses corrupts an image, and how to fix it. We cannot make a perfect image, but we can come close.

8.2 Chromatic aberrations

Chromatic aberration is caused by the dependence of refractive index on the wavelength of light. Every wavelength sees a slightly different optical system. This causes light to disperse, causing effects like rainbows.

Chromatic aberration can be quantified in terms of the Abbe number, which is a scalar measure of dispersion. It describes the curvature of the wavelength vs. refractive index plot. We measure this by taking three points along the curve (conventionally, F - 486nm, D - 589nm, C - 656nm) and taking a second derivative by taking two measures of slope, then a slope between them.

$$V = \frac{1/F_D}{1/f_F - 1/f_C} = \frac{n_D - 1}{n_F - n_C} \tag{8.1}$$

A high Abbe number corresponds to low dispersion, and a low Abbe number corresponds to high dispersion.

Usually, lenses are rated to somewhere in the middle of the wavelength range, around green light. This means when light deviates from this, chromatic aberration becomes significant. For example, when there is red and blue light not coming perfectly into focus, they may combine to form purple fringe effects.

Axial chromatic aberration (or longitudinal colour) is the specific name for this aberration due to different focal length for different colours. This can be seen in the lensmaker's formula, where we let $n = n(\lambda)$,

$$\frac{1}{f} = \frac{n(\lambda) - n_0}{n_0}\left(\frac{1}{R_1} - \frac{1}{R_2}\right) \tag{8.2}$$

The change in focal length is

$$\Delta f(\lambda) = \frac{n(\lambda_0) - n(\lambda)}{n(\lambda) - 1} f_G \tag{8.3}$$

In terms of optical power, $\phi_F - \phi_C \equiv \Delta\phi = \frac{\phi_D}{V}$.

8.3 Correcting chromatic aberrations

A lens that corrects chromatic aberration is called an achromat. It consists of placing a lens with high dispersion (say, flint glass) with negative curvature on the left-facing surface and a flat right surface. This is next to a converging lens made of a material with lower dispersion (such as crown glass).

To have red and blue focus to the same plane, given the Abbe numbers of the two materials, we can choose these lenses so that they focus well. We know that optical power adds,

$$\phi_{tot} = \phi_1 + \phi_2 \tag{8.4}$$

and dispersion,

$$\frac{\phi_1}{V_1} + \frac{\phi_2}{V_2} = 0 \tag{8.5}$$

We can solve for ϕ_1 and ϕ_2.

Suppose we wanted to correct for three wavelengths. We would be able to find a solution for ϕ_1 through ϕ_3, but this begins to stretch the thin lens approximation. More wavelengths would definitely break the approximation. It is possible to make corrections for any number of wavelengths, but this would induce many other aberrations.

In general, we can design for no axial colour in other ways. We can design a system such that all light comes to the same back focal plane in ways other than adding more optical elements. (Missed out on the method, or if there was one explained)

Due to the different focal lengths for different wavelengths of light, chromatic aberration can be used as a depth sensor. A lens can be designed to intentionally have bad chromatic effects, so that it is sensitive to changes in depth.

8.4 Axial and lateral colour

The other type of chromatic aberration is lateral or transverse. This is a change in the size or height of an image with colour. Axial colour affects the chief ray, and lateral colour affects the marginal ray. If a system has axial colour, then the choice of the stop position determines the amount of lateral colour, and if a stop is at the single lens there is no lateral colour. An achromat can correct both.

What is the optimal separation between two lenses of focal lengths f_1 and f_2 to minimize lateral chromatic aberration?

We can set up the equation for the power and the aberration. To find the power, we need to find the composite focal length, which we do via the ABCD matrix representation,

$$\begin{bmatrix} 1 & 0 \\ -\frac{1}{f_2} & 1 \end{bmatrix} \begin{bmatrix} 1 & L \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ -\frac{1}{f_1} & 1 \end{bmatrix} \implies \phi(L) = \frac{1}{f_1} + \frac{1}{f_2} - \frac{L}{f_1 f_2} \tag{8.6}$$

Using the lens maker's formula, we can rewrite this,

$$\frac{1}{f_1} = C_1(n-1), \frac{1}{f_2} = C_2(n-1) \tag{8.7}$$

We eventually get (derivation done fully in the slides)

$$L = \frac{1}{2}(f_1 + f_2) \tag{8.8}$$

8.5 Anomalous dispersion

Anomalous dispersion bends light the other way, i.e. having red bend more than blue. This is usually done by causing dispersion off a diffraction grating. This is the method that modern spectrometers use (over prisms) because it is much more space efficient.

8.6 Monochromatic aberrations, quantifying aberrations

Next time: monochromatic aberrations. These either cause blurring or distortion. A distorted image is still sharp, but with positioning errors. This can be fixed digitally. Blurring refers to when the image is no longer sharp, and this cannot be fixed digitally.

These aberrations come from the higher-order terms in the small angle approximation. So far we have only worked with first-order theory, but now we will study third-order theory, in which the approximation is $\sin\theta \approx \theta - \frac{\theta^3}{3}$. If the third-order aberrations can be corrected, the image is likely to be fine.

Aberrations can be characterized by a spot diagram, the ray mapping from a single point object, and by a point spread function, a map of intensity distribution imaged from a single point object across the image plane. Aberrations can be reduced by using a higher refractive index, decreasing the angle of incidence, or creating aspheric surface.

8.7 Third-order geometric optics

It is possible to write equations governing the third-order approximation to geometric optics. For example, refraction at a curved surface has the following equation in third order,

$$\frac{n_1}{s_0} + \frac{n_2}{s_i} = \frac{n_2 - n_1}{R} + \alpha h^2 \tag{8.9}$$

where α is a coefficient depending on the radii of curvature and refractive indices (the important part is the h^2 dependence.)

EE 118: Introduction to Optical Engineering **Spring 2019**

Lecture 9: Aberrations II (Monochromatic)

Lecturer: Laura Waller *25 February* *Aditya Sengupta*

9.1 Logistics

Problem Set 4 is due Wednesday, and Exam 1 is on Monday March 4. One hand-written double sided cheat sheet is allowed.

9.2 Monochromatic Aberrations

Last time, we introduced aberrations. An ideal $4f$ system, as an example, maps a point to a point. Introducing aberrations means points are no longer ideally mapped to points, because rays are not perfectly collimated. Today, we will study aberrations from the lens (haha) of third-order theory. These are called Seidel aberrations. These can be characterized by spot diagrams or by point-spread functions, both defined last time.

There are five primary Seidel aberrations:

1. Spherical aberrations
2. Coma
3. Astigmatism
4. Field curvature
5. Distortion

9.3 Spherical aberrations

This arises from lenses not being paraxial near the edges of lenses. There is an aplanatic point, with no spherical aberrations, only at $r = R/n_{lens}$. Spherical aberrations can be reduced using an asphere, or by using multiple elements. Spherical aberrations are space-invariant (like time-invariance of LTI systems), meaning that the PSF is the same no matter where the object is. The aberration comes from the lens, so it does not depend on space. This also means that spherical aberration is present both at the center and at the edges of the image.

9.4 Coma

This causes a comet or teardrop shape in a point image (it has an apparent tail). It is like spherical aberration for off-axis sources, due to their having different effective focal lengths and magnifications. This is a function of wavelength.

9.5 Astigmatism

This is characterized by an asymmetric PSF. If light is put through a system starting from a horizontal vs. a vertical plane, and the horizontal light is in focus but the vertical light is out of focus (for example). Non-axisymmetric systems are astigmatic even for on-axis points.

9.6 Field curvature

The problem that field curvature is associated with is that a lens maps curved surfaces to curved surfaces. In the small-angle approximation, this is fine, but at short focal lengths, the image seems curved. This means the image is out of focus at the edges, that is, some of the field of view is out of focus.

Is it possible to choose two thin lenses to cancel field curvature? To do this, we use the equation

$$\Delta x = \frac{y_i^2}{2} \sum_{j=1}^{m} m \frac{1}{n_j f_j} \tag{9.1}$$

$$\Delta x = 0 \implies n_1 f_1 + n_2 f_2 = 0 \tag{9.2}$$

This is called Petzval's condition. It is true if there is negligible spacing between the two lenses. In case there is, consider the case $f_1 = f_2$ and $n_1 = n_2$. Then, the effective focal length is

$$\frac{1}{f_1} + \frac{1}{f_2} - \frac{d}{f_1 f_2} \implies f = \frac{f_1^2}{d} \tag{9.3}$$

9.7 Distortion

This is a varying amount of magnification with the distance from the optical axis. This has two sub-cases: barrel distortion (decreasing magnification with increasing distance, such as in a fisheye/wide-angle lens) and pincushion distortion (increasing magnification with increasing distance, such as in a telephoto lens).

Distortion and coma are related. An orthoscopic system of thin lenses is one in which the optical center and the center of the aperture stop are coincident. Stopping down the aperture in an orthoscopic system reduces coma but not distortion. Symmetric lens systems about the stop will have no distortion, meaning that both can be managed. However, symmetric systems require that $M = 1$.

If half of a lens is covered, then the image becomes fainter, but no part of the image disappears. All the rays converge on the same point, so cutting off part of the lens only removes some of the rays, rather than changing the geometry of the image.

