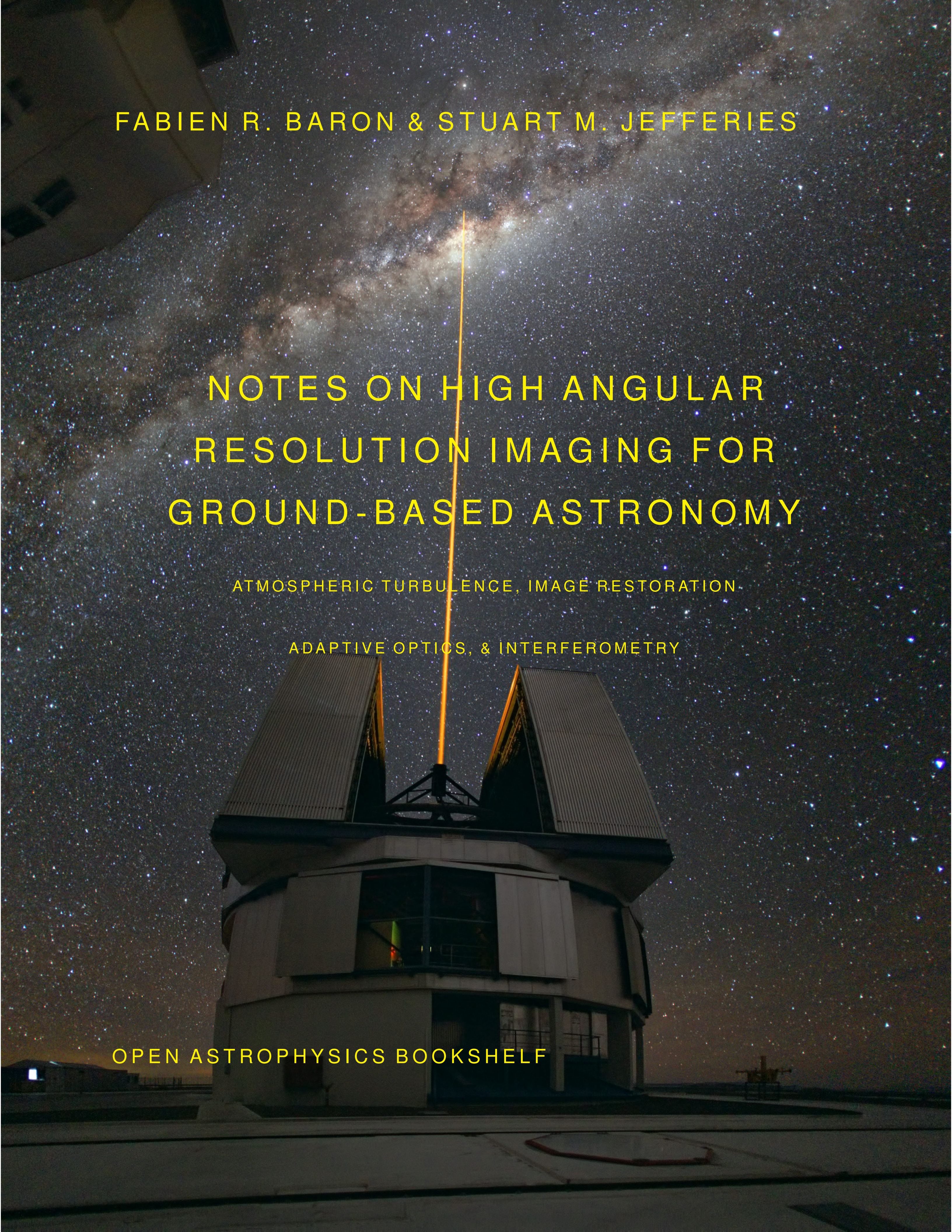


FABIEN R. BARON & STUART M. JEFFERIES

NOTES ON HIGH ANGULAR RESOLUTION IMAGING FOR GROUND-BASED ASTRONOMY

ATMOSPHERIC TURBULENCE, IMAGE RESTORATION

ADAPTIVE OPTICS, & INTERFEROMETRY

OPEN ASTROPHYSICS BOOKSHELF

About the cover: In mid-August 2010 European Southern Observatory's Photo Ambassador Yuri Beletsky snapped this photo at ESO's Paranal Observatory, Chile. A group of astronomers were observing the center of the Milky Way using the laser guide star facility at Yepun, one of the four Unit Telescopes of the Very Large Telescope (VLT). Yepun's laser beam crosses the southern sky and creates an artificial star at an altitude of 90 km high in the Earth's mesosphere. The Laser Guide Star (LGS) is part of the VLT's adaptive optics system and is used as a reference to correct the blurring effect of the atmosphere on images. The color of the laser is precisely tuned to energize a layer of sodium atoms found in one of the upper layers of the atmosphere – one can recognize the familiar color of sodium street lamps in the color of the laser. This layer of sodium atoms is thought to be a leftover from meteorites entering the Earth's atmosphere. When excited by the light from the laser, the atoms start glowing, forming a small bright spot that can be used as an artificial reference star for the adaptive optics. Using this technique, astronomers can obtain sharper observations. For example, when looking towards the center of our Milky Way, researchers can better monitor the galactic core, where a central supermassive black hole, surrounded by closely orbiting stars, is swallowing gas and dust. Taken with a wide angle lens, this photo covers about 180° of the sky.
Credit & Copyright: ESO/Yuri Beletsky, CC BY 4.0

© 2017 - 2021 Fabien Baron & Stuart M. Jefferies
git version c2f32663 . . .

Except where explicitly noted, this work is licensed under the Creative Commons Attribution-ShareAlike 4.0 International (CC BY-SA 4.0) license.

Preface

These notes are currently being written while teaching a graduate-level astronomy course on adaptive optics, image restoration and optical interferometry at Georgia State University. The text layout uses the `tufte-book` (`https://tufte-latex.github.io/tufte-latex/`) LaTeX class: the main feature is a large right margin in which the students can take notes; this margin also holds small figures and sidenotes. Exercises are embedded throughout the text. These range from "reading exercises" to longer, more challenging problems. In addition, there are several numerical exercises in Python and Julia (available from `https://julialang.org`). Because the exercises are spread throughout the text, there is a "List of Exercises" in the front matter to help with looking for specific problems.

PLEASE BE ADVISED THAT THESE NOTES ARE UNDER ACTIVE DEVELOPMENT; to refer to a specific version, use the eight-character stamp labeled "git version" on the copyright page.

Contents

1 Fourier Optics 1

1.1 Fourier transform definitions 1

The spatial harmonic function - plane wave relationship 2

Transfer function of free space 3

Fourier Transform using a lens 5

Optical spatial filtering 5

1.2 Characterizing instrument aberrations 5

Point source 5

The pupil field 6

Optical path length 6

Zernike polynomials 7

Point spread function 9

Airy pattern 10

Polychromatic PSF 11

1.3 Image formation 13

Imaging is convolving 13

Fast convolution 13

Optical transfer function 13

1.4 Sparse aperture characteristics and OTF 14

1.5 Atmosphere and turbulence 16

The Kolmogorov model of turbulence 17

C_n^2 profile 18

Fried parameter: r_0 19

Atmospheric time constant: τ_0 20
Isoplanatic angle: θ_0 20
Layer model for atmosphere 21
Scintillation 21
1.6 Imaging through atmospheric turbulence 23
Long-exposure images 24
Atmospheric transfer function 26
Speckle imaging 26
Shift-and-add 26
Speckle Interferometry 27
1.7 Some practical matters 28
Convolution and correlation 28
Autocorrelation 28
Computing convolution and correlation using the discrete Fourier Transform 28
Image sampling 29
Simulating Atmospheric Phase Screens 30

2 Image Restoration 33
2.1 Image preprocessing 33
2.2 Multiplicative deconvolution 34
2.3 Blind deconvolution 35
2.4 Diversity imaging methods 37
2.5 Some practical matters 39
Cost functions 39
Optimization routines 40
Analytic derivatives for pupil plane model of PSF 40
Modeling PSFs at different wavelengths 41

3 Adaptive Optics Compensation 43
3.1 Wavefront sensing 43
Gradient-Based Sensors 44
Curvature-Based Sensors 46
Phase Retrieval 47

3.2 Wave front reconstruction 47

Zonal reconstruction 48

Modal reconstruction 49

3.3 Deformable mirrors 49

Segmented DMs 50

Continuous face-sheet DMs 50

Bimorph DMs 50

MEMS DMs 50

Adaptive secondary mirrors 51

3.4 Close loop operation 51

3.5 Artificial guide stars 51

Rayleigh beacons 53

Sodium beacons 53

4 Optical interferometry: theory 55

4.1 Coherent and incoherent light 55

4.2 A two-element interferometer 57

Monochromatic interference: fringes 57

Polychromatic light, Coherence Envelope 58

4.3 Amplitude interferometry 58

Spatial coherence: the Van Cittert-Zernike theorem 58

Temporal coherence: the Wiener-Khintchine theorem 60

The birth of stellar optical interferometry 60

4.4 Hanbury Brown and Twiss effect: intensity interferometry 61

The HBT effect 61

The Narrabri Stellar Intensity Interferometer 62

A resurgence of interest 63

4.5 From visibilities to models and images: the inverse problem 63

1

Fourier Optics

Fourier optics provides a description of the propagation of light based on harmonic analysis (the Fourier transform) and linear systems. Whereas geometrical ray optics is a convenient tool to determine imaging characteristics such as the location of the image and the image magnification, a complete description of the imaging system requires the wave properties of light and associated processes like diffraction to be included. These processes determine the resolution of optical devices, the image contrast, and the effect of spatial filters. The theory of Fourier Optics provides the cornerstone for the analysis of diffraction, coherence, and imaging. It is also fundamental to wavefront control, propagation through random media, and holography.

1.1 Fourier transform definitions

The analytical Fourier transform of an arbitrary function $g(x, y)$ is given by

$$G(u, v) = \int_{-\infty}^{\infty}\int g(x, y) \exp\left[-i2\pi(ux + vy)\right] dx\ dv \tag{1.1}$$

where $G(u, v)$ is the transform result and u, v are independent frequency variables associated with the spatial coordinates x and y, respectively, and $i = \sqrt{-1}$. We note the Fourier coordinates (u, v) have units of inverse length and are called spatial frequencies. The shorthand notations for the Fourier transform operation is
$\mathscr{F}\{g(x, y)\} = G(u, v)$.

The analytic inverse Fourier transform is given by

$$g(x, y) = \int_{-\infty}^{\infty}\int G(u, v) \exp\left[i2\pi(ux + vy)\right] du\ dv\ . \tag{1.2}$$

Equation 1.2 shows that an arbitrary function $g(x, y)$ can be analyzed as a superposition, or integral of harmonic functions of different spatial frequencies, u and v, and complex amplitudes $G(u, v)$. The shorthand notation for the inverse Fourier transform is $\mathscr{F}^{-1}\{G(u,v)\} = g(x,y)$.

Implementation of the Fourier transform on a computer is usually achieved using the Fast Fourier Transform (FFT) which is a computationally efficient version of the Discrete Fourier Transform (DFT). If we denote the sampled version of the continuous function $g(x,y)$ by $\tilde{g}(x,y)$ (see section 1.7), then the DFT of $\tilde{g}(x,y)$ is given by

$$\tilde{G}(p,q) = \sum_{m=-M/2}^{M/2-1} \sum_{n=-N/2}^{N/2-1} \tilde{g}(m,n) \exp\left[-i2\pi\left(\frac{pm}{M} + \frac{qn}{N}\right)\right], \tag{1.3}$$

where $\tilde{G}(p,q)$ represents the DFT of $\tilde{g}(m,n)$, $p = u/\Delta u$, $q = v/\Delta v$, $m = x/\Delta x$, $n = y/\Delta y$, $\{p,q,m,n\}$ are integers, $\{M,N\}$ are the number of samples of size $\Delta x, \Delta y$, in the x and y directions, respectively, and $\Delta u, \Delta v$ are the corresponding spectral sample sizes in the Fourier domain.