9.8 Mathematical description of aberrations

Seidel aberrations are specifically arrived at by truncating the Taylor series of $\sin(\theta)$. An alternate way of describing aberrations is the Zernike polynomials, which are made for cylindrical imaging systems.

Aberrations connect rays to waves. Aberrations are wavefront distortions, but they can also be described as ray bending. The ideal output is considered a spherical wavefront reference; a spherical wave in the exit pupil plane is needed to get a point in the object plane. If we consider an imaging system to be a black box, we can measure aberrations as the difference between the expected and actual outputs.

In paraxial optics, everything lies in one plane. However, with aberrations, skew rays that have some component along the third dimension are possible. Rays still originate from points along one axis in the object plane, but they can be taken anywhere in the 2D pupil plane. Not every aberration will do this, but some of them can, so they need to be allowed in our mathematical model. Skew rays are still rotationally symmetric, so polar coordinates are an easier way to describe this system. We need to consider one dimension in the object plane, say y, and two in the pupil plane, (r, ϕ). The normalized coordinates are (h, ρ, ϕ), where h is the y height divided by the max object height, and $\rho = r *$ pupil radius. We expect both of these coordinates to be between 0 and 1.

An aberration can be considered a transverse error vector in the image plane. We can describe it by components (ϵ_x, ϵ_y), which is a measure of how far off the paraxial ray and the actual ray are from one another.

The direction of a ray is always normal to the wavefront surface, so ray errors are proportional to the derivative of the wavefront:

$$\epsilon_x = \frac{1}{n'\theta'_a}\frac{\partial W}{\partial \rho_x} \tag{9.4}$$

and the same for y.

We can describe ray errors using deviations from the sphere. The wavefront aberration function is the optical path distance between the ideal and actual wavefront, $W(h, \rho, \phi) = W_r(h, \rho, \phi) - W_p(h, \rho, \phi)$.

Wavefront aberrations can take on restricted forms because they have to be rotationally symmetric. With this restriction, the wavefront aberration must be of the form

$$W(h, \rho, \phi) = \sum_{i,j,k} W_{ijk} h^i \rho^j \cos^k(\phi) \tag{9.5}$$

where $i + j$ must be even.

In an expanded form, this is

$$W(h, \rho, \phi) = W_{000} + W_{020}\rho^2 + W_{111}h\rho\cos\phi + W_{200}h^2 + W_{040}\rho^4 + W_{131}h\rho^2(\rho\cos\phi) + W_{220}h^2\rho^2 + W_{222}h^2\rho^2\cos^2\phi + W_3$$

The indices that sum to 4 correspond to the Seidel aberrations. These are 4th order terms because ray aberrations are derivatives of the wavefront aberrations.

The W_{111} term is tilt, and the W_{000} and W_{200} terms are both piston terms. $W_{020}\rho^2$ is defocus, as it depends on a location in the pupil but is space-invariant. Tilt can be analyzed based on the previous deviation formula. Say $\phi = 0$, so that

$$W_{111}h\rho\cos\phi = W_{111}h\rho \tag{9.7}$$

The deviation in the x direction is then

$$\epsilon_x = \frac{1}{NA}\frac{\partial W}{\partial \rho_x} = 0 \tag{9.8}$$

and in the y,

$$\epsilon_y = \frac{1}{NA'}\frac{\partial W}{\partial \rho_y} = W_{111}h \tag{9.9}$$

as we expect for an angle equal to zero.

Similarly, we can analyze the defocus term.

$$W_{020}\rho^2 = W_{020}(\rho_x^2 + \rho_y^2) \tag{9.10}$$

$$\epsilon_x = \frac{1}{NA} 2W_{020}\rho_x \tag{9.11}$$

$$\epsilon_y = \frac{1}{NA'} 2W_{020}\rho_y \tag{9.12}$$

Therefore there are linear deviations in the x and y components, as expected for a defocus term.

EE 118: Introduction to Optical Engineering **Spring 2019**

Lecture 10: Aberrations III

Lecturer: Laura Waller *27 February* *Aditya Sengupta*

Last time, we covered the two ways to describe aberrations, in terms of rays and wavefronts. Ray errors are described using the deviation from the ideal sphere.

Piston: constant wavefront error. W_{000}, W_{200} are piston, W_{111} is tilt, and W_{020} is defocus. Today, we will cover the fourth-order wavefront aberrations, which are the third-order Seidel aberrations.

10.1 Third-order Seidel aberrations

The five Seidel aberrations are spherical, coma, astigmatism, field curvature, and distortion. The sixth term in the aberration expansion is just an additional piston term.

$$W_4(h, \rho, \phi) = W_{040}\rho^4 + W_{131}h\rho^2(\rho\cos\phi) + W_{220}h^2\rho^2 + W_{222}h^2\rho^2\cos^2\phi + W_{311}h^3\rho\cos\phi + W_{400}h^4 \quad (10.1)$$

040 - spherical, 220 - field curvature, 222 - astigmatism.

The $h\rho^3\cos\phi$ is a tilt term multiplied by a defocus, which is coma (a change in magnification based on where light hits the pupil). The $h^2\rho^2$ term is field curvature, the $h^2\rho^2\cos^2\phi$ term is astigmatism because there is dependence on angle magnitude, and the $h^3\rho\cos\phi$ term is distortion.

Consider the spherical aberration term in Cartesian coordinates.

$$W = W_{040}\rho^4 = W_{040}\left(\rho_x^4 + 2\rho_x^2\rho_y^2 + \rho_y^4\right) \quad (10.2)$$

$$\epsilon_y = \frac{4W_{040}}{NA}\left(\rho_y^3 + \rho_x^2\rho_y\right), \epsilon_x = \frac{4W_{040}}{NA}\left(\rho_x^3 + \rho_y^2\rho_x\right) \quad (10.3)$$

Similarly, we can look at the coma term.

$$W = W_{131}h\rho^3\cos\phi = W_{131}h\rho^2\rho_y = W_{131}\rho^2\Delta m \quad (10.4)$$

We can therefore find the error term,

$$\epsilon_y = \frac{W_{131}}{NA}h(3\rho_y^2 + \rho_x^2) \quad (10.5)$$

and the same expression in x (with the indices exchanged).

For field curvature, we have $W = W_{220}h^2(\rho_x^2 + \rho_y^2)$. The error term is

$$\epsilon_y = \frac{2W_{220}}{NA} h^2 \rho_y \tag{10.6}$$

We analyze this similarly to a defocus term. (The Buralli notes cover these derivations in detail; I'm not sure I really followed everything.)

The last one we wil look at here is astigmatism,

$$W = W_{222} h^2 \rho^2 \cos^2 \phi = W_{222} h^2 \rho_y^2 \tag{10.7}$$

We see there is no x deviation, and the y deviation is $\frac{2W_{222}}{NA} h^2 \rho_y$.

10.2 Review

Topics you should know about for the exam:

1. Aberrations: both mathematical and intuitive. E.g. drawing spot diagrams and ϵ vs ρ.
2. Ray transfer matrices
3. Imaging conditions
4. Finding principal planes, aperture stop, entrance/exit pupils, numerical aperture
5. Tracing marginal and chief rays, back and front focal planes, effective focal length
6. Circle of confusion, defocus, vignetting
7. Refraction and waves
8. Approximations (paraxial)

10.2.1 Problems

The Coma aberration term is given by $W_{131} h \rho^3 \cos \phi$. Intuitively, coma changes magnification based on where light hits the pupil (varying quadratically). Coma is zero at the center of the field of view in a rotationally symmetric imaging system, which we can see by substituting in $\rho = 0$ to the above expression. If $W_{131} = -\frac{\lambda}{4}$, with $\lambda = 500\text{nm}$, then the transverse ray errors can be derived. As in the Homework 4 problem, assume an image distance of 20, a focal length of 10 and an f-number of 2. Then

$$\theta_0' = \frac{D}{2 \cdot 20} = -\frac{10/2}{40} = -\frac{1}{8} \tag{10.8}$$

$$\epsilon_x = \frac{1}{-\frac{1}{8}} \cdot -\frac{1}{8} h \frac{\partial \rho_y \rho^2}{\partial \rho_x} = h \rho_y \cdot 2\rho_x \tag{10.9}$$

$$\epsilon_y = h(3\rho_y^2 + \rho_x^2) \tag{10.10}$$

For a fan of rays, the direction not being plotted is zero, so ϵ_x can just be considered identically zero for the plot. (See slides for the derivation.)

(Seeing Nemo question from the slides)

We can make the ABCD matrix for the system,

$$\begin{bmatrix} x \\ \theta \end{bmatrix} = \begin{bmatrix} 1 & s \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ -\frac{1-n}{-R} & 1 \end{bmatrix} \begin{bmatrix} 1 & R/n \\ 0 & 1 \end{bmatrix} \begin{bmatrix} x_i \\ n\theta_i \end{bmatrix} \tag{10.11}$$

$$\begin{bmatrix} x \\ \theta \end{bmatrix} = \begin{bmatrix} \frac{1}{n} & \frac{R+s}{n} \\ \frac{1-n}{R} & 1 + \frac{s(1-n)}{R} \end{bmatrix} \begin{bmatrix} x_i \\ n\theta_i \end{bmatrix} \tag{10.12}$$

The term that relates the input angle to the output position is $\frac{R+s}{n}$, which should be made zero by the standard imaging condition. The condition is therefore $s = -R$. We get

$$\begin{bmatrix} \frac{1}{n} & 0 \\ \frac{1-n}{R} & n \end{bmatrix}$$

We get a magnification of $M = \frac{x_o}{x_i} = n$, so the image is erect, and the image is formed where the actual object is present. In reverse, the system becomes

$$\begin{bmatrix} x_i \\ \theta_i \end{bmatrix} = \begin{bmatrix} 1 & \frac{R}{n} \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ \frac{1-n}{R} & 1 \end{bmatrix} \begin{bmatrix} 1 & s \\ 0 & 1 \end{bmatrix} \begin{bmatrix} x_o \\ n\theta_o \end{bmatrix} \tag{10.13}$$

$$\begin{bmatrix} x_i \\ \theta_i \end{bmatrix} = \begin{bmatrix} 1 - \frac{s(n-1)}{R} & \frac{s'}{n} + s - \frac{ss'(n-1)}{nR} \end{bmatrix} \tag{10.14}$$

$$\frac{1-n}{R} \quad 1 - \frac{s'(n-1)}{nR} \begin{bmatrix} x_o \\ \theta_o \end{bmatrix} \tag{10.15}$$

EE 118: Introduction to Optical Engineering **Spring 2019**

Lecture 11: Quiz 1

Lecturer: Laura Waller *4 March* *Aditya Sengupta*

No content here, this is just being included for completeness.

EE 118: Introduction to Optical Engineering **Spring 2019**

Lecture 12: Wave Optics

Lecturer: Laura Waller *6 March* *Aditya Sengupta*

12.1 Photons

Light in its particle form can be considered to be made up of photons. Photons are stable, chargeless, massless elementary particles travelling at the speed of light. They have energy $E = h\nu$. They can exert force; when an EM wave impinges on a material, it interacts with the charges in it and exerts a force. Photons have momentum $\vec{p} = \hbar\vec{k} = \frac{h}{2\pi}\vec{k}$, and a frequency which is related to the wavelength by $c = \lambda\nu$. This gives us an energy-wavelength relationship, $E = \frac{hc}{\lambda}$.

We can get a sense of scale pf photons via these equations. For example, consider the question: *How many photons of visible light are generated by a 100W light bulb that is left on for 1 hour?* The energy per photon is

$$E_{photon} = \frac{hc}{\lambda} = \frac{6.63 \times 10^{-34} \times 3 \times 10^8}{500 \times 10^{-7}} \approx 4 \times 10^{-19}\text{J} \tag{12.1}$$

and the total energy is 100 W times 3600s. Therefore the number of photons is

$$\#\text{ photons} = \frac{100 \times 3600}{4 \times 10^{-19} = 9 \times 10^{23}\text{ photons}} \tag{12.2}$$

12.2 EM Waves

Recall that the wave equation allows us to check if a function $\psi(x,t)$ is of the required form of a wave,

$$\frac{\partial^2\psi}{\partial x^2} = \frac{1}{v^2}\frac{\partial^2\psi}{\partial t^2} \tag{12.3}$$

In general, any function of the form $\psi(x,t) = f(x - vt)$ is a valid wave. For example, $\psi(x,t) = e^{-a(x-vt)^2}$ is a valid wave; it is a bell-shaped curve travelling in the x direction.

Electromagnetic waves are a bit more specific. They have associated E and B fields, that are orthogonal and whose cross product describes the direction of propagation. They are specific in that they follow Maxwell's equations, which are as follows:

$$\nabla \times \vec{E} = -\frac{\partial \vec{B}}{\partial t} \tag{12.4}$$

$$\nabla \times \vec{H} = \frac{\partial \vec{D}}{\partial t} + \vec{J} \tag{12.5}$$

$$\nabla \vec{D} = \rho \tag{12.6}$$

$$\nabla \cdot \vec{B} = 0 \tag{12.7}$$

In words, and in order, these are: the curl of the electric field is the negative of the magnetic flux density; the curl of the magnetic field is the time derivative of the electric displacement plus the electric current density; the divergence of the electric (displacement) field is equal to the electric charge density, and the divergence of the magnetic field is zero (there is no magnetic charge).

More simply, they say that electric fields are generated by both electric charges and time-varying magnetic fields, and vice versa for magnetic fields.

We can get the magnetic field from the electric field with Maxwell's equations. Consider a plane-wave electric field, $\vec{E} = \hat{x}E_0 \cos(kz - \omega t)$. To find the magnetic field, we take the curl of the electric field and divide by $\frac{-1}{\mu_0}$ to get the time-derivative of $\vec{H}$. (By definition $\vec{B} = \mu_0 \vec{H}$). We get

$$\frac{\partial \vec{H}}{\partial t} = -\frac{1}{\mu_0}\left(\hat{x}(0-0) - \hat{y}(0 - \frac{\partial E_x}{\partial z} + \hat{z}\left(0 - \frac{\partial E_x}{\partial y}\right))\right) \tag{12.8}$$

$$\vec{H} = \int \frac{\partial \vec{H}}{\partial t} dt = \frac{kE_0}{\mu_0 \omega}\cos(kz - \omega t)\hat{y} \tag{12.9}$$

$$\vec{H} = \hat{y}\left(\frac{k}{\mu_0 \omega}\right)\left|\vec{E}\right| \tag{12.10}$$

The Poynting vector describes energy flow along the direction of propagation (it is parallel to propagation), $\vec{S} = c^2 \varepsilon_0 \vec{E} \times \vec{B}$. The magnitude of the Poynting vector represents the power per unit area crossing some surface.

We can compute the Poynting vector for the previously determined E and B fields, as an example. Let $\vec{E} = E_0 \cos(k \cdot r - \omega t)$ (to make the direction generic) and $\vec{B} = B_0 \cos(k \cdot r - \omega t)$. Then the Poynting vector magnitude, first in general in terms of the electric field, is

$$|S| = c^2 \varepsilon_0 |\vec{E}||\vec{B}| = \vec{B} = \frac{\vec{k}}{\omega} \times \vec{E} \implies |B| = \frac{k}{\omega}|E| \tag{12.11}$$

$$|S| = c\varepsilon_0 |E|^2 \tag{12.12}$$

For the plane wave, this becomes

$$|S| = c\varepsilon_0 E_0^2 \cos^2(kr - \omega t) \tag{12.13}$$

A detector cannot instantaneously measure this energy, so we end up measuring a time-averaged intensity $I = \frac{1}{T}\int_t^{t+T} |S| dt'$,

$$I = \frac{c\varepsilon_0}{2} E_0^2 \tag{12.14}$$

For a linear isotropic dielectric, the electric displacement field is related to the electric field by $\vec{D} = \varepsilon_0 \vec{E} + \vec{P} = \varepsilon_0(1+\chi)\vec{E}$. It turns out that $n = \sqrt{1+\chi}$. For an anisotropic dielectric, χ is a 3x3 matrix, which physically means that the refractive index depends on the polarization.

The situation becomes even weirder in nonlinear optics, which has $P = \varepsilon_0(\chi_1 + \chi_2 E + \chi_3 E_2 + \dots)E$.

12.3 Electric and magnetic forces

These are determined in terms of fields by the Lorentz force equation,

$$\vec{F} = q(\vec{E} + \vec{v} \times \vec{B}) \tag{12.15}$$

(There are several examples in the slides that I can't replicate without looking up a number of diagrams)

EE 118: Introduction to Optical Engineering **Spring 2019**

Lecture 13: Polarization

Lecturer: Laura Waller *11 March* *Aditya Sengupta*

13.1 Finishing the previous lecture

13.1.1 Changing magnetic flux

Faraday's experiment, consisting of a magnetic core being powered by a battery (with a switch to close and open the circuit) and an ammeter to detect current thus produced, showed that a changing magnetic flux induces a current.

The electrical continuity law says that electric current and charge density are conserved,

$$\vec{\boldsymbol{\nabla}} \cdot \vec{J} = -\frac{\partial \rho}{\partial t} \tag{13.1}$$

13.1.2 Plane waves are solutions to the wave equation

To find when a plane wave solves the wave equation, we can substitute in $\vec{E} = \hat{x}E_0 \cos(kz - \omega t)$ satisfies the wave equation and derive a condition from that,

$$\vec{\boldsymbol{\nabla}}^2 E - \mu_0 \varepsilon_0 \frac{\partial^2 E}{\partial t^2} = 0 \tag{13.2}$$

$$\hat{x}E \frac{\partial^2}{\partial z^2} \cos(kz - \omega t) - \mu_0 \varepsilon_0 \frac{\partial^2}{\partial t^2} \left(E_0 \hat{x} \cos(kz - \omega t)\right) = 0 \tag{13.3}$$

$$-k^2 \cos(kz - \omega t) = \mu_0 \varepsilon_0 (-\omega^2) \cos(kz - \omega t) \tag{13.4}$$

Therefore the condition is $k^2 = \mu_0 \varepsilon_0 \omega^2$. This is called the dispersion relation. It relates spatial and tempoeral frequency, which are proportional to one another with proportionality constant $\mu_0 \varepsilon_0 = \frac{1}{c^2}$.