The inverse discrete Fourier transform (DFT^{-1}) is given by

$$\tilde{g}(m,n) = \frac{1}{MN} \sum_{p=-M/2}^{M/2-1} \sum_{q=-N/2}^{N/2-1} \tilde{G}(p,q) \exp\left[i2\pi\left(\frac{pm}{M} + \frac{qn}{N}\right)\right], \tag{1.4}$$

The spatial harmonic function - plane wave relationship

A traveling plane wave is represented by

$$U(\vec{r},t) = A \exp\left[i\left(k\vec{r} + \omega t\right)\right] \tag{1.5}$$

where $\vec{r} = \sqrt{x^2+y^2+z^2}$ and the wavevector $k = 2\pi/\lambda = \sqrt{k_x^2+k_y^2+k_z^2}$. Considering only the spatial component we have

$$U(x,y,z) = A \exp\left[i2\pi\left(k_x x + k_y y + k_z z\right)\right]. \tag{1.6}$$

At $z = 0$ this becomes

$$U(x,y,0) = A \exp\left[i2\pi\left(ux + vy\right)\right], \tag{1.7}$$

with $u = k_x/2\pi$ and $v = k_y/2\pi$. This has the form of a harmonic function. To see this we can project a plane wave onto the $x-y$ plane. Here the x and y wavelengths of the harmonic function are given by $\Lambda_x = 1/u$ and $\Lambda_y = 1/v$, respectively, and the plane wave needs to travel at angles $\theta_x = \sin^{-1}\left(\lambda/\Lambda_x\right) = \sin^{-1}\left(\lambda u\right)$, and $\theta_y = \sin^{-1}\left(\lambda/\Lambda_y/\right) = \sin^{-1}\left(\lambda v\right)$ for the projection to model the harmonic function. That is, a harmonic function with spatial frequencies u and v at the plane $z = 0$ is consistent with a plane wave traveling at angles θ_x and θ_y. The harmonic function pattern works like a grating. (see Fig. 1.1). In general, we can decompose the spatial distribution of $f(x,y)$

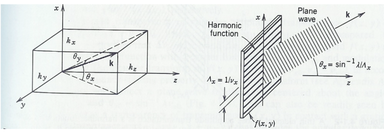


Figure 1.1:

at $z = 0$ into spatial harmonics, and treat each component separately. In this case we have

$$U(x,y,z) = \int_{-\infty}^{\infty}\int F(u,v)\exp\left[i2\pi(ux+vy)\right]\exp\left(ik_z z\right)\,du\,dv \tag{1.8}$$

which shows that an incident plane wave can be decomposed into many plane waves, each traveling at angles $\theta_x = \sin^{-1}\left(\lambda\nu_x\right)$, and $\theta_y = \sin^{-1}\left(\lambda\nu_y\right)$, with a complex envelope $F(u,v)$, the Fourier transform of $f(x,y)$.

Transfer function of free space

In Fourier optics the propagation of light between two planes is regarded as a linear system whose input $f(x,y)$ and output $g(x,y)$ are the complex amplitudes of the wave in the two planes. The output is related to the input via a transfer function. For propagation in free space, the transfer function $H(\nu_x,\nu_y)$ is given by

$$\begin{aligned} H(u,v) = \frac{g(x,y)}{f(x,y)} = \frac{U(x,y,z_1)}{U(x,y,0)} &= \exp\left[ik_z d\right] \\ &= \exp\left[i2\pi\left(\frac{1}{\lambda^2} - u^2 - v^2\right)^{1/2} d\right] \\ &= \exp\left[ikd\sqrt{1-(\lambda u)^2-(\lambda v)^2}\right], \end{aligned} \tag{1.9}$$

where d is the distance between the two planes. This is known as the Rayleigh-Sommerfeld transfer function.

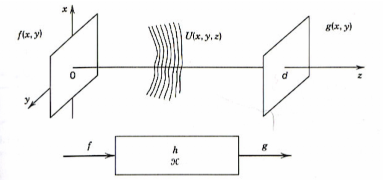


Figure 1.2: The propagation of light between two planes is regarded as a linear system.

Given an arbitrary input function $f(x,y)$ we can obtain the output function $g(x,y)$ by: (1) Determining the complex envelopes of the plane wave components in the input plane by Fourier transform

$$F(u,v) = \int_{-\infty}^{\infty}\int f(x,y)\exp\left[-i2\pi(ux+vy)\right]\,du\,dv \tag{1.10}$$

(2) Generating the complex envelopes of the plane wave components in the output plane using $H(u,v)F(u,v)$, and (3) Summing the plane waves in the output plane

$$g(x,y) = \int_{-\infty}^{\infty}\int H(u,v)F(u,v)\exp\left[i2\pi(ux+vy)\right]\,dx\,dy\,. \tag{1.11}$$

.

That is, $g(x,y)$ is a convolution of $f(x,y)$ with the Rayleigh-Sommerfeld impulse response, $h(x,y) = \mathscr{F}^{-1}\{H(u,v)\}$ (see Section 1.7). The Rayleigh-Sommerfeld expression requires that the distance between the source and the observation positions must be much greater than a wavelength.

Two approximations that are commonly used are the Fresnel and Fraunhofer approximations. The Fresnel approximation is a small angle approximation, that is,

$$2\pi \left(\frac{1}{\lambda^2} - u^2 - v^2 \right)^{1/2} z = \frac{2\pi z}{\lambda} \left(1 - \phi^2 \right)^{1/2} \simeq \frac{2\pi z}{\lambda} \left(1 - \frac{\phi^2}{2} \right) \quad (1.12)$$

and the Transfer Function, Eqn 1.9, becomes

$$H(u,v) = \exp ikz \ \exp -i\pi\lambda z \left(u^2 + v^2 \right) \quad (1.13)$$

which is valid provided $|\lambda u| \ll 1$ and $|\lambda v| \ll 1$. This is just a restriction to small angles. The impulse response is given by

$$h(x,y) = \frac{\exp ikz}{i\lambda z} \exp \left[\frac{ik}{2z} (x^2 + y^2) \right] , \quad (1.14)$$

The Fresnel diffraction expression in the observation plane is given by

$$U_2(x,y) = h(x,y) \times \int \int \left\{ U_1(\xi,\eta) \exp \left[\frac{ik}{2z} (\xi^2 + \eta^2) \right] \right\} \exp \left[-i \frac{2\pi}{\lambda z} (x\xi + y\eta) \right] d\xi d\eta \quad (1.15)$$

where (ξ, η) are coordinates in the source plane and U_1 is the field in the source plane.

The Fraunhofer approximation is a much more stringent approximation for use with very long propagation paths, i.e.,

$$z \gg \frac{\pi}{\lambda} \left(\xi^2 + \eta^2 \right) \ . \quad (1.16)$$

There is no transfer function in this case, and the Fraunhofer diffraction expression is given by a scaled version of the Fourier transform of the initial field, i.e.,

$$U_2(x,y) = h(x,y) \times \int \int U_1(\xi,\eta) \exp \left[-i \frac{2\pi}{\lambda z} (x\xi + y\eta) \right] d\xi d\eta \quad (1.17)$$

The Fresnel number is givn by

$$N_F = \frac{w^2}{\lambda z} \quad (1.18)$$

where w is the half-width of the aperture (e.g., the radius for a circular aperture) and z is the distance to the observation plane. The Fresnel region is given by $N_F \lesssim 1$ while the Fraunhofer region is give by $N_F \ll 1$.

Fourier Transform using a lens

A remarkable property of a converging lens is its ability to perform two-dimensional Fourier transforms.

The transmittance function for an ideal simple lens is given by

$$t_A(x,y) = \exp\left[-i\pi\left(x^2+y^2\right)/\lambda f\right] \tag{1.19}$$

where f is the focal length of the lens. If the field incident on the lens is $U_1(x_1,y_1)$ then the field exiting the lens is $U_1(x_1,y_1)t_A(x_1,y_1)$. Inserting this into Eqn. 1.15 for $U_1(x_1,y_1)$ and setting $z=f$ we have

$$U_2(x,y) = h(x,y)\times\int\int U_1(\xi,\eta)\exp\left[-i\frac{2\pi}{\lambda f}(x\xi+y\eta)\right]d\xi d\eta \tag{1.20}$$

which shows that the field at the focal plane of an ideal positive lens is simply the Fraunhofer pattern of the incident filed with $z=f$. If we place the source field at the front focal plane of a positive lens, a distance f from the lens, we get

$$U_2(x,y) = \frac{\exp(ikf)}{i\lambda f}\times\int\int U_1(\xi,\eta)\exp\left[-i\frac{2\pi}{\lambda f}(x\xi+y\eta)\right]d\xi d\eta \tag{1.21}$$

which shows the focal plane field is a scaled Fourier transform of the input field. A lens has the effect of bringing the far-field angular distribution into the focal plane.

Optical spatial filtering

By making a Fourier transform of an image using a lens, it is possible to change the information in amplitude and phase that is supported by this image.

For that purpose, a filter can be placed in the spatial frequency plane (u; v). A second lens, placed after the spatial frequency plane, can then be used to display the Fourier transform of F(u; v)T(u; v), where T(u; v) describes the amplitude transmission of the filter. The basic optical processor is shown in Fig. 1.3.

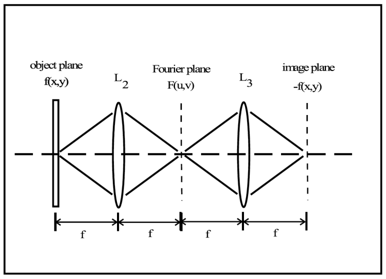


Figure 1.3: Fourier Transform by a lens. L1 is the collimating lens, L2 is the Fourier transform lens, u and v are normalized coordinates in the transform plane.

1.2 *Characterizing instrument aberrations*

Point source

A point source is an object whose angular size is below the angular resolution of the instrument. It is said to be *unresolved*. Mathematically, a point source is a 2D delta function.

$$\delta(x,y) = \begin{cases} 1 & \text{if } (x,y)=(x_0,y_0) \\ 0 & \text{elsewhere} \end{cases} \tag{1.22}$$

The pupil field

The pupil function describes how a light wave is affected upon transmission through an optical imaging system. More specifically, it is a two-dimensional complex function of the position in the pupil that indicates the relative change in amplitude and phase of the light wave. Imperfections in the optics typically have a direct effect on the pupil function, it is therefore an important tool to study optical imaging systems and their performance. The pupil function $P(u,v)$ can be split into its modulus and phase:

$$P(u,v) = A(u,v) \exp i\Phi(u,v), \tag{1.23}$$

where the modulus $A(u,v)$ includes the aperture function (percentage of light blocked) and throughput losses, and the phase $\Phi(u,v)$ describes the aberrations of the system. The pupil function of an ideal system (no aberrations, 100% transmission) is equal to one at every point within the pupil, and zero outside of it. In the case of a circular pupil of unit radius, the pupil function is $P(u,v) = \Pi(u,v)$, where:

$$\Pi(u,v) = \begin{cases} 1 & \forall\,(u,v) | \sqrt{u^2+v^2} \leq 1 \\ 0 & \text{elsewhere} \end{cases}. \tag{1.24}$$

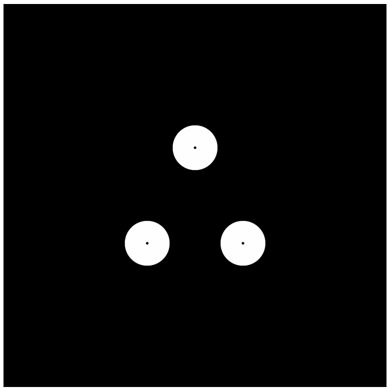


Figure 1.4: A Golay array with three sub-apertures.