13.1.3 Phase velocity

The phase velocity describes the speed at which the profile of the wave moves. It is given by $\frac{\omega}{k}$ for the aforementioned plane wave.

13.2 Polarization

Transverse waves oscillate in the xy plane while travelling along z. Polarized light has a common direction of oscillation. We can describe polarization in terms of the ratio of intensity of light that is polarized to the total intensity of light. Longitudinal waves cannot be polarized. An EM wave can be polarized by selective absorption. For example, a Polaroid absorbs more light in one polarization than the other. Other ways include reflection and scattering. Birefringence (double refraction) in crystalline materials can also cause polarization.

Polarization describes the orientation of oscillation. Oscillations can be resolved into orthogonal components, with some phase offset and relative amplitudes.

$$\vec{E} = \hat{x}E_{0x}\cos(kz - \omega t) + \hat{y}E_{0y}\cos(kz - \omega t + \varepsilon) \tag{13.5}$$

There are four major types of polarization: linear, circular, elliptical, and unpolarized. We want to find a specific form for the $\vec{E}$ field equation for each of these cases.

13.3 Linear polarization

This can also be called plane polarization. The electric field vector traces out a plane, and the x and y components are in phase. This gives us the constraint $\varepsilon = n\pi$, so the form of the electric field reduces to

$$\vec{E}_{linear} = [\hat{x}E_{0x} + \hat{y}E_{0y}]\cos(kz - \omega t) \tag{13.6}$$

We represent the electric field by a double-sided vector (a line with two arrows) that represent the extreme values of the electric field. In general, if we want an EM wave to be linearly polarized and inclined at θ relative to the horizontal axis, then $\vec{E} = (\hat{x}E_0\cos\theta + \hat{y}E_0\sin\theta)\cos(kz - \omega t)$.

13.4 Dichroism

Natural dichroism occurs in crystals like tourmaline. They absorb one polarization strongly along a principal axis. This is not perfect, though, because it also absorbs some of the other polarization. It is strongly wavelength dependent, making strange colours. We usually use plastic polarizers, that have engineered dichroism. They consist of a large number of metallic wires in parallel, which absorb the components of the electric field that hit it edge-on. Those that come in vertically polarized either just go through the wire or cause transverse oscillations in the wire that cause it to re-radiate. Photographers use polarizers to limit stray reflections.

13.5 Polarization on reflection

When an unpolarized EM wave is reflected by a smooth surface, it may become completely or partially polarized. The amount of polarization depends on the angle of incidence; if the angle is zero, there is no polarization. There is some particular angle such that the wave becomes completely polarized. For other angles, there is partial polarization. In general, waves oriented parallel to the surface will reflect more. This is the concept behind polarized sunglasses. They block horizontally polarized light to reduce glare from horizontal surfaces.

I'VE DECIDED WE SHOULD BE "COOLER" THAN WE ARE.
WE'RE NOT COOL?
SURE WE'RE COOL. BUT WE'RE NOT AS COOL AS WE COULD BE.
COOL PEOPLE WEAR DARK GLASSES!
IT'S COOL TO BUMP INTO THINGS?
YOU DON'T MOVE, YOU JUST HANG AROUND.
© 1986 Universal Press Syndicate
3-28
WATTERSON

Polarization can be used to redirect or split light. We can also image polarization vectors. Our eyes cannot see polarization, but we can design cameras to do this; in front of every pixel on a CCD, four polarizers with different directions can be placed, to spatially map out the polarization of an image.

13.6 Partial polarization

Partial polarization can be thought of as a superposition of specific amounts of randomly polarized and fully polarized light. This can be seen by comparing pictures of the sky with or without a polarizer; the former is noticeably more blue. Light from the sun is unpolarized, and hits particles in the atmosphere in a preferential direction. They oscillate more in the perpendicular plane. Therefore, as observed from the side, light is polarized. Blue light scatters more, therefore it is less polarized; since multiple scattering destroys polarization, blue light makes it through the polarizer.

13.7 Polarizer angles

Crossed polarizers block all light; if light is linearly polarized in one direction, then none of it makes it through a second polarizer that is oriented 90 degrees relative to the first one. In general, the intensity that makes it through two polarizers oriented at an angle θ relative to one another is

$$I_{out} = I_{in} \cos^2 \theta \tag{13.7}$$

This is Malus' Law. We can use this to predict the intensity of light that passes through one linear polarizer, by integrating:

$$I_{av} = \frac{1}{2\pi} \int_0^{2\pi} I_o \cos^2 \theta d\theta = \frac{I_o}{2} \tag{13.8}$$

Interestingly, if we have three polarizers, with both consecutive pairs oriented at 45 degrees relative to one another, the intensity at the end is not zero even though there is still an angle change of 90 degrees overall (whereas if there were just two at 90 degrees, no light would get through).We can use Malus' Law to find the intensity at the end.

$$I_2 = I_1 \cos^2 45^\circ = I_0 \cos^4 45^\circ \times \frac{1}{2} = \frac{I_0}{8} \tag{13.9}$$

The factor of $\frac{1}{2}$ comes in from going from unpolarized to polarized light.

EE 118: Introduction to Optical Engineering **Spring 2019**

Lecture 14: Polarization II

Lecturer: Laura Waller *13 March* *Aditya Sengupta*

14.1 Circular and Elliptical Polarization

Last time, we saw that polarization describes the direction of oscillation. Linear polarization is characterized by $\epsilon = 0$, where ϵ is the phase offset between the x and y electric fields, and circular polarization is characterized by $\epsilon = \frac{\pi}{2}$. Circularly polarized light has a vector tip tracing out a circle in $E_x - E_y$ space. Returning to the generic equation for light, we want to try and make it into that of a circle:

$$\vec{E} = \hat{x}E_{0x}\cos(kz - \omega t) + \hat{y}E_{0y}\cos(kz - \omega t + \epsilon) \tag{14.1}$$

To make this into circular polarization, we set $E_{0x} = E_{0y}$ and $\epsilon = \frac{\pi}{2}$ in order to generate one sine component and one cosine component.

Circular polarization decomposes into two perpendicular EM waves of equal amplitude, but with a 90 degree phase difference. To check the direction of circular polarization, freeze space (set $z = 0$) and substitute in values for time, e.g. $t = 0$ and $t = 0.1$. From this, we can see whether the field would move clockwise or counterclockwise.

Right circularly polarized light has the form

$$\vec{E} = \hat{x}E_0\cos(kz - \omega t) + \hat{y}E_0\sin(kz - \omega t) \tag{14.2}$$

and left circularly polarized light has the form

$$\vec{E} = \hat{x}E_0\cos(kz - \omega t) - \hat{y}E_0\sin(kz - \omega t) \tag{14.3}$$

If left and right circularly polarized light are added, we get

$$\vec{E}_L + \vec{E}_R = 2E_0\left(\hat{x}\cos(kz - \omega t)\right) \tag{14.4}$$

which is linearly polarized.

For some general linear combination of circular polarizations, e.g. where $E_{0x} \neq E_{0y}$, we get elliptical polarization. In terms of ϵ, E_{0y}, and E_{0x}, we can show that the equation we get from adding two electric fields is that of an ellipse. We start with

$$E_x = E_{0x}\cos(kz - \omega t), E_y = E_{0y}\cos(kz - \omega t + \epsilon) \tag{14.5}$$

From the latter, we get

$$\frac{E_y}{E_{0y}} = \cos(kz - \omega t)\cos(\epsilon) - \sin(kz - \omega t)\sin(\epsilon) \tag{14.6}$$

and from the former,

$$\frac{E_x}{E_{0x}} = \cos(kz - \omega t) \tag{14.7}$$

Therefore, we get

$$\frac{E_y}{E_{0y}} = \frac{E_x}{E_{0x}}\cos(\epsilon) - \sqrt{1 - \cos^2(kz - \omega t)}\sin(\epsilon) \tag{14.8}$$

$$\frac{E_y}{E_{0y}} = \frac{E_x}{E_{0x}}\cos(\epsilon) - \sqrt{1 - \frac{E_x^2}{E_{0x}^2}}\sin(\epsilon) \tag{14.9}$$

Squaring both sides after rearranging, we get

$$\left(\frac{E_y}{E_{0y}} - \frac{E_x}{E_{0x}\cos\epsilon}\right)^2 = 1 - \frac{E_x^2}{E_{0x}^2}\sin^2\epsilon \tag{14.10}$$

or

$$\left(\frac{E_y}{E_{0y}}\right)^2 + \left(\frac{E_x}{E_{0x}}\right)^2 - 2\left(\frac{E_y}{E_{0y}}\right)\left(\frac{E_x}{E_{0x}}\right)\cos\epsilon = \sin^2\epsilon \tag{14.11}$$

In the limiting case $E_{0x} = E_{0y}$ and $\epsilon = \frac{\pi}{2}$, we get the unit circle.

Other types of polarization are a subset of elliptical polarization. For varying values of ϵ, we get different eccentricities of the ellipse traced out by the electric field vector.