The circular pupil is the most common type of aperture in astronomy. However, the continuing quest for finer angular resolution demands larger and larger apertures. Unfortunately, the manufacturing cost for monolithic optics increases faster than the diameter of the optic squared, and we are already close to the limit of what is financially feasible. Consequently, for applications that do not require extremely high sensitivity, the notion of a sparse aperture offers a novel solution for decreasing the volume and weight of a telescope while allowing for large pupil sizes. Golay[1] proposed that there is an optimal imaging configuration for sparse arrays that depends on the number of sub-apertures.

Optical path length

The optical path length (OPL) or optical distance is the product of the geometric length of the path light through the system, by the index of refraction of the medium through which it propagates. A difference in optical path length between two paths is often called the optical path difference (OPD). Optical path length is important because it determines the phase of the light and governs interference and diffraction of light as it propagates.

The OPD corresponds to the phase shift undergone by the light emitted from two previously coherent sources when passed through

[1] M. Golay, "Point Arrays Having Compact Non-redundant Autocorrelations", J.Opt Soc. Am., vol. 61, pp 272, 1971.

mediums of different refractive indices. For example, a wave passed through glass will appear to travel a greater distance than an identical wave in air. This is because the source in the glass will have experienced a greater number of wavelengths due to the higher refractive index of the glass. The OPD can be calculated from the following equation:

$$\mathrm{OPD} = d_1 n_1 - d_2 n_2 \tag{1.25}$$

where d_1 and d_2 are the distances of the ray passing through medium 1 or 2, n_1 and n_2 the refractive indices. In a medium of constant refractive index n the OPL for a path of physical length d is just $\mathrm{OPL} = nd$. If the refractive index varies along the path, the OPL is given by

$$\mathrm{OPL} = \int_C n(s)\mathrm{d}s, \tag{1.26}$$

where $n(s)$ is the local refractive index as a function of distance, s, along the path C. Finally, the phase shift introduced by an OPD is given by:

$$\Phi = \frac{2\pi}{\lambda}\mathrm{OPD} \quad (\mathrm{rad}) \tag{1.27}$$

provided the OPD and λ are expressed in the same units. Consequently, when considering e.g. $1/2$ wavelength of piston aberration, we are talking about a phase shift of $\Phi = \frac{2\pi}{\lambda} \times \frac{1}{2}\lambda = \pi$ radians.

Zernike polynomials

The Zernike polynomials are orthogonal two-dimensional functions, which form a basis defined over the circular support area; typically the pupil planes in classical optical imaging at visible and infrared wavelengths through systems of lenses and mirrors of finite diameter. They were introduced by Frits Zernike in the 1930s to model the pupil wavefront. [2]

A circular wavefront profile Φ associated with aberrations may be mathematically modeled using a linear sum of Zernike polynomials.

$$\Phi = \sum_{j=1}^{N} a_j Z_j \tag{1.28}$$

where j is the main index (Noll index) of Zernike polynomials Z_j, and a_j its scalar strength in the wavefront.

These Zernike coefficients are linearly independent, thus individual aberration contributions to an overall wavefront may be isolated and quantified separately. Hence an aberrated wavefront profile may be decomposed into Zernike polynomials since:

$$\forall i, \quad Z_i \cdot \Phi = \sum_{j=1}^{N} a_j Z_i \cdot Z_j = a_i Z_i \cdot Z_i = \pi a_i \tag{1.29}$$

[2] At this time, the perfect point spread function in the presence of diffraction had already been described by George Biddell Airy, as early as 1835. It took almost hundred years to arrive at a comprehensive theory and modeling of the point image of aberrated systems (Zernike and Nijboer). The analysis by Nijboer and Zernike describes the intensity distribution close to the optimum focal plane. Recent analytic results have made it possible to extend Nijboer and Zernike's approach for point-spread function evaluation to a large volume around the optimum focal point. This extended Nijboer-Zernike (ENZ) theory is instrumental in studying the imperfect imaging of three-dimensional objects in confocal microscopy or astronomy under non-ideal imaging conditions. The ENZ-theory has also been applied to the characterization of optical instruments with respect to their aberration by measuring the through-focus intensity distribution and solving an appropriate inverse problem.

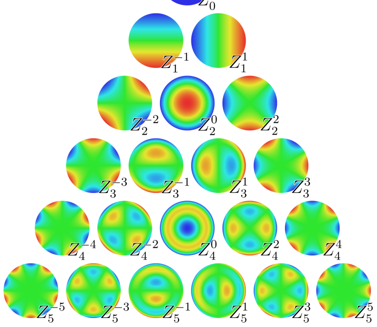


Figure 1.5: The first 21 Zernike polynomials, ordered vertically by radial degree and horizontally by azimuthal degree. Each represent a possible mode of the pupil phase.

where it is assumed $Z_i \cdot Z_j = \pi\delta_{ij}$ (π is often chosen for normalization).

Going beyond simple indexing, there are even and odd Zernike polynomials. The even Zernike polynomials are defined as:

$$Z_n^m(\rho,\theta) = R_n^m(\rho)\,\cos(m\,\theta) \tag{1.30}$$

and the odd Zernike polynomials as

$$Z_n^{-m}(\rho,\theta) = R_n^m(\rho)\,\sin(m\,\theta) \tag{1.31}$$

where m and n are nonnegative integers with $n \geq m$, θ is the azimuthal angle, ρ is the radial distance $0 \leq \rho \leq 1$, and R_n^m are the radial polynomials defined below. Zernike polynomials have the property of being limited to a range of -1 to $+1$, i.e. $|Z_n^m(\rho,\theta)| \leq 1$. The radial polynomials $R_n^m$are defined as

$$R_n^m(\rho) = \sum_{k=0}^{\frac{n-m}{2}} \frac{(-1)^k\,(n-k)!}{k!\,\left(\frac{n+m}{2}-k\right)!\,\left(\frac{n-m}{2}-k\right)!}\,\rho^{n-2\,k} \tag{1.32}$$

for $n-m$ even, and are identically 0 for $n-m$ odd.

Table 1.1: The first few Zernike modes, ordered by Noll index j, normalized so that $\int_0^{2\pi}\int_0^1 Z_j^2\,\rho\,d\rho\,d\theta = \pi$.

Noll index j	Radial degree n	Azimuthal degree m	Z_j	Classical name
1	0	0	1	Piston
2	1	1	$2\rho\cos\theta$	Tip (X-Tilt)
3	1	-1	$2\rho\sin\theta$	Tilt (Y-Tilt)
4	2	0	$\sqrt{3}(2\rho^2-1)$	Defocus (longitudinal position)
5	2	-2	$\sqrt{6}\rho^2\sin 2\theta$	Oblique astigmatism
6	2	2	$\sqrt{6}\rho^2\cos 2\theta$	Vertical astigmatism
7	3	-1	$\sqrt{8}(3\rho^3-2\rho)\sin\theta$	Vertical coma
8	3	1	$\sqrt{8}(3\rho^3-2\rho)\cos\theta$	Horizontal coma
9	3	-3	$\sqrt{8}\rho^3\sin 3\theta$	Vertical trefoil
10	3	3	$\sqrt{8}\rho^3\cos 3\theta$	Oblique trefoil
11	4	0	$\sqrt{5}(6\rho^4-6\rho^2+1)$	Primary spherical
12	4	2	$\sqrt{10}(4\rho^4-3\rho^2)\cos 2\theta$	Vertical secondary astigmatism
13	4	-2	$\sqrt{10}(4\rho^4-3\rho^2)\sin 2\theta$	Oblique secondary astigmatism
14	4	4	$\sqrt{10}\rho^4\cos 4\theta$	Vertical quadrafoil
15	4	-4	$\sqrt{10}\rho^4\sin 4\theta$	Oblique quadrafoil

The first few Zernike polynomials are show in Table 1.2; ρ is the normalized pupil radius with $0 \leq \rho \leq 1$, θ is the azimuthal angle around the pupil with $0 \leq \theta \leq 2\pi$, and the fitting coefficients $a_0, \ldots, a_8$ are the wavefront errors in wavelengths.

Their advantages are the simple analytical properties inherited from the simplicity of the radial functions and the factorization in radial and azimuthal functions; this leads, for example, to closed-form expressions of the two-dimensional Fourier transform in terms of Bessel functions. They are commonly used in adaptive optics, where they can be used to effectively cancel out atmospheric turbulence. For example, one of the Zernike terms (for m = 0, n = 2) is called defocus. By coupling the output from this term to a control system, an automatic focus can be implemented. Another advantage is that it is often useful to describe the wave front aberrations by the rms value σ averaged over the pupil. If the wave front is modeled as a Zernike series then

$$\sigma = \sqrt{\sum_{j=2}^{\infty} a_j^2} \tag{1.33}$$

Their disadvantage, in particular if high n are involved, is the unequal distribution of nodal lines over the unit disk, which introduces ringing effects near the perimeter ($\rho \approx 1$), which often leads attempts to define other orthogonal functions over the circular disk (e.g., disk harmonics). Wavefronts with very steep gradients or very high spatial frequency structure, such as produced by wave propagation through atmospheric turbulence, are not well modeled by Zernike polynomials, which tend to low-pass filter the fine three-dimensional spatial definition in the wavefront. In this case, other fitting methods such as fractals or singular value decomposition may yield improved fitting results.

EXERCISE 1.1—

1. Compute the Hadamard product of the first 20 Zernikes using the library provided with this book, and display the resulting 20×20 triangular array. How is the orthogonality affected on a non-circular pupil (e.g. annular) ?
2. Perform the Zernike decomposition of the provided wavefront. Given the results of the previous exercise, what would happen if attempting the same for an annular pupil ?

Point spread function

A point of light in the object plane is spread out in the image plane according to the imaging system's impulse response function. If the illumination is spatially incoherent then the impulse response is called the point spread function (PSF), i.e. the PSF is the image of an unresolved object through the system. The PSF can also be understood as the power-spectrum of the normalized pupil field function, i.e.,

$$PSF = \left| \mathscr{F}^{-1}(P_n) \right|^2 = \mathscr{F}^{-1}\left(P_n \star P_n\right) \tag{1.34}$$

where $\star$ denotes correlation,

$$P_n(u,v) = A_n(u,v) \exp i\Phi(u,v), \tag{1.35}$$

and

$$A_n(u,v) = \frac{A(u,v)}{\sqrt{\sum\limits_{u,v} A^2(u,v)}} = \frac{|P(u,v)|}{\sqrt{\sum\limits_{u,v} |P(u,v)|^2}}. \tag{1.36}$$

This normalization ensures that the PSF has unit volume. This is done because the PSF just redistributes the light from each point in the observed object: there is no gain or loss of signal.
A PSF has a number of other noteworthy characteristics:

1. For a circular aperture of diameter D the first zero in the PSF's diffraction intensity pattern occurs at $1.22\lambda/D$. This corresponds to the angular separation at which two point sources become resolved and is known as the Rayleigh resolution limit.[3]

2. The width of the PSF at half its maximum intensity is called the full width at half maximum (FWHM). For a circular aperture the FWHM is given by $1.03\lambda/D$. Note, the peak of a Gaussian profile has a similar shape to that of the Airy function and for a standard deviation $\sigma_G = 0.44\lambda/D$, has the same FWHM.