We can analyze equations to find their state of polarization. For example, $\vec{E} = \hat{x}E_0\cos(kz - \omega t) - \hat{y}2E_0\sin(kz - \omega t)$ has elliptical polarization because the components are out of phase with different magnitudes; the polarization is left-handed, which can be seen by setting $z = 0$ and advancing time slightly from $t = 0$. $\vec{E} = \hat{x}E_0\cos(kz - \omega t) + \hat{y}2E_0\sin(\omega t - kz + \pi/2)$ has its components in phase (because $\sin(x + \pi/2) = \cos(x)$) so this describes linear polarization with angle $\theta = \arctan 2$.

14.2 Birefringence

In birefringent materials, each polarization component (x, y) incurs a different phase delay due to an anisotropic molecular structure. This causes different polarizations to see different refractive indices.

Birefringence relates to anisotropy in that the binding forces between atoms in a crystal are stronger or weaker based on the location in the material. We describe birefringent materials in terms of their ordinary and extraordinary axes. Uniaxial crystals have one refractive index for light polarized along the optical axis, and another for light polarized perpendicularly to it. Along the optical axis, the refraction is *extraordinary*; perpendicular to it, the refraction is *ordinary.*

Birefringent materials show different colours when placed between two crossed polarizers. Not all light is blocked in this case, because the birefringent material in the middle changes the polarization to somewhere in between both of the polarizers' axes. Wave pates use birefringent materials to change the state of polarization. Half-wave plates cause a phase delay of π, which is given by the following equation:

$$\Delta\varphi = \frac{2\pi}{\lambda} d|n_o - n_e| = \frac{2\pi}{\lambda} OPL \tag{14.12}$$

In summary, the optical axis of a birefringent crystal is the direction which suffers no birefringence. A quarter-wave plate has $d|n_o - n_e| = \frac{(2m+1)\lambda_0}{2}$ and converts between linear and circular polarization, and a half-wave plate has $d|n_o - n_e| = \frac{(4m+1)\lambda_0}{4}$ and it rotates linearly polarized light from θ to $-\theta$. Liquid crystal displays use crossed polarizers; the electrically polar nature of LC molecules means the lowest energy configuration state is with electric dipoles aligned with the applied electric field. This causes a change in the refractive indices for different incident polarizations. Therefore, a liquid crystal cell between crossed polarizers modulates amplitude with applied voltage per pixel.

EE 118: Introduction to Optical Engineering **Spring 2019**

Lecture 15: Fresnel Equations

Lecturer: Laura Waller *18 March* *Aditya Sengupta*

15.1 Example of predicting polarization

Randomly polarized input light is sent through a linear polarizer at 45 degrees to the $x-$axis, followed by two quarter-wave plates with refractive indices n_f and n_s. After the first linear polaizer, the intensity becomes $I_{in}/2$, and the polarization state is given by

$$\hat{e} = \sqrt{\frac{I_{in}}{2}}\hat{x}\cos(kz - \omega t) + \hat{y}\cos\Big(kz - \omega t + \frac{\pi}{2}\Big) \tag{15.1}$$

Then, through the waveplates, we get

$$I_{out} = \frac{I_{in}}{2} \tag{15.2}$$

$$\hat{e}_{out} = \sqrt{\frac{I_{in}}{2}}(\hat{x} - \hat{y})\cos(kz - \omega t) \tag{15.3}$$

Therefore, we get linearly polarized light with the given direction and intensity.

15.2 Polarization Changes on Reflection and Refraction

Polarization changes on reflection and refraction. The relative amounts of reflected and refracted light are described by the Fresnel equations. At interfaces, we describe wave polarizations in terms of the transverse electric field (which is perpendicular to the plane) and the transverse magnetic field (parallel to the plane). If the polarization is somewhere in between, we describe it in terms of a TE component and a TM component.

The TM case is the dual of the TE case; see the slides for the precise variable mapping.

15.3 Dipole model for polarization-dependent reflection

It is easier to re-radiate light in particular directions, based on the alignment of the reflected ray with a dipole axis. If the two directions are close, then the reflected beam is weaker. With no reflection, we know that $\theta_t + \theta_r = 90$. The θ_r at which this happens is called Brewster's angle, and it satisfies the relation

$$n_i \sin\theta_p = n_t \sin\theta_t \tag{15.4}$$

$$\tan\theta_p = \frac{n_t}{n_i} \tag{15.5}$$

15.4 Fresnel Equations

These give the ratios of the reflected and transmitted electric fields in the TE and TM directions.

$$r_\perp = \left(\frac{E_{0r}}{E_{0i}}\right)_\perp = \frac{n_i \cos\theta_i - n_t \cos\theta_t}{n_i \cos\theta_i + n_t \cos\theta_t} \tag{15.6}$$

$$r_\parallel = \frac{n_t \cos\theta_i - n_i \cos\theta_t}{n_i \cos\theta_t + n_t \cos\theta_i} \tag{15.7}$$

$$t_\perp = \left(\frac{E_{0t}}{E_{0i}}\right)_\perp = \frac{2n_i \cos\theta_i}{n_i \cos\theta_i + n_t \cos\theta_t} \tag{15.8}$$

$$t_\parallel = \frac{2n_i \cos\theta_i}{n_i \cos\theta_t + n_t \cos\theta_i} \tag{15.9}$$

Intuitively, they describe how waves change at boundaries. Previously, we talked about how $\vec{E} = E_0 e^{i(kz-\omega t)}$ describes a plane wave in a homogeneous medium. At interfaces, there will be three components: incident, reflected, and transmitted. For continuity, their intensities all have to add together.

From this perspective, Snell's law is a consequence of the tangential electric field being continuous across the boundary (which is an important rule). Therefore, to solve problems, we look at boundary conditions at an interface. This is how the Fresnel equations are derived.

We can plot the reflection coefficients over a range of incidence angles, which is done in the slides.

Using Snell's law, we can simplify the Fresnel equation significantly. For example, $r_\perp = \frac{\sin(\theta_t - \theta_i)}{\sin(\theta_t + \theta_i)}$.

The ratio of energies can also be given by the reflection and transmission coefficients. These energy ratios are denoted R and T. Energy conservation is equivalent to the statement that $R + T = 1$, or explicitly

$$1 = \frac{E_{0r}^2}{E_{0i}^2} + \frac{E_{0t}^2}{E_{0i}^2}\frac{n_t \cos\theta_t}{n_i \cos\theta_i} \tag{15.10}$$

Energy being conserved across a boundary means both the parallel and perpendicular components must conserve energy: $R_\perp + T_\perp = 1$ and $R_\parallel + T_\parallel = 1$.

If waves with the same polarizations add, then they interfere and we get $I = |E_1 + E_2|^2$. If they have opposite polarization, there is no interference.

EE 118: Introduction to Optical Engineering **Spring 2019**

Lecture 16: Optics and Image Quality in the Human Eye

Lecturer: Laura Waller *20 March* *Aditya Sengupta*

(Guest lecturer: Austin Roorda)

The human eye is an optical system with a number of components. Light first hits the first surface of the cornea, which has a radius of curvature of 7.7mm and an index of refraction of 1.376. Therefore, its power comes out to be +48.83 D. At the back surface of the cornea, the index reduces slightly and the radius of curvature becomes 6.8mm, to get a power of -5.88D. The total power of the cornea is about 43 D.

The pupil governs image quality, amount of light, and depth of focus. The pupil is perfectly located to maximize the field of view of the eye. If you trace light through the pupil, which turns out to be the field stop, it just hits the edges of the retina.

Humans can operate in a really wide range of illuminances in natural environment, from about 10^{-6} to 10^8 cd/m^2. So an optical system that can operate at all of these ranges would have to be really well engineered.

After these two, we hit the crystalline lens, which has a gradient index of refraction. This increases the overall power as it causes the light to curve as it passes through. For a homogeneous lens to have the same power, it would have to have a higher refractive index than the peak index in the gradient.

To change the eye's focal length, a sphincter muscle (which shrinks when activated) is used. The relaxed eye is under tension at the equator from the ciliary body. This keeps the surfaces flat enough so that for a typical eye, distant objects are easily visible. In the accommodated eye, the ciliary muscle constricts and relaxes the tension on the equator of the lens. This increases the surface curvature and power of the lens. This increases the eye's power from about 22 to about 32 dioptres.

With age, the eye tends to harden and lose some of its accommodation power. The human eye starts around a range of 10 dioptres, and over time this accommodation reduces.

The retina samples the image by millions of rods and cones. The fovea (maximum concentration of cones) is 5 degrees from the optical axis, which does something. (Sorry, I'm not quite getting all the details here - his standard slides are available online.)

The human eye perceives the sizes of objects based on visual angles. For example, the full moon occupies half the visual angle of your fingernail.

EE 118: Introduction to Optical Engineering **Spring 2019**

Lecture 17: Interference

Lecturer: Laura Waller *1 April* *Aditya Sengupta*

Light + light = darkness. If you combine two waves that are π radians out of phase, then the crests and troughs match up and destructively interfere to create nothing. Noise-cancelling headphones work on this principle; they listen to ambient noise and create an opposite wave to destructively interfere with it and create silence. However, sound and light differ in that we cannot measure the actual electric field of the light; we can only measure time-averaged interference. Instead of measuring E directly, we measure I.