3. The ratio of the peak intensity of the PSF to the peak intensity of the unaberrated PSF (i.e. the PSF you would get if the aberration phase $\Phi = 0$) is called the Strehl ratio S:

$$S = \frac{\max\{\mathrm{PSF}(\Phi)\}}{\max\{\mathrm{PSF}(\Phi = 0)\}} \leq 1 \tag{1.37}$$

   This ratio provides a simple meaningful way of expressing the effect of wave front aberrations on image quality: the higher the Strehl ratio, the better the image resolution. $S = 1$ corresponds to a diffraction-limited image (the best). If σ – the *rms* value averaged over the wavefront Φ – is known, then we can use the Maréchal formula to approximate the Strehl: $S \simeq \exp(-\sigma^2)$.

4. The PSF integral over the circle of radius β is called the encircled energy. This characteristic is important for observations of faint objects, where one wants to concentrate the photons in as small an area as possible.

Airy pattern

The Airy disc (or Airy pattern) is the best focused spot of light that a perfect lens with a circular aperture can make, limited only by the diffraction of light. Given by the squared modulus of the Fourier transform of the circular aperture Π:

$$I(\theta) = I_0 \left(\frac{2J_1(\frac{2\pi}{\lambda} R \sin\theta)}{\frac{2\pi}{\lambda} R \sin\theta} \right)^2 = I_0 \left(\frac{2J_1(x)}{x} \right)^2 \tag{1.38}$$

where I_0 is the maximum intensity of the pattern at the Airy disc center, J_1 is the Bessel function of the first kind of order one, R is the radius of the aperture, and θ is the angle of observation, i.e. the angle

[3] The angular resolution in arcseconds is given by $\theta = 0.25\lambda/D$, where λ is in microns and D is in meters.

between the axis of the circular aperture and the line between aperture center and observation point. Expressed in linear units, $x = \frac{2\pi R}{\lambda}\frac{r}{d}$, where r is the radial distance from the optics axis in the observation plane and d the observation distance. Note that the limit for $\theta \to 0$ (or for $x \to 0$) is $I(0) = I_0$. We can now check the results from the previous section: the zeros of $J_1(x)$ are at $x \approx 3.8317, 7.0156, 10.1735, 13.3237, \ldots$. From this, it follows that the first dark ring in the diffraction pattern occurs where $\frac{2\pi R}{\lambda}\sin\theta = 3.8317$, or $\sin\theta \approx 1.22\frac{\lambda}{D}$. The next dark rings are at $2.23\frac{\lambda}{D}$ and $3.24\frac{\lambda}{D}$. The half maximum of the central Airy disk (where $J_1(x) = x/2\sqrt{2}$) occurs at $x \approx 1.61633$, leading to the FWHM at $1.028\frac{\lambda}{D}$. The maximum of the first ring occurs at $x \approx 5.13562$, or $1.63\frac{\lambda}{D}$.

The intensity I_0 at the center of the diffraction pattern is related to the total power P_0 incident on the aperture by:

$$I_0 = \frac{E_A^2 R^2}{2d^2} = \frac{P_0 A}{\lambda^2 d^2} \tag{1.39}$$

where E is the source strength per unit area at the aperture, A is the area of the aperture ($A = \pi R^2$) and d is the distance from the aperture. The intensity at the maximum of the first ring is about 1.75% of the intensity at the center of the Airy disk. The expression for $I(\theta)$ above can be integrated to give the total power contained in the diffraction pattern within a circle of given size:

$$P(\theta) = P_0\left[1 - J_0^2\left(\frac{2\pi}{\lambda}R\sin\theta\right) - J_1^2\left(\frac{2\pi}{\lambda}R\sin\theta\right)\right] \tag{1.40}$$

where J_0 and J_1 are Bessel functions. Hence the fractions of the total power contained within the first, second, and third dark rings (where $J_1(\frac{2\pi}{\lambda}R\sin\theta) = 0$) are 83.8%, 91.0%, and 93.8% respectively. [4]

Polychromatic PSF

Equations 1.34 and 1.35 define the PSF for monochromatic light. However, many observations in astronomy are taken through filters with appreciable spectral bandwidth. In this case we need to use the polychromatic point spread function (PPSF), which is obtained by integrating the monochromatic intensities along the sampled spectrum taking into account both the spectral responses of the detector and any filters used, and (ideally) the spectral emission of the source. Practically, this is done by computing the PSF at a series of discrete wavelengths (color planes) and summing across all color planes weighted by the spectral response function, $S(\lambda)$:

$$PPSF = \frac{\sum_j S(\lambda_j) PSF(\lambda_j)}{\sum_j S(\lambda_j)}. \tag{1.41}$$

[4] IMAGE: 3 DIFFERENT APERTURES (CIRCULAR, SQUARE AND HEXAGONAL) AND THEIR CORRESPONDING PSFS.
IMAGE: IMAGE OF POINT SOURCE WITH: NO ABERRATIONS, 1 WAVE PTV SPHERICAL, 1 WAVE PTV ASTIGMATISM, 1 WAVE PTV SPHERICAL+ASTIGMATISM+COMA

EXERCISE 1.2— Generate PSFs for:

1. Circular apertures of radii $R_1 < R_2 < R_3$.
2. Circular apertures with a secondary hole.
3. Circular apertures with a secondary hole and a support spider.
4. A sparse aperture called the Golay-3 pupil, with the following subaperture center (x, y) coordinates: $\left(-\frac{1}{2}, -\frac{\sqrt{3}}{6}\right)$, $\left(\frac{1}{2}, -\frac{\sqrt{3}}{6}\right)$ and $\left(0, \frac{\sqrt{3}}{3}\right)$.
5. A sparse aperture called the Golay-6 pupil, with the following subaperture center (x, y) coordinates: $\left(1, \frac{\sqrt{3}}{3}\right)$, $\left(\frac{3}{2}, -\frac{\sqrt{3}}{6}\right)$, $\left(0, -\frac{2\sqrt{3}}{3}\right)$, $\left(-1, -\frac{2\sqrt{3}}{3}\right)$, $\left(-1, \frac{\sqrt{3}}{3}\right)$, $\left(-\frac{1}{2}, \frac{5\sqrt{3}}{6}\right)$.
6. Look at different PSFs for the first 12 Zernikes.
7. Calculate Strehl for the previous PSFs. Be mindful of the normalization of the unaberrated pupil and of the normalization of the Fourier transform functional.

1.3 Image formation

Imaging is convolving

By virtue of the linearity property of optical imaging systems, the image of the superposition of two objects O_1 and O_2 is the superposition of the two images:

$$I(O_1 + O_2) = I(O_1) + I(O_2). \tag{1.42}$$

Since the object-plane field is itself a weighted sum over 2D impulse response functions[5],

$$O(x_0, y_0) = \sum_{(u,v)} O(u,v)\, \delta(x_0 - u, y_0 - v) \tag{1.43}$$

the image plane field is the weighted sum over the images of these impulse response functions,

$$I(x_i, y_i) = \sum_{(u,v)} O(u,v)\, \mathrm{PSF}(x_i - u, y_i - v) \tag{1.44}$$

With the following definition for the discrete 2D convolution:

$$\begin{aligned}(f * g)[n,m] &= \sum_p \sum_q f[p,q]\, g[n-p, m-q] \\ &= \sum_p \sum_q f[n-p, m-q]\, g[p,q]. \end{aligned} \tag{1.45}$$

we realize the image is given as the convolution of the object by the PSF:

$$I = O * PSF \tag{1.46}$$

Fast convolution

The *convolution theorem* states that the Fourier transfor$\mathscr{F}\{f * g\} = \mathscr{F}\{f\} \cdot \mathscr{F}\{g\}$, where $\cdot$ denotes the Hadamard multiplication. This also works the other way around: $\mathscr{F}\{f \cdot g\} = \mathscr{F}\{f\} * \mathscr{F}\{g\}$. A major application of this theorem is found in the computation of the image as a function of object and PSF:

$$I = O * \mathrm{PSF} = \mathscr{F}^{-1}\{\mathscr{F}\{O\} \cdot \mathscr{F}\{\mathrm{PSF}\}\} \tag{1.47}$$

which tends to be faster than the direct calculation from equation 1.44.

Optical transfer function

The optical transfer function (OTF) of an optical system specifies how different spatial frequencies are transmitted through the system. The optical transfer function is the Fourier transform of the point spread function. When writing the convolution equation in Fourier space, one can see that the spatial frequencies of the image are those of the object filtered by the OTF of the system:

---

[5] In general, a signal can be decomposed as a weighted sum of basis signals. For example, in a Fourier series, any periodic signal, even a rectangular pulse, can be represented by a sum of sine and cosine functions. However, a signal can also be represented as a sum of scaled and shifted impulse (delta) functions. As a demonstration, consider the 1D signal $x(0) = 1, x(1) = 3, x(2) = 2$. Since the impulse function, $\delta(n)$ is 1 at $n = 0$, and 0 at $n \neq 0$, then the signal at $x(1)$ can be written as $3\delta(n-1)$ because $\delta(n-1)$ is 1 at $n = 1$ and 0 elsewhere. Similarly the signals at $x(0)$ and $x(2)$ can be written as $1\delta(n-0)$ and $2\delta(n-2)$). We can thus write the entire signal as a sum of scaled and shifted delta functions, i.e. $x(n) = \sum_k x(k)\delta(n-k)$.

$$\mathscr{F}\{I\} = \mathscr{F}\{O\} \cdot \mathrm{OTF}, \tag{1.48}$$

meaning that the only spatial frequencies from the object that will be kept in the image are the ones which will be non-zero in the OTF. Since the point spread function is the square absolute of the inverse Fourier transformed pupil function, the optical transfer function can also be calculated directly from the pupil function: it is in fact the auto-correlation of the pupil function[6], i.e.,

$$\mathrm{OTF} = P \star P \tag{1.49}$$

where the discrete 2D cross-correlation is given by:

$$(f \star g)[n, m] \;= \sum_p \sum_q f^*[p, q]\; g[p+m, q+n]. \tag{1.50}$$

The OTF is in general a complex function having both a magnitude and a phase portion. The former is referred to as the modulation transfer function (MTF), and the latter the phase transfer function (PTF). The MTF is the magnitude response of the optical system to sinusoids of different spatial frequencies. As such, it provides a useful measure of true or effective resolution, since it accounts for the amount of blur and contrast over a range of spatial frequencies.

Given the shape of the pupil, its MTF can be estimated without computer by doing the autocorrelation by hand as shown on Figure 1.3 for a simple single aperture: simply by sliding the pupil onto of itself, looking for the non-zero intersection, and summing the surface of the intersecting area. The shape of the autocorrelation peaks in the MTF is described by the following equation as a function of the radial frequency u:

$$\mathrm{OTF}(u) = \frac{2}{\pi}\left(\arccos(|u|) - |u|\sqrt{1-u^2}\right). \tag{1.51}$$

Note that frequencies higher than D/λ are absent from the MTF in this case.

Figure 1.3 presents the slightly more complex situation of a pupil with two sub-apertures. In this case the MTF presents two half-strength peaks surrounding the main peak.