Suppose there is a point P at which two electric fields combine, $\vec{E}_P = \vec{E}_1 + \vec{E}_2$. The intensity there is

$$I = \varepsilon_0 c \left\langle \vec{E}_P^2 \right\rangle = \varepsilon_0 c \left\langle \vec{E}_1 \cdot \vec{E}_1 + \vec{E}_2 \cdot \vec{E}_2 + 2\vec{E}_1 \cdot \vec{E}_2 \right\rangle = \varepsilon_0 c \left\langle I_1 + I_2 + I_{12} \right\rangle \tag{17.1}$$

The I_{12} term represents interference. It is the dot product of the electric fields. For the usual form of electric field,

$$\vec{E}_1 \cdot \vec{E}_2 = \vec{E}_{01}\vec{E}_{02} \cos(ks_1 - \omega t + \phi_1)\cos(ks_2 - \omega t + \phi_2) \tag{17.2}$$

We simplify this by introducing constant phases $\alpha = ks_1 + \phi_1$, $\beta = ks_2 + \phi_2$, and using a trig identity,

$$I_{12} = \varepsilon_0 c \vec{E}_{01}\vec{E}_{02} \left\langle \cos(\beta - \alpha) \right\rangle = \varepsilon_0 c \vec{E}_{01}\vec{E}_{02} \left\langle \cos \Delta\phi \right\rangle \tag{17.3}$$

Overall, if we time-average the cosine expressions in I_1 and I_2, we get

$$I = I_1 + I_2 + I_{12} = \frac{1}{2}\varepsilon_0 c \left(E_{01}^2 + E_{02}^2\right) + 2\sqrt{I_1 I_2} \left\langle \cos \Delta\phi \right\rangle \tag{17.4}$$

To maximize the contrast, we want I_1 and I_2 close together. When they are equal, say $I_1 = I_2 = I_0$, I maxes out at $4I_0$ and is at minimum 0.

Unidirectional destructive interference, where waves combine destructively in one direction and constructively in the other, is possible.

17.1 Interferometry

To characterize the phase of a wave, we combine it with a known wave and measure the space- or time-varying intensity of the combination. A Michelson interferometer characterizes interference in time. A wave hits a beam splitter, and half goes through each arm and the halves coherently combine. By varying the lengths of the arms or any optical obstacles on the way, a phase shift can be induced and interference effects can be induced with it. Half the light goes down a *reference arm* and the other half goes down a *measurement arm* perpendicular to it.

To get the phase, we induce a controlled phase shift. This allows us to completely specify the phase up to a modular factor of 2π. This "wrapped phase" can be recovered by interferometry, and it can then be unwrapped more or less by guessing that jumps of 2π represent phase wrapping rather than being a part of the wave itself.

Interferometers measure optical path length differences. (Out of time, I missed some stuff)

The same as in time can be done in space. This was done by Young's double-slit experiment, that showed the interference pattern.

EE 118: Introduction to Optical Engineering **Spring 2019**

Lecture 18: Interference in thin films

Lecturer: Laura Waller *3 April* *Aditya Sengupta*

We want to derive an expression for the intensity of light that has undergone interference. Consider the case of a pinhole in an opaque screen, with a little bit of light coming through a slit and spreading out from there as a spherical wave. The field at a point x' due to a pinhole at $x = x_0$ is

$$E = \frac{e^{ik(z+l)}}{i\lambda l} e^{i\pi \frac{(x'-x_0)^2}{\lambda l}} \tag{18.1}$$

Now, if there is a pinhole at $x = -x_0$, this expression only changes to have $x' + x_0$ instead of $x' - x_0$. If we have both of these together, the interference pattern can be described by the squared magnitude of their sum,

$$g(x') = \frac{e^{ik(z+l)}}{i\lambda l} \left(\exp\left(\frac{i\pi}{\lambda l}(x' - x_0)^2 \right) + \exp\left(\frac{i\pi}{\lambda l}(x' + x_0)^2 \right) \right) \tag{18.2}$$

By expanding, and taking the absolute value squared while dropping oscillatory terms whose magnitude is 1, we end up with

$$I(x') = |g(x')|^2 = \frac{4}{\lambda^2 l^2} \cos^2\left(\frac{2\pi}{\lambda l} x_0 x' \right) = \frac{2}{\lambda^2 l^2}\,(1) + \cos\left(2\pi \frac{x'}{\Lambda} \right) \tag{18.3}$$

where $\Lambda = \frac{\lambda l}{2x_0}$ is the spatial period.

The double-slit experiment can be analyzed, rather than in terms of the electric fields, in terms of rays. For any point on the imaging screen, whether the point is bright or dark depends on the phase shift between the two light waves, which can be determined by the relative optical path lengths. Constructive interference occurs when $\Delta OPL = m\frac{\lambda}{2}$ for m even, and destructive interference occurs for the same condition but when m is odd.

If the pinholes are made larger, then the overall effect would be the same as if multiple slits were put together, creating an overall blur. If the distance betweens lits were increased, the distance between dark fringes would decrease, and if the wavelength of light were increased, the distance between light fringes would also increase.

Huygens' principle can be seen as a consequence of combining many pinholes in this way. Each point on a wavefront acts as a secondary light source emitting a spherical wave.

What if more than two plane waves were to interfere? This becomes a more difficult roblem to analytically answer; a "speckle", the result of this, is essentially the sum of many random vectors.

Thin-film interference is seen in real life often. This occurs when there is a thin film, say oil on water, of a different refractive index than the bulk medium. The thickness of the thin film should be roughly constant to within a couple wavelengths; call this thickness d. Consider light coming into the thin film surface at angle θ_1. The Fresnel equations tell us how much is reflected and transmitted. Light reflects off the other end of the thin film and refracts outwards at the top interface. At the same time, the initially reflected light is also present. Therefore there are two reflected rays with an optical path difference between them. From geometry, we eventually get $\Delta OPL = 2dn_2 \cos\theta_2$.

Thin-film interference is why colour fringes are created in glass that has internal cracks. This is a combination of constructive or destructive reflections, where the optical path difference is an integer multiple of $\frac{\lambda}{2}$ as before.

EE 118: Introduction to Optical Engineering **Spring 2019**

Lecture 19: Coherence

Lecturer: Laura Waller *8 April* *Aditya Sengupta*

19.1 Thin-film coatings continued

Thin-film coatings can be constructed to make an anti-reflective lens. Recall that $\Delta OPL = 2dn_2 \cos\theta_2$; to cause destructive interference in order to zero out reflections, we want $\Delta OPL = n_2 \frac{\lambda}{2}$. Additionally, the intensities have to be the same. Assuming this is the case, we get

$$d = \frac{\lambda}{4\cos\theta_2} \tag{19.1}$$

The thickness is around 138 nm (for $\lambda = 550$nm and $\theta = 10°$.), or any thickness that yields this modulo 2π. This can be precisely made.

A perfect anti-reflection coating is impossible to make, because of secondary reflections. More light could be made to reflect by adding more thin-film layers; multilayer mirrors that employ this technique are common in optics labs.

19.2 Coherence

The requirements for two beams to interfere were that they both have the same wavelength, same polarization, and both be coherent. If incoherent waves combine, then we just get $\langle I \rangle = I_1 + I_2$; if completely coherent waves combine, then as we derived before $\langle I \rangle = I_1 + I_2 + 2\sqrt{I_1 I_2} \cos\langle\Delta\varphi\rangle$.

Suppose we have a wave whose wavefronts are not quite periodic in time. We characterize this by temporal coherence; how correlated is the wave to the wave after a time-step? Similarly, spatial coherence describes fluctuations in space, such as the shape of wavefronts. Measuring coherence becomes difficult because it is a measure of correlation. We cannot make direct measurements of this, only averaged ones.

First consider spatial coherence. Consider two particular points P_1 and P_2 on the wave. If their temporal variation is the same, or if they are negatives of one another, they are highly correlated; if the temporal variation at fixed points is very different, then they are not correlated. We can filter incoherent light to make it coherent; a pinhole aperture makes a wave spatially coherent.

Temporal coherence measures time correlations. Consider a time-versus-amplitude plot; if it is started in parallel with a perfect sine wave of the same period and amplitude, the coherence time is the maximum duration over which the perfect sine wave is in phase with the wave.

We previously said that the Michelson interferometer was a time-based interferometer; we can use it to measure the coherence time of a wave. Consider a wave with a set of random intensities over time, which the interferometer delays and recombines with the original wave. Over time, the fringe visibility (the measure of correlation, or contrast in the interference pattern) reduces as the waves get less and less correlated. The distance over which the fringe visibility drops from 1 to some conventional point (say 0.6) is the temporal coherence. Strangely, it has dimensions of length, because it is always measured physically. Sometimes short coherence is preferable.

Spatial coherence can be measured with a Young's interferometer.

Light from incoherent sources becomes more spatially coherent if you move farther away from it. Close to the slits, light from two interfering sources is mostly from one or another. Farther away, both slits start to see an even mix.

An extended source has low spatial coherence.

Coherent modes: partially coherent beams can be decomposed into a basis set of coherent modes, in which each mode is coherent with itself but not with others. As an example of engineering spatial coherence, a rotating diffuser creates partial coherence. Each random speckle field is a coherent mode, and the time averaged field is perfectly coherent. We can create a quasi-monochromatic partially coherent beam by doing this.

EE 118: Introduction to Optical Engineering **Spring 2019**

Lecture 20: Diffraction I: Intro to Fourier Optics

Lecturer: Laura Waller *10 April* *Aditya Sengupta*

20.1 Diffraction

Diffraction is a different type of interference. When we talk about diffraction, we are usually talking about light propagating. When light hits an object, it spreads out like water waves creating a "wake", appearing to bend around an object. Red waves do not scatter as much as blue waves, and blue waves scatter to higher angles than red waves.

Diffraction can be predicted by Huygen's principle, but a mathematical description of diffraction via Huygen's principle is difficult. To make this easier, we write it as a convolution,

$$g(x) \circledast h(x) = \int g(x')h(x-x')dx' \tag{20.1}$$

Coherent light propagates first in the near field (where evanescent waves must be considered) to the Fresnel region (paraxial approximation) to the fractional Fourier region (similar to Fresnel but with physical spreading) to finally the Fraunhofer region (where optics can be described by a Fourier transform).

An optical Fourier transform can be done in two ways; by far-field propagation or by a $2f$ system. We denote a Fourier transform by $g(x", y") = \widetilde{G}\left(\frac{x"}{\lambda z}, \frac{y"}{\lambda z}\right)$. When done by a lens (where the object and image distances are both f), we set $z = f$.

20.2 Fourier analysis

We know about spatial frequency, and that it is related to the angle of propagation by $\sin\theta = \lambda f_x$. The Fourier transform decomposes a function of space into its constituent spatial frequencies.

$$\widetilde{G} = \mathcal{F}(g) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x,y)\exp\left[-i2\pi(f_X x + f_Y y\right] dxdy \tag{20.2}$$

We will deal with 2D spatial frequencies, where the transformation is unitary. $g(x, y)$ is a complex function and it is usually separable in x and y. The Fourier transform is linear, meaning it has homogeneity and superposition: $\mathcal{F}(\lambda g) = \lambda\mathcal{F}(g)$ and $\mathcal{F}(g_1 + g_2) = \mathcal{F}(g_1) + \mathcal{F}(g_2)$. Intuitively, wide functions of space have thin Fourier representations and vice versa.

By taking a Fourier transform, we want to decompose a pattern into a linear basis function set of spatial frequencies (an inverse Fourier transform). Some common Fourier inputs are the rectangle function, sinc (rect and sinc are Fourier pairs), the signum function, the triangle function, and the comb function (also known as an impulse train). Fourier transforms have many useful properties, which can be read about online like here:

https://ocw.mit.edu/resources/res-6-007-signals-and-systems-spring-2011/video-lectures/lecture-9-fo
MITRES_6_007S11_lec09.pdf.

Far away from a light source, we take a Fourier transform of the input slit; when $R > a^2/\lambda$, this describes the intensity distribution on a far-field observation plane. For example, the intensity due to a rectangular slit is a sinc squared ($I \propto |E|^2$).

If illumination is not monochromatic, then diffraction patterns like rainbows (based on the shape of the grating) are created.

Suppose the transmission of a wavefront is described by a function $g_t(x, y)$, such that $g_+(x, y) = g_-(x, y) \times g_t(x, y)$. The transparency has two effects: attenuation and phase delay. The transmittance function can be described as a complex transmission function, $g_t(x, y) = a(x, y)e^{i\varphi(x,y)}$ where the modulus is the attenuation and the phase is the phase delay. This is assuming that the features on the transparency are greater than one wavelength.

Consider a rectangular aperture,

$$g_{in}(x, y) = \text{rect}\left(\frac{x}{x_0}\right)\text{rect}\left(\frac{y}{y_0}\right) \tag{20.3}$$

The Fourier transform is

$$G_{in}(u, v) = x_0 y_0 \text{sinc}(x_0 u)\text{sinc}(y_0 v) \tag{20.4}$$

This lets us say that g_{out} is proportional to sincs in both the x and y directions.

The Fourier transform of a sinusoid can be found using Euler's formula; if $g(x) = cos(2\pi f_0 x)$ then $G(f_x)$ is a sum of two deltas. If there is a sinusoid in only one dimension, the Fourier transform is just two dots along that direction. The higher the frequency, the closer together the dots (double check that). If the sinusoid rotates in the $x - y$ plane then the dots rotate with it.

Diffraction gratings have orders based on the optical path difference. This is given by $d \sin\theta = n\lambda$.

Diffraction gratings can be split into amplitude gratings, which continue forever and cause amplitude variation keeping the phase constant, and phase gratings, which cause phase variation. In the first, there is a periodic sinusoid in the amplitude, and in the second, it is in the phase. $g_1(x) = \frac{1}{2}\left(1 + m\cos\left(2\pi\frac{x}{\Lambda} + \phi\right)\right)$ and $g_2(x) = \exp\left(i\frac{m}{2}\sin\left(2\pi\frac{x}{\Lambda} + \phi\right)\right)$.

Consider a sinusoidal amplitude grating like the first one above. Using Euler's formula we can rewrite this,

$$g_t(x) = \frac{1}{2} + \frac{m}{4}\exp\left(i2\pi\frac{\sin\theta}{\lambda}x\right) + \frac{m}{4}\exp\left(-i2\pi\frac{\sin\theta}{\lambda}x\right) \tag{20.5}$$

$$g_t(x) = \frac{1}{2} + \frac{m}{4}\exp(i2\pi u_0 x) + \frac{m}{4}\exp(-i2\pi u_0 x) \tag{20.6}$$

where $\sin\theta = \frac{\lambda}{\Lambda}$ and $u_0 = \frac{1}{\Lambda}$.

The last bit of terminology here is the diffraction efficiency (a function of n), the amplitude of the nth order squared. $\eta_0 = \left(\frac{1}{2}\right)^2$, and $\eta_{\pm 1} = \left(\frac{m}{4}\right)^2$.

Any grating can be decomposed into a sum of sinusoids, and the resultant image found in this way (demo).

EE 118: Introduction to Optical Engineering **Spring 2019**

Lecture 21: Wave Optical Systems

Lecturer: Laura Waller *15 April* *Aditya Sengupta*

We want to look at imaging from a systems perspective. Suppose we have the property of linearity for an optical system H, i.e. we can add and scale solutions. Then, we can use Green's method to solve a system of linear equations using the Fourier transform. Put another way, we put in an impulse function and get out a point spread function; if the input is a sum of impulses, the output must be a sum of impulse responses.

Convolution in real space is multiplication in Fourier space, which makes convolution computationally very efficient. Imaging systems usually have space invariance (as an assumption; if aberrations are present this is not the case) which allows us to convolve an input and a point-spread function to get a system output. If H is not space invariant, convolution cannot be used.

We will use the systems perspective to analyze the optical Fresnel region, which is closer than the Fraunhofer region. On the way to far-field, intensity variations occur due to wave overlapping. In particular, 3 plane waves of infinite extent will create intensity variations that are periodic along z. This is called the Talbot effect.

At $z = 10000\lambda$ and $z = 20000\lambda$ from a sinusoidally varying grating as an object, the pattern is the same, but the spatial frequency doubles at 5000λ. This is called "self-imaging", where a periodic function repeats itself along z with propagation.

The Talbot effect can be physically explained as a consequence of a plane wave incident on a grating, to create three plane waves: of 0th, 1st, and -1st orders. At a distance z, the phase delay is $\pi \frac{z^2}{\Lambda\lambda^2}$ (verify) and so at certain lengths, we get constructive or destructive interference. The Talbot effect can be used to create repeating 3D intensity patterns for lithography.

For small objects that are not gratings, in general, the far-field image can be predicted by the Fourier transform, and in the middle, the image stretches out till it becomes its own Fourier transform.

Propagation is a convolution. Let the input be some complex field $g = Ae^{i\varphi}$. We Fresnel propagate by a distance z. By Huygen's principle, we know that the complex field is the sum of all the spherical wavelets produced at $z = 0$:

$$g'(x', y') = \frac{z}{i\lambda} \int \int g(x, y) \frac{e^{ikr}}{r^2} dx dy \tag{21.1}$$

where $r = \sqrt{z^2 + (x' - x)^2 + (y' - y)^2}$. The propagation of the wave field can be treated as a 2D convolution of the complex field with a spherical wave, $g'(x, y) = g(x, y) \circledast h(x, y)$ where $h(x, y) \propto \frac{e^{ikR}}{r}$. Huygen's principle predicts spherical wavelets, and we can expand the above $h(x, y)$ to

$$h(x, y) \propto ik(x^2 + y^2 + z^2)^{1/2} = ikz \left(1 + \frac{1}{2}\left(\frac{x^2 + y^2}{z^2}\right) - \frac{1}{8}\left(\frac{x^2 + y^2}{z^2}\right)^2 + \dots\right) \tag{21.2}$$

For phase, $R \approx z\left(1 + \frac{x^2+y^2}{2z}\right)$.

This $h(x, y)$ acts like a divergent lens.