1.4 *Sparse aperture characteristics and OTF*

Sparse-aperture pupils are often characterized by their *fill factor*, which is the ratio of the area of a given aperture to the area of a filled aperture

---

[6] The autocorrelation function and the power spectrum are a Fourier transform pair.

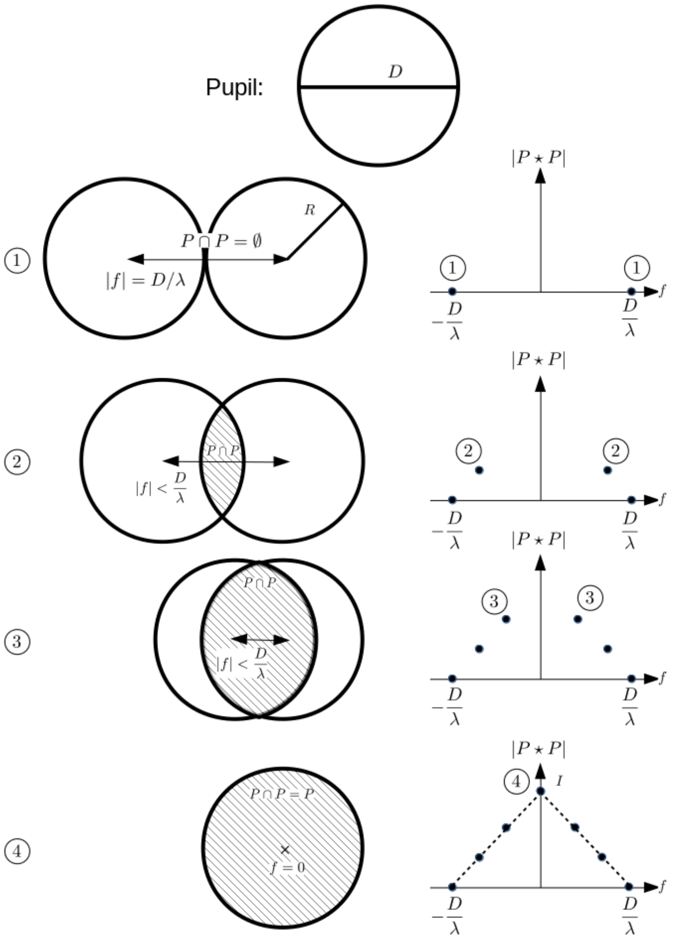

Pupil:
D
R
$P \cap P = \emptyset$
$|f| = D/\lambda$
$|f| < \frac{D}{\lambda}$
$P \cap P$
$P \cap P = P$
$f = 0$
I
$|P \star P|$
f
$-\frac{D}{\lambda}$
$\frac{D}{\lambda}$
①
②
③
④

Figure 1.6: How to build "by hand" the autocorrelation of a single aperture pupil P. For simplicity we assume here the pupil is uniform and real-valued (zero phase). We apply Equation 1.50 by sliding the pupil on top of itself (here from left to right), and computing the integral of the overlapping surface (hashed). Note that since the pupil is point-symmetric, so is the MTF. We obtain the well-known result that the frequency cutoff is $\frac{D}{\lambda}$.

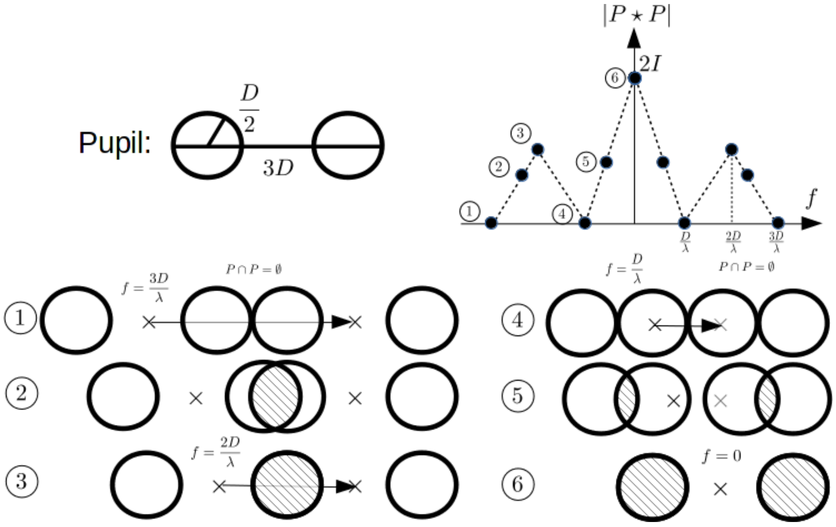

Pupil:
$\frac{D}{2}$
$3D$
$|P \star P|$
$2I$
f
$\frac{D}{\lambda}$
$\frac{2D}{\lambda}$
$\frac{3D}{\lambda}$
$f = \frac{3D}{\lambda}$
$P \cap P = \emptyset$
$f = \frac{D}{\lambda}$
$P \cap P = \emptyset$
$f = \frac{2D}{\lambda}$
$f = 0$
①
②
③
④
⑤
⑥

Figure 1.7: How to build "by hand" the autocorrelation of a pupil with two subapertures. The same principles as for Figure 1.3 apply. This pupil is not point-symmetric, so the shape of the MTF slide depends on the sliding direction: this is obvious since this pupil will provide better resolution along the horizontal axis than the vertical one.

having comparable theoretical resolution. Pupils with low fill factor are said to be *diluted* pupils; conversely, pupils with high fill factor are said to be *compact* pupils.

Non-redundant apertures consists of sub-apertures positioned so that no two pairs of sub-apertures have similar separation vectors, i.e. so that the different pairs do not give rise to overlapping peaks in the OTF. *Redundant* apertures have overlapping frequencies in the OTF.

A desirable feature for sparse aperture imaging is to enhance resolution while minimizing the total light collection area (since large mirrors are expensive). Hence, good non-redundant apertures such as the Golay family of pupils cover the spatial frequency/OTF plane as much as possible without leaving holes (zeros): they are non-redundant yet maximally compact pupils.

1.5 *Atmosphere and turbulence*

The wavefront from an astronomical source can be considered flat at the top of the atmosphere. But as it propagates to the ground the optical turbulence corrupts its phase and amplitude, affecting measurements from ground-based telescopes. The perturbations of the optical path length $l(\vec{x})$ are considered as achromatic[7]. However, the phase of the optical wave $\phi(\vec{x}) = \frac{2\pi}{\lambda} l(\vec{x})$ strongly depends on the wavelength λ.

Optical turbulence is caused by the mechanical mixing of layers of air with different temperatures and hence density. As the refractive index of air changes with density this turbulence creates a continuous screen of spatially and temporally varying refractive indices. Although each of the refractive index inhomogeneities in the turbulent layers may be small the wavefront passes through a large number of them and the cumulative effect can be quite large. The cumulative refractive index variations delay parts of the incoming wavefront with respect to others. The net effect is that the wavefront becomes aberrated. If we assume a horizontal turbulent layer at altitude, h, above the ground and that the layer thickness, δh, is large compared to the eddy size of the refractive index inhomogeneities, but small enough so that we can ignore diffraction effects within the layer (thin screen approximation), then the phase fluctuations, $\phi(r)$, induced by the turbulent layer is related to the refractive index fluctuations, $n(h, r)$, along the propagation path by

$$\phi(r) = k \int\limits_{h}^{h+\delta h} n(h, r) dh \tag{1.52}$$

where k is the wave number, $2\pi/\lambda$, with λ being the wavelength of the light and r is a spatial coordinate. The wave function after the layer is then,

---

[7] Air is slightly dispersive, but this is usually neglected.

$$\psi(r) = \chi(r) \exp i\phi(r) \propto \exp i\phi(r) \tag{1.53}$$

It is these aberrations in the wavefront which act to distort images from astronomical telescopes. Now the refractive index fluctuations are random and the spatial structure of a random process is given by *structure functions*. The structure function $D_f(r_1, r_2)$ of a random variable f measured at positions $r_1,\ r_2$ is defined by

$$D_x(r_1, r_2) = \left\langle \left| f\left(r_1\right) - f\left(r_2\right) \right|^2 \right\rangle, \tag{1.54}$$

where the angled brackets, $\langle ... \rangle$, represent the spatial average. That is, the structure function measures the expectation value of the difference of the values of f measured at two positions, r_1 and r_2. With this in mind, we are therefore not interested in the absolute value of the phase, only the difference between its value at two points, which is caused by the spatial variance of the refractive index. The refractive index structure function, $D_n(\rho)$, is thus the spatial variance in the difference of refractive index n as a function of separation ρ.

$$D_n(\rho) = \left\langle \left| n\left(r\right) - n\left(r+\rho\right) \right|^2 \right\rangle. \tag{1.55}$$

The Kolmogorov model of turbulence

A description of the nature of the wavefront perturbations introduced by the atmosphere is provided by the Kolmogorov model developed by Tatarski, based partly on the studies of turbulence by the Russian mathematician Andrei Kolmogorov. This model is supported by a variety of experimental measurements and is widely used in simulations of astronomical imaging. The model assumes that the wavefront perturbations are brought about by variations in the refractive index of the atmosphere. In the Kolmogorov model of turbulence the refractive index structure function is given by

$$D_n(\rho) = C_n^2\, \rho^{2/3} \quad \text{for } \ l_0 < \rho < L_0 \tag{1.56}$$

where the quantity C_n^2 is known as the refractive index structure coefficient (see section below on C_n^2) and can be considered to be a measure of the small-scale turbulence, and l_0 and L_0 are the inner and outer scales of turbulence, respectively. The former is the size below which viscous effects and energy is dissipated into heat. The latter is the size above which isotropic behavior is violated. We note that the inner scale size of turbulence l_0 varies inversely with C_n^2 and has values that range in size from 1 - 2 mm up to 1 cm or more near the ground.

The power spectral density for refractive index fluctuations over inertial subrange ($1/L_0 < \kappa < 1/l_0$) in three dimensions is defined by

$$P_n(\kappa) = 0.033\, C_n^2\, \kappa^{-11/3} \tag{1.57}$$

where κ is the spatial frequency.

The refractive index variations lead directly to phase fluctuations described by $\phi(\mathbf{r})$, but any amplitude fluctuations are only brought about as a second-order effect while the perturbed wavefronts propagate from the perturbing atmospheric layer to the telescope. For all reasonable models of the Earth's atmosphere at optical and infra-red wavelengths the instantaneous *imaging* performance is dominated by the phase fluctuations $\phi(\mathbf{r})$. The amplitude fluctuations described by $\chi(\mathbf{r})$ have negligible effect on the structure of the images seen in the focus of a large telescope, but will of course affect *photometric* measurements.

As the phase perturbations are random, we use the phase structure function to describe the spatial structure of the phases, i.e.

$$D_\phi(|\rho|) = \left\langle |\phi(\mathbf{r}) - \phi(\mathbf{r}+\rho)|^2 \right\rangle_{\mathbf{r}} \tag{1.58}$$

where $D_\phi(\rho)$ is the atmospherically induced variance between the phase at two parts of the wavefront separated by a distance ρ in the aperture plane, and $\langle ... \rangle$ represents the ensemble average.

The power spectral density for phase fluctuations is given by

$$P_\phi(\kappa) = |\mathscr{F}(\phi)|_\kappa^2 = 0.023\, r_0^{-5/3}\, \kappa^{-11/3}. \tag{1.59}$$

where r_0 is the Fried parameter (described in section below).

C_n^2 *profile*

A more thorough description of the astronomical seeing at an observatory is given by producing a profile of the turbulence strength as a function of altitude h, called a C_n^2 profile.

The model most often used is the generalized Hufnagel-Valley (HV) model:

$$\begin{aligned} C_n^2(h) &= A \exp(-h/H_A) + B \exp(-h/H_B) \\ &\quad + C\, h^{10} \exp(-h/H_C) \\ &\quad + D \exp\left(-(h-H_D)^2/2d^2\right) \end{aligned} \tag{1.60}$$

Here A is the coefficient for the surface (boundary layer) turbulence strength ($\mathrm{m}^{-2/3}$) and H_A is the height for its *1/e* decay (meters), B and H_B similarly define the turbulence in the troposphere (up to about 10 km), C and H_C define the turbulence peak at the tropopause, and D and H_D define one or more isolated layers of turbulence, with d being the layer thickness (meters). The coefficients for the HV(5-7) model are[8]: $A = 1.7 \times 10^{-14}$ $\mathrm{m}^{-2/3}$, $H_A = 100$ m, $B = 2.7 \times 10^{-16}$ $\mathrm{m}^{-2/3}$, $H_B = 1500$ m, $C = 3.59 \times 10^{-53}$ $\mathrm{m}^{-2/3}$ $H_C = 1000$ m, $D = 0$. This model highlights the fact that the dominant locations for index of refraction fluctuations that affect astronomers are the atmospheric boundary layer and the troposphere. However, each observing site will have its own set of typical values for the coefficients.

Values of C_n^2 of $\sim 10^{-17}\mathrm{m}^{-2/3}$ or less indicate weak turbulence, while values of $\sim 10^{-13}\mathrm{m}^{-2/3}$ or more indicate strong turbulence.

C_n^2 profiles are generally measured when deciding on the type of adaptive optics system which will be needed at a particular telescope, or in deciding whether or not a particular location would be a good site for setting up a new astronomical observatory. Typically, several methods are used simultaneously for measuring the C_n^2 profile and then compared.

[9]

Fried parameter: r_0

For simplicity, the phase fluctuations in Tatarski's model are often assumed to have a Gaussian random distribution. This is because for atmospheric layers that are much thicker than the turbulence cells, many independent variables contribute to the phase shift (see Equation 1.52). Therefore the Central Limit Theorem implies that ϕ has Gaussian statistics. In this case the structure function can be described in terms of a single parameter r_0, i.e.,

$$D_\phi(\rho) = 6.88 \left(\frac{|\rho|}{r_0} \right)^{5/3} . \tag{1.61}$$

The parameter r_0, which is often known as the Fried parameter, named after David L. Fried, indicates the strength of the phase fluctuations and is frequently used to describe the atmospheric conditions at astronomical observatories. Although often regarded as a constant, r_0 is a statistical parameter that varies considerably over short periods of time, sometimes changing by a factor of 2 within a few seconds. [10] r_0 can be determined from a measured C_n^2 profile using:

$$r_0 = \left[0.423\, k^2 (\cos\zeta)^{-1} \int_0^\infty C_n^2(h) dh \right]^{-3/5} \tag{1.62}$$

where the turbulence strength $C_n^2(h)$ varies as a function of height h above the telescope, and ζ is the angular distance of the astronomical source from the zenith. The key points of equation (1.62) are that the optical effects of turbulence are less (i.e. r_0 is larger) at smaller zenith angles and at longer wavelengths. Equation (1.62) shows that r_0 varies with wavelength as $\lambda^{6/5}$ since $k = \frac{2\pi}{\lambda}$. We note the standard wavelength for reporting r_0 values is 0.5 μm.

[8] Hardy, 1998

[9] IMAGE: C_N^2 PROFILE

[10] IMAGE: Graph of r_0 verses time showing variability on short time scales, as well as r_0 spread during an observing night

A commonly used definition for r_0 is that it corresponds to the aperture diameter D for which the variance σ^2 of the wavefront phase averaged over the aperture comes approximately to unity:

$$\sigma^2 = 1.0299 \left(\frac{D}{r_0}\right)^{5/3} \tag{1.63}$$

Equation (1.63) shows that r_0 corresponds to the length-scale over which the phase perturbations due to atmospheric turbulence becomes significant (15 to 20 cm at visible wavelengths at good observatories).

We note that r_0 is often referred to as the characteristic scale of atmospheric turbulence. This is not quite true as the Kolmogorov law does not have any characteristic scale. However, only the perturbations of order r_0 are relevant for long-exposure imaging (see section 1.6).

Atmospheric time constant: τ_0

A second parameter that is used to describe the turbulence conditions at a site is the atmospheric time constant τ_0. This parameter, which is also know as the atmospheric coherence time and the Greenwood time constant (after Darryl Greenwood), corresponds to the time-scale over which the changes in the turbulence become significant. This occurs when the wavefront phase decorrelates by ~ 1 radian RMS). If the evolution of the atmosphere is slow enough that it can be considered "frozen" at short time scales, and the turbulence is just advected across the aperture by the wind in the atmosphere (the Taylor Hypothesis), then τ_0 is simply proportional to r_0 divided by the mean wind speed, i.e.

$$\tau_0 = 0.314\, r_0 / \bar{V} \tag{1.64}$$

where

$$\bar{V} = \left[\frac{\int C_n^2(h)|V(h)|^{5/3} dh}{\int C_n^2(h) dh}\right]^{3/5}. \tag{1.65}$$

Equation (1.65) shows that the main contribution to $\bar{V}$ is from high-velocity winds where C_n^2 can be large.

Images taken with exposure times $\leq \tau_0$ are called "short-exposure" images and their phase aberrations are considered frozen (fixed). At longer exposure times the aberrations are averaged and for exposure times $>> \tau_0$ we obtain the long-exposure PSF. Like r_0, τ_0[11] varies with wavelength as $\lambda^{6/5}$.

Isoplanatic angle: θ_0

The viewing angle over which the wave front error can be considered constant is called the *isoplanatic* angle.

[11] r_0 determines the spacing of the actuators needed in an adaptive optics system, and τ_0 determines the correction speed required to compensate for the effects of the atmosphere.

However, atmospheric turbulence is continuously changing and the instantaneous phase aberrations observed in the aperture therefore change with the viewing direction. That is, when beams propagate in different directions through a series of turbulence layers, each beam will sample slightly different turbulence. Propagation through turbulence is therefore, in general, *anisoplanatic*.

The isoplanatic angle is given by

$$\theta_0 = \left[2.91\, k^2\ (\sec\zeta)^{8/3}\ \int_0^\infty C_n^2(h) h^{5/3} dh\right]^{-3/5} . \tag{1.66}$$

A simpler expression, which is often used, is

$$\theta_0 \simeq 0.314\ \left(\frac{r_0}{\bar{h}}\right), \tag{1.67}$$

where

$$\bar{h} = \sec\zeta\ \left[\frac{\int h^{5/3} C_n^2(h) dh}{\int C_n^2(h) dh}\right]^{3/5} . \tag{1.68}$$

We note that the isoplanatic angle is hugely variable and in extreme turbulence situations can approach the diffraction limit of the telescope.[12] Like r_0 and τ_0, θ_0 varies with wavelength as $\lambda^{6/5}$.

Layer model for atmosphere

For simplicity we commonly approximate the C_N^2 profile by a number of layers where each layer propagates in the direction of the local wind velocity and is characterized by a turbulence strength that is constant within the layer. In addition, the thickness of a layer is assumed to be large compared to the scale size of the turbulence cells, so that Gaussian statistics apply, but still small enough for diffraction effects to be negligible.

The Fried parameter for an individual turbulence layer r_{0_i} is related to the overall Fried parameter r_0 by

$$r_0^{-5/3} = \sum_{i=1}^{N} r_{0_i}^{-5/3}. \tag{1.69}$$

Scintillation

When the propagation path length L from a turbulent layer is comparable to, or longer than, the Fresnel propagation length

$$L_F = r_0^2/\lambda\ , \tag{1.70}$$

the rays diffracted at the turbulence cells in the layer interfere with each other which causes intensity variations, or *scintillation*, in the

[12] For angular separations smaller than the isoplanatic angle, the variance between difference paths is small and we can treat the PSF as spatially invariant and model the image as a convolution of the object and PSF. However, for angular separations larger than the anisoplanatic model, the PSF varies spatially across the field-of-view (FOV) and the convolution model is only approximate, with the approximation becoming worse as the FOV increases.

pupil plane in addition to phase variations. For turbulence conditions of r_0 = 20 cm and $\lambda = 0.5\mu$m, the Fresnel propagation length L_F = 80 km. This is significantly larger than the height of the layers that provide the major contributions to the C_N^2 integrals, and scintillation effects are negligible on imaging compared to phase effects as long as the zenith angle is not too large.

The characteristic size of the scintillation speckles is given by the Fresnel scale:

$$r_F = \sqrt{L\lambda}. \tag{1.71}$$

As a wavefront propagates away from a turbulent layer, increasing L and hence r_F, the spatial intensity fluctuations become larger in terms of spatial extent. This is not dependent on the strength of the layer, which only affects the magnitude of the intensity fluctuations and not their spatial properties; for example, a turbulent layer at 10 km observed at 500 nm results in speckles of size ~ 7 cm, irrespective of the strength of the turbulence.

Equation (1.70) is equivalent to saying when the Fresnel scale r_F is commensurate with or larger than the Fried scale r_0 we have to consider the effects of scintillation. As $L = h_L / \cos z$, where h_L is the height of the turbulent layer, it is primarily the high altitude turbulent layers in the atmosphere and observations at high zenith angles that contribute to the intensity variations of the wave front. The resulting 'scintillation noise' is what limits the precision of photometric measurements.

The power spectral density for scintillation from a turbulence layer i at height h_i and thickness Δh_i in the atmosphere is given by

$$S_\chi(\kappa) = 1.54\, \lambda^2\, \kappa^{-11/3}\, C_n^2\, \Delta h_i \times \sin^2\left[\pi\lambda h_i \kappa^2\right]. \tag{1.72}$$

The effects of scintillation can be quantified by determining the variance of intensity fluctuations $\sigma_I^2 = \frac{\langle I^2\rangle - \langle I\rangle^2}{\langle I\rangle^2}$. As the scintillation noise is caused by the intensity speckles entering and leaving the telescope pupil there are three regimes of scintillation noise: short exposures and longer exposures on medium/large telescopes ($D > r_F$, so larger than a few tens of centimeters), and the specific case of small telescopes.

- For short exposures on medium/large telescopes, the intensity speckles appear frozen in the pupil and no temporal averaging occurs. The expression of the variance is shown to be [13]:

$$\sigma_I^2 = 17.34\, d^{-7/3}\, (\cos z)^{-3} \int C_n^2(h) h^2 dh\ . \tag{1.73}$$

  where d is the diameter of the telescope. Note how this expression shows the expected strong decrease of scintillation amplitude with aperture size and that it is independent of the observing wavelength.

[13] Kenyon et al., 2006

- For longer exposures on medium/large telescopes, the speckles traverse the pupil during an exposure and the scintillation noise will be reduced by temporal averaging. The amount by which the scintillation noise is reduced is dependent on the wind speed:

  $$\sigma_I^2 = 10.66\, d^{-4/3} t^{-1} \left(\cos z\right)^{\alpha} \int \frac{C_N^2(h) h^2}{V_\perp(h)} dh\,, \tag{1.74}$$

  where $V_\perp(h)$ is the wind velocity profile and α the airmass exponent (taken to be ~ -3.5 in practice). Here again scintillation amplitude decreases with aperture size, and no wavelength dependency is apparent.