We can use this to make an algorithm for digital propagation. Let there be a complex object with $g(x, y) = A(x, y)e^{i\varphi(x,y)}$. Put this through a system $h_z(x, y)$, and we get $g'(x', y') = g(x, y) \circledast h_z(x, y)$. We take the intensity, $I_z(x, y) = |g(x, y) \circledast h_z(x, y)|^2$. This contains both phase and amplitude information.

The Fresnel propagation kernel varies with distance; the greater the distance, the greater the number of rings in the PSF and the more blurred the intensity gets.

The Fresnel number, $F = \frac{\Delta x^2}{\lambda z}$, describes an amount of diffraction. This is a good indicator of whether imaging is coherent.

Digital refocusing uses convolution. Consider the image of a bug in a complex impulse function created by ripples in water. The distorted image is the convolution of these two. If distortion is caused by defocus, the image can be brought back into focus by backpropagating.

EE 118: Introduction to Optical Engineering **Spring 2019**

Lecture 22: Diffraction: Fresnel to Fraunhofer Propagation

Lecturer: Laura Waller *17 April* *Aditya Sengupta*

Recall that we propagate a wave by convolving it with a point-spread function. Last time we covered digital Fresnel propagation, which is done by convolution: $g_{out}(x', y') = g_{in}(x, y) \circledast h_z(x, y)$. Then we can take the intensity.

Suppose there is an circular shaped disk in a beam of light. We are interested in finding out whether light constructively interferes at the center. We get an optical path length from the edge of the disk of $n\sqrt{r^2 + z^2}$, which is cylindrically symmetric, so the difference in OPL and therefore the phase difference is 0. Therefore we get constructive interference. This is called the Poisson spot or Arago spot. It is a coherent effect resulting from constructive interference, so it gives evidence that light is a wave. It is due to Fresnel diffraction, therefore it occurs when the Fresnel number $\frac{r^2}{\lambda z} > 1$. We do not see this often in everyday life because if the light is not coherent, then the fringes and spot are blurred.

In between the Fresnel and Fraunhofer regions, we analyze optics in the fractional Fourier region. This is similar to Fresnel but takes into account physical spreading. A fractional Fourier transform requires that we define the order of a Fourier transform, which is just the number of times a Fourier transform is applied to itself: $\mathcal{F}^2(f) = \mathcal{F}(\mathcal{F}(f))$. The p/qth Fourier transform is the operation which when applied to itself q times yields $\mathcal{F}^p(f)$.

We can show that Fresnel propagation goes to Fraunhofer propagation at long distances. Consider Fresnel propagation by convolution with a PSF for propagation by distance z, $h(x) = \frac{e^{ikz}}{i\lambda z} e^{ik\frac{x^2}{2z}}$. We get

$$g_{out} = \int g_{in}(x) \frac{e^{ikz}}{i\lambda z} e^{i\frac{k}{2z}(x'-x)^2} dx \tag{22.1}$$

$$g_{out} = \frac{e^{ikz}}{i\lambda z} e^{\frac{ik}{2z}x'^2} \int g_{in} x e^{ikx^2/2z} e^{-ikx'x/z} dx \tag{22.2}$$

Drop the x^2 term because of the Fraunhofer far-field condition. Eventually we get

$$g_{out} = \frac{e^{ikz}}{i\lambda z} e^{ikx'^2/2z} \int g_{in}(x) e^{-i\pi x' x/\lambda z} dx \tag{22.3}$$

which is just the statement of the Fourier transform with spatial frequency $f_x = \frac{x'}{\lambda z}$, as we expect.

The Fourier transform is not space invariant, so it cannot be written as a convolution and no transfer function exists. Fresnel optics still applies, and Fraunhofer optics is one subset of Fresnel.

As mentioned previously, an optical Fourier transform can be done either by a 2f system ($g(x, y) = G\left(\frac{x}{\lambda f}, \frac{y}{\lambda f}\right)$) or by carrying out far-field propagation ($g(x, y) = G\left(\frac{x}{\lambda z}, \frac{y}{\lambda z}\right)$). We saw how to derive the far-field propagation method. We can derive the $2f$ system one by carrying out Fresnel propagation twice, with a spherically-shaped phase delay representing the lens in the middle.

The total phase delay is

$$\varphi(x,y) = kn\Delta(x,y) + k(\Delta_0 - \Delta(x,y)) \tag{22.4}$$

We split the lens into three parts to get this, $\Delta(x,y) = \Delta_1(x,y) + \Delta_2(x,y) + \Delta_3(x,y)$. From geometry, we get

$$\Delta(x,y) = \Delta_0 - R_1\left(1 - \sqrt{1 - \frac{x^2+y^2}{R_1^2}}\right) - R_2\left(1 - \sqrt{1 - \frac{x^2+y^2}{R_2^2}}\right) \tag{22.5}$$

Applying the paraxial approximation, this simplifies to

$$\Delta(x,y) = \Delta_0 - \frac{x^2+y^2}{2}\left(\frac{1}{R_1} - \frac{1}{R_2}\right) \tag{22.6}$$

Ignoring the constant phase term, we get a lens transformation with the above phase,

$$t_l(x,y) = \exp\left(-j\frac{k}{2f}(x^2+y^2)\right) \tag{22.7}$$

The field just before the lens is given by

$$g_{len^-}(x') = \int g(x) e^{\frac{i\pi}{\lambda f}(x'-x)^2} dx \tag{22.8}$$

and just after the lens it is

$$g_{lens^+}(x') = g_{lens^0} e^{i\pi x'^2/\lambda f} \tag{22.9}$$

At the output plane, we get

$$g_{out}(x'') = \int g_{lens^+}(x') e^{(i\pi/\lambda f)(x''-x')^2} dx' \tag{22.10}$$

which in terms of the input $g(x)$ comes out to

$$g_{out}(x'') = \int g(x) \exp\left(-i2\pi x x''/\lambda f\right) dx \tag{22.11}$$

which is again a Fourier transform with frequency $f_x = x''/\lambda f$.

Parseval's theorem relates the energy in terms of space with that in terms of frequency,

$$\int_{-\infty}^{\infty} |f(x)|^2 dx = \frac{1}{2\pi} \int_{-\infty}^{\infty} |F(\omega)|^2 d\omega \tag{22.12}$$

If you do two Fourier transforms with a $4f$ system, we get the original image back with a sign shift: $g''(x'') = g(-\frac{f_2}{f_1}x'')$.

A mask in Fourier space acts as a filter. For a desired point-spread function, the Fourier transform of that PSF can be the Fourier plane transparency (in the middle of a $4f$-type system), which will take an input wave to a point to the desired PSF and finally back to the desired filtered field: $g''(x'') = g(-(f_2/f_1)x'') \circledast T(x')$. (See slides for examples of Fourier filtering).

EE 118: Introduction to Optical Engineering **Spring 2019**

Lecture 23: $4f$ systems

Lecturer: Laura Waller *22 April* *Aditya Sengupta*

23.1 Fourier masks

We previously saw that a mask in Fourier space acts as a filter. This allows us to design for a desired point-spread function by constructing the corresponding Fourier mask. We can use this to figure out the filters that would cause certain outputs. For example, if a transform takes a point to three vertically-patterned points in space, we know that the point-spread function is those three vertically-patterned points. We expect that the Fourier mask looks like $f(x') = 1 + \cos(f_x x')$. We can figure out this spatial frequency based on known focal lengths. Say $f_1 = 100$mm and $f_2 = 125$mm in the $4f$ system being used to do this transform. We say $d \sin\theta = \lambda$, and by trigonometry we see that $\tan\theta = \frac{0.75/2}{f}$ (where 0.75mm is the distance between the two first orders). We get $d = 177\mu m$.

For a system with higher orders (in the slides, every odd order is shown, which corresponds to a rectangular mask) the approximation of sines and cosines matching each of the spot ends up forming the Fourier series of the real mask.

23.2 LTI System Convolution

If a system is LTI, then every point sees the same point spread function, and convolution can be reversed by using linearity. Suppose we have output and input images, and we want to find the PSF with which the input was convolved. In Fourier space, $g_{out} = h \circledast g_{in}$ transforms to $G_{out} = HG_{in}$, so we can deconvolve just by dividing in Fourier space. In this case, it is necessary for the Fourier transform not to have any zero values, otherwise the input would blow up to infinity at those points (by dividing $G_{in} = \frac{G_{out}}{H}$).

Engineering the point-spread function allows us to reconstruct images via deconvolution. Applying and deconvolving from a uniform blur allows for an extended depth of field, for example. (A lot of this lecture was demos and examples, not theory)

The PSF of a circular aperture is an Airy disk. If there are many circular objects in real space, their Airy disks overlap in Fourier space. If two points are sufficiently far apart that their Airy disks sum to just one larger Airy disk, then the points are not resolved, and otherwise they are. The Nyquist sampling criterion applies to the resolution of the pixels. A larger PSF causes more blurring.

For a given lens diameter, blue has the best resolution in a microscope; wavelength and NA both affect this, and the smaller the wavelength the smaller the diffraction limit. The Rayleigh criterion formalizes this,

$$res = 0.61\frac{\lambda}{NA} \tag{23.1}$$

where 0.61 is half of the first minimum of the Airy function.

The modulation transfer function, a measure of contrast as a function of spatial frequency, also characterizes resolution.