- For telescopes smaller than a few tens of centimeters ($D < r_F$) there will be a significant wavelength dependence since the size of the speckles is itself wavelength-dependent.[14]

  $$\sigma_I^2 = 19.2\, \lambda^{-7/6} \left(\cos z\right)^{-11/6} \int C_N^2(h) h^{5/6} dh\,, \tag{1.75}$$

Another way to look at scintillation is as a fluctuation in the spatial intensity distribution in the pupil caused by the random turbulent atmosphere locally focusing and defocussing the wave front. Scintillation is therefore a second order effect as it depends on the curvature (i.e., second derivative) of the wave front, whereas phase perturbations in the pupil plane are a first-order effect as they depend on the first derivative of the wave front.

Finally, Young's approximation is a semi-empirical formula that gives an estimation of the scintillation noise variance:

$$\sigma^2 = 10^{-5} C^2 d^{-\frac{4}{3}} t^{-1} \left(\cos z\right)^{-3} \exp\left(-2h_{obs}/H\right) \tag{1.76}$$

where d is the diameter of the telescope, t is the exposure time of the observation, z is the zenith distance, h_{obs} is the altitude of the observatory and H the scale height of the atmospheric turbulence, which is generally accepted to be approximately 8000 m. C has been empirically measured at several major observatories to have a mean value of $\simeq 1.5$ in SI units [15].

1.6 Imaging through atmospheric turbulence

The atmosphere has two distinct states of motion: laminar and turbulent. Mixing of the air does not occur in laminar flow but turbulent flow is characterized by dynamic mixing and it acquires random subflows called eddies. The different size eddies cause different imaging effects. Eddies larger than the beam size cause tilt variations (beam wander) while eddies smaller than the beam size cause it to break up as different sections of the beam are refracted in different directions (beam spread). That is, the atmosphere can be visualized as a path of continually changing lenses refracting light as it travels to the telescope.[16] In other words, the atmospheric turbulence may be regarded as random phase aberrations added to the telescope. These aberrations constantly change in time, and consequently so does the PSF.

---

[14] Dravins et al., 1998

[15] Osborn et al., 2015

The wave front variance for beam spread is found by removing the contribution from the two-axis tilt variations from equation (1.63) and is given by

$$\sigma^2_{\text{tiltcomp}} = 0.134 \left(\frac{D}{r_0}\right)^{5/3} \tag{1.77}$$

This is the origin of the common statement that "tip-tilt is the largest contributor to wave front error". However, if (D/r_0) is large, removing tip/tilt won't yield anywhere near a diffraction-limited image.

Equations (1.63) and (1.77) show that the ratio D/r_0 provides a measure of the strength of the atmospheric turbulence. When the quantity D/r_0 is a few (small) we have low turbulence. When D/r_0 is large we have strong turbulence.

Let's now take a look at the morphology of a short-exposure, turbulence degraded, image. If we consider r_0 as being roughly the average physical dimension of a single isoplanatic "patch", and a patch gives rise to a diffraction-limited "speckle" in the PSF of width $\sim \lambda/D$, then the total number of speckles in the short-exposure PSF from all the patches in the telescope pupil of diameter D is $\sim (D/r_0)^2$ and they are spread over an area of width $\sim \lambda/r_0$. The latter is known as the "seeing" disk and its location depends on the instantaneous tip/tilt perturbation in the wave front. In summary, there are two components that contribute to the morphology of a short-exposure turbulence-degraded image: the diffraction limit of the aperture which depends on λ/D, and the short-exposure image spread which depends on λ/r_0. For long-exposure images there is a third component related to the image displacement caused by non-zero tip/tilt aberrations in the atmospheric wave front. This component depends on both D and r_0.[17]

Long-exposure images

Long-exposure images are images that are acquired over exposure times that are long compared to the atmospheric coherence time τ_0. The long-exposure PSF, which has the three components described at the end of the last section, can (are) modeled using two functions representing a "core" and a "halo". Roughly speaking when $\lambda > 20\ \sigma_{OPD}$ we have "core-only" PSFs. Here σ_{OPD} is the rms variation of the optical path difference. When $\lambda < 4\ \sigma_{OPD}$ we have "halo-only PSFs", and when $4\ \sigma_{OPD} \leq \lambda \leq 20\ \sigma_{OPD}$ we have PSFs with "cores and halos".

---

[16] This is only illustrative: there are no discontinuities in the atmosphere.

[17] IMAGE: TWO ROWS WITH INSTANTANEOUS "SPECKLE" PSFS IN THE TOP ROW AND THE LONG-EXPOSURE (500 IMAGES) PSFS IN THE BOTTOM ROW. FOUR COLUMNS FOR Ro OF 50cm, 20cm, 10cm, AND 5cm WITH A TELESCOPE OF APERTURE D=100 cm

We'll start with looking at the cases where the PSFs display a "core".

The mean-square deflection angle, or angle of arrival, due to tip/tilt variations has a Gaussian distribution with a variance given by

$$\alpha^2 = 0.18 \left(\frac{\lambda}{D}\right)^2 \left(\frac{D}{r_0}\right)^{5/3}. \tag{1.78}$$

Since $r_0 \propto \lambda^{6/5}$ then the mean-square deflection angle varies as $\sim D^{-1/3}$ and is independent of λ. However, relative to the Airy disk, image motion gets worse for larger D and smaller λ, i.e.

$$\frac{\alpha}{\lambda/D} = 0.424 \left(\frac{D}{r_0}\right)^{5/6} \propto \frac{D^{5/6}}{\lambda}. \tag{1.79}$$

The effect of a random tilt error α is to spread each point of an image into a Gaussian profile with standard deviation α. Therefore the FWHM of the central core of a long-exposure PSF, including the tilt error, can be approximated by

$$\begin{aligned} FWHM_{core} &= 2.354 \sqrt{\alpha^2 + 0.1936\,(\lambda/D)^2} \\ &= 1.03 \left(\frac{\lambda}{D}\right) \sqrt{0.93\,(D/r_0)^{5/3} + 1} \end{aligned} \tag{1.80}$$

where the second term represents the variance of a Gaussian profile with the same FWHM as an Airy function (see section 1.1).

Equation (1.80) shows that for $D/r_0 \sim 1$ or less, the core of the PSF is close to diffraction-limited and the long-exposure degradation to the image is caused by image motion (wander).

Next we move to the "halo-only" image case. As the ratio D/r_0 increases, both the image motion and *the number of speckles* increase and the FWHM of the core of the long-exposure PSF approaches that of the halo from the short-exposure image spread, [18] i.e.,

$$FWHM_{Halo} = 0.98\lambda/r_0. \tag{1.81}$$

Thus, for telescopes with diameters $D \sim r_0$ or smaller, the resolution of long-exposure images is determined by diffraction and is given by $\sim \lambda/D$. For telescopes with D greater than r_0, the long-exposure image resolution is given by $\sim \lambda/r_0$. That is, the resolution is determined primarily by the atmosphere and is independent of telescope diameter, remaining constant at the value given by a telescope of diameter equal to r_0.[19] We note that the long-exposure PSF is independent of the viewing direction (i.e. it is isoplanatic) because the turbulence and its structure function are statistically the same everywhere in the field.

[18] The FWHM of the long-exposure PSF is known in ground-based astronomy as the "seeing" parameter $\beta(\lambda) = 0.98\lambda/r_0$. In arcseconds, $\beta(\lambda) \simeq 20\lambda/r_0$. where λ is in microns and r_0 is in centimeters.

[19] Due to the wavelength dependence of r_0, the "seeing" depends on $\lambda^{-1/5}$ and we can obtain slightly higher-resolution long-exposure imagery by observing at longer wavelengths when using large telescopes.

Atmospheric transfer function

For long exposure images, theory leads to the expression.

$$\mathrm{OTF}_{\mathrm{LongExposure}} = \mathrm{OTF}_{\mathrm{Tel}} \cdot \mathrm{OTF}_{\mathrm{Atm}} \tag{1.82}$$

For large telescopes with good optical quality the resolution is entirely defined by the atmosphere, so we neglect the first term and $OTF_{\mathrm{LE}} \approx OTF_{\mathrm{Atm}}$. The atmospheric OTF is related to the statistics of atmospheric phase aberrations through the phase structure function $D_\phi(\vec{r})$:

$$OTF_a(\vec{f}) = \exp[-0.5 D_\phi(\lambda \vec{f})]. \tag{1.83}$$

Speckle imaging

As discussed at the beginning of this section, the instantaneous atmospheric PSF consists of a number of "speckles" that depends on the level of turbulence as measured by the ratio D/r_0. Thus, imaging at frame rates that are commensurate with, or less than, the atmospheric coherence time is referred to as *speckle imaging*. The main advantage of speckle imaging is that each frame contains spatial information out to the diffraction limit of the telescope. Speckle imaging can be divided into two approaches: "shift and add" (imaging stacking) and speckle interferometry.

Shift-and-add

One approach to recovering the high spatial-frequency information in a set of speckle images is to shift all the short-exposure images to a common center-of-mass and then average them. This procedure is effectively removing the image motion due to the tip/tilt components of the atmospheric wave front. However, it has been found that a better approach which provides a higher Strehl ratio, at least for images of objects with point-like structures, is to align the images to the brightest speckle and then average.

For low levels of turbulence the best resolution is provided by a variation of the shift-and-add approach, known as *"lucky imaging"*. Since atmospheric turbulence is a random process there are moments in time when the seeing is significantly better or worse than average. If we only select the speckle images where the seeing is essentially diffraction-limited ("lucky images", we can produce a better image than would be possible by combining all the images without frame selection.

The probability that we can observe an image with diffraction-limited resolution is given by:

$$\mathrm{P}_{\mathrm{Lucky}} \simeq 5.6\ \exp\left[-0.1557\,(D/r_0)^2\right] \ \text{for} \quad D/r_0 \geq 3.5 \tag{1.84}$$

This shows that for turbulence with $D/r_0 = 6$ there is a 1 in 50 chance of acquiring an image with diffraction-limited resolution. For higher levels of turbulence, the probability diminishes quickly: at $D/r_0 = 10$ the probability is reduced to $\sim$1 in 1,000,000. That is, lucky imaging is a technique that only feasible for relatively benign turbulence conditions with $D/r_0 \leq 7$.
20

Speckle Interferometry

Speckle interferometry yields the high resolution autocorrelation of the object and is a popular technique for studying binary stars.

Since $I = O * \mathrm{PSF}$ and the PSF irradiance varies randomly while the object irradiance remains constant, then

$$\left\langle |\mathscr{F}\{I\}|^2 \right\rangle = \left\langle |\mathscr{F}\{O \cdot PSF\}|^2 \right\rangle = |\mathscr{F}\{O\}|^2 \cdot \left\langle |\{\mathscr{F}\{PSF\}|^2 \right\rangle \qquad (1.85)$$

where $< .. >$ denotes the ensemble average. Thus the autocorrelation of the object is computed from the recorded speckle data in two steps. First, we average the modulus square of the Fourier transforms (i.e. the power spectra) of all the speckle images and then compensate the resulting spectrum for the transfer function of the atmosphere and instrument. The latter can come from either a model or (better) from a time series of speckle images of an unresolvable star close to the object. This step yields the power spectrum of the object, what Michelson observed as "visibility".

Second, we take the inverse Fourier transform of the corrected ensemble averaged power spectrum to provide the autocorrelation function of the object with a resolution limited only by diffraction, not anymore by the turbulent atmosphere. That is,

$$\mathrm{ACF} = \mathscr{F}^{-1}\{\left\langle |\mathscr{F}\{I\}|^2 \right\rangle / \left\langle |\mathscr{F}\{PSF\}|^2 \right\rangle\} = \mathscr{F}^{-1}\{|\mathscr{F}\{O\}|^2\} \qquad (1.86)$$

Note that the ratio in equation (1.86) can only be taken in regions of frequency space where $\left\langle |\mathscr{F}\{PSF\}|^2 \right\rangle$ is non-zero and has signal-to-noise greater than unity. The quantity $\left\langle |\mathscr{F}\{PSF\}|^2 \right\rangle$ is referred to as the speckle transfer function. The speckle transfer function relates the modulus squared of the object spectrum to the second moment of the image spectrum. The speckle transfer function is finite out to spatial frequencies approaching the diffraction-limited cut-off frequency, and hence allows diffraction-limited information to be measured.
21

If we want to reconstruct an actual image of the object from the recorded speckle images, then we can do so if the object and the point source are in the same isoplanatic patch, and they are imaged at the

[20] IMAGE: AN EXAMPLE RAW IMAGE OF A BINARY STAR ACQUIRED THROUGH $D/R_0 = 5$ TURBULENCE, THE AVERAGE OF 500 IMAGES TURBULENCE, THE SAA IMAGE BASED ON CENTER-OF-MASS, THE SAA IMAGE BASED ON BRIGHTEST SPECKLE, AND THE SAA IMAGE BASED ON THE TOP 5% LUCKY IMAGES.

[21] IMAGE: AVERAGE POWER SPECTRUM OF DATA USED IN SAA EXAMPLE, ACF OF DATA, ACF OF RAW DATA CORRECTED BY TRANSFER FUNCTION

same time (i.e. simultaneously). The approach, called *speckle holography*, uses the images of the point source as keys for the deconvolution of the images of the object. Deconvolution is discussed in the next chapter. We can also reconstruct an image of the object by using *speckle masking* to recover the spectral phase information on the object from the speckle data. Speckle masking is a method which involves estimation of the closure phases from each of the short exposure speckle images. This will be discussed in the chapter on interferometry. The object spectral phase information is combined with its spectral amplitude information from speckle interferometry and inverse transformed to yield the image of the object.

1.7 Some practical matters

Convolution and correlation

Table 1.2: Some characteristics of convolution and correlation of two real functions

	Convolution	Correlation
Notation	$g = f * h$	$Corr(f,h) = f \star h$
Fourier Transform pair	$f * h \leftrightarrow FH$ (Convolution Theorem)	$f \star h \leftrightarrow F^* H$ (Correlation Theorem)
Properties	$f * h = h * f$	$f \star h \neq h \star f$
Discrete 1-D description (*)	$g_i = \sum_j f_j h_{i-j}$	$Corr(f,h)_i = \sum_j f_{j-i} h_j$ $(= \sum_j f_j h_{i+j})$
(*) Moving strip approach	h scans f: sequence *reversed*	f scans h: sequence *not* reversed
Example of use	Image formation	Image registration

Autocorrelation

Table 1.3: Some characteristics of autocorrelation of a real function

	Autocorrelation
Notation	$Corr(f,f) = f \star f$
Fourier Transform pair	$f \star f \leftrightarrow FF^* = \|F\|^2$ (Wiener-Khinchin Theorem)
Properties	$f \star f$ is an even function
Discrete 1-D description (*)	$Corr(f,f)_i = \sum_j f_{j-i} f_j$ $(= \sum_j f_j f_{i+j})$
(*) Moving strip approach	f scans f: sequence *not* reversed
Example of use	Generation of OTF from pupil function

Example: Let $f = \{2\ 2\ 3\ 3\ 4\}$ and $h = \{1\ 1\ 2\}$, then
$f * h = \{2\ 4\ 9\ 10\ 13\ 10\ 8\}$, $h \star f = \{4\ 6\ 10\ 11\ 14\ 7\ 4\}$, and
$h \star h = \{2\ 3\ 6\ 3\ 2\}$.

Computing convolution and correlation using the discrete Fourier Transform

The convolution and (auto)correlation of two functions can be computed using the Fourier Transform and the Fourier Transform pair relations listed in tables 1.7 and 1.7. However, it is important to ensure that the array sizes used in the transforms are large enough to accommodate the resultant convolution/correlation which has a size in each dimension that is equal to the sum of the maximum sizes of each of the two functions in the same dimension. Practically, this is achieved by padding the arrays that describe the two functions to be convolved/correlated with zeros.

Image sampling

A digital image represents a sampled version of a continuous optical image. Converting from a continuous optical image $g(x, y)$ to its digital representation $\tilde{g}(x_i, y_i)$ requires averaging over the pixel area and sampling. That is,

$$\tilde{g}(x_i, y_i) = \{g(x, y) * rect(\Delta x', \Delta y')\} \times \delta(x, y : \Delta x, \Delta y) \tag{1.87}$$

where $rect(\Delta x', \Delta y')$ and $\delta(x, y : \Delta x, \Delta y)$ are the rect and sampling functions, respectively, $\Delta x, \Delta y$ are the sampling intervals (i.e. the spacing between the pixel centers in the x- and y-directions), and $\Delta x' \leq \Delta x, \Delta y' \leq \Delta y$ are the dimensions of the photosensitive region of a pixel. Note, the fraction of the detector surface that is sensitive to light is called the fill factor (unrelated to the pupil's fill factor !) and is given by:

$$\textit{fill factor} = 100\% \times (\Delta x' \Delta y')/(\Delta x \Delta y). \tag{1.88}$$

According to the Nyquist-Shannon sampling theorem, in order to preserve the spatial resolution of the original continuous image, the sampling intervals should be no greater than one-half the size of the smallest resolvable feature of the optical image. This "critical" (or "Nyquist"), sampling is equivalent to acquiring samples at twice the highest spatial frequency contained in the image. We note that the use of a smaller sampling interval does not yield more spatial information. However, this "over-sampling" is relatively harmless, the only downsides being increased storage requirements and computation times for image processing. The same is not true when using a larger sampling interval. Here the subsequent 'under-sampling' can be highly detrimental as it results in both a loss of information and aliasing artifacts in the recorded imagery. Aliasing is where high spatial frequency signals appear as low spatial frequency signals in the sampled signal.

For critical sampling of an image acquired with an instrument with a circular aperture we have $2\Delta x = 1.22\lambda/D = 2\Delta y$. The Fourier spectrum of the sampled image is a band-limited function, due to the multiplication of the object spectrum by the OTF, with cut-off spatial frequencies of $u_{co} = 1/(2\Delta x)$ and $v_{co} = 1/(2\Delta y)$, and frequency resolutions of $\Delta u = 1/(M\Delta x)$ and $\Delta v = 1/(N\Delta y)$, where M, N are the number of pixels in the camera along the x- and y- directions, respectively.

As an example, if a detector with an array of $M \times N$ square pixels, with $M = N = N_{pix}$, provides critical sampling of an image acquired through a circular aperture at wavelength λ_1, then the diameter of the OTF, on the same pixel grid, is N_{pix} and the diameter of pupil is $N_{pix}/2$. To simulate the PSF for an image acquired with the same imaging system and detector at another wavelength $\lambda_2 > \lambda_1$ we have to scale the size of the OTF, and thus the pupil, by (λ_1/λ_2) because the cut-off frequency varies with wavelength ($\propto D/\lambda$). Note, for $\lambda_2 < \lambda_1$ the image would be under-sampled.

Simulating Atmospheric Phase Screens

There are four steps to generating a simulated atmospheric phase screen. First we create a 2D complex array of random numbers having a Gaussian distribution with mean = 0 and standard deviation $\sigma = 1$. Second, we multiply this array by the amplitude of the wavefront phase spectral density. We note that the sampled approximation to Equation (1.59) is given by

$$P_\phi(i,j) = 0.023\ (2D/r_0)^{5/3}\ (i^2 + j^2)^{-11/6}. \qquad (1.89)$$

where i and j are the sample indices[22]. Third, we ensure that the D.C. (i.e. zero spatial frequency) bin is set to zero to ensure a zero mean phase at the end of the process. [23] Fourth, we inverse Fourier Transform the resultant array to produce two independent realizations of a Kolmogorov phase screen: one in the real part of the transform, the other in the imaginary part. We note that for the simulated screen to be accurate over the aperture we are interested in we need to generate the screen on an array that is at least four-times larger than the aperture. This is necessary for two reasons. First, we need to use a number of points in the simulation that can capture the outer scale length of the turbulence L_0 which typically has a value of around 15 to 25 m, depending upon the site. [24] Without this, the low spatial frequencies are not adequately represented and the observed phase structure function for the phase screen does not well represent the expected phase structure function (Equation (1.61)). Second, using the FFT approach for generating a phase screen results in phase screens that are periodic in x and y and which do not appear to have any apparent overall slope. This is not realistic. However, the shortcoming can be mitigated by only using the central section of the screen.

The downside of the FFT approach is that when the aperture is much smaller than L_0 we need to produce very large screens which is computationally expensive. In this case there more computationally

[22] Lane et al. 1992

[23] We also need to ensure that the D.C. bin is in the correct position (i.e. its is in the corner of the array) in preparation for the inverse Fourier transform in the final step.

[24] Sedmak 2004

efficient methods for generating phase screens that can be used [25]. As the long-exposure seeing probes the medium to large-scale properties of the phase screens, we can compare the measured FWHM of the long-exposure image generated by summing over a large number of speckle PSFs to the theoretical FWHM of the long -exposure image, to validate the fidelity of the phase screens.

[26]

EXERCISE 1.3—

1. Convolve the two 1D functions $(3,4,5)$ and $(2,1,3)$ using the serial product and the FFT methods.
2. Generate a PSF via the autocorrelation of the pupil function.
3. Show how the images of M31, Saturn, and the USAF 1951 pattern are affected by having 1 wave of defocus in the imaging system.
4. Show how an image of a triple star system is affected by having $\frac{1}{2}$ wave of vertical coma in the imaging system.
5. Compare the MTFs for 1 wave of defocus and $\frac{1}{2}$ wave of vertical coma.
6. Show that the Golay-6 pupil is a good pupil for imaging. How can we check this on its OTF ?
7. Make an image of an object through the Golay-3 pupil. What happens when we increase the sub-aperture distance ?
8. How can we determine the size of the aperture used to generate an image, from just the image?

---

[25] Lane et al. 1992

[26] IMAGE: SIMULATED TURBULENCE PHASE SCREEN

2

Image Restoration

Achieving the highest possible resolution imagery from a ground-based telescope requires removing the deleterious effects of image blur. This can be accomplished in two ways: adaptive optics compensation of the incoming wave front (see chapter 3) and numerical image restoration. Ideally both approaches are used in combination. However, as numerical restoration is more accessible, in that it only requires a computer and not an assembly of complex optical hardware, we will address it first.

2.1 Image preprocessing

Before we embark on a journey into methods for image restoration we need to point out the success of any type of image restoration depends *strongly* on the quality of the image preprocessing performed beforehand. The types of preprocessing can be divided into three categories: calibration, windowing, edge extension and sectioning.

1. Calibration: Raw image data are subject to variations in the pixel-to-pixel sensitivity of the detector and/or by distortions in the optical path and sit on a pedestal due to the dark current signal in the detector. These effects result in artifacts in the image need to be taken into account, either before embarking in image restoration, or as a part of the image restoration process. The correction for pixel-to-pixel variations is known as a "flat-field" correction and requires measurement of a source of light with constant, uniform intensity across the telescope pupil. The dark current signal is obtained by acquiring images with no incident light. We note that dark current can vary with time if the temperature of the detector changes. The raw data are processed by subtracting a suitable dark frame and dividing by the flat-field frame (which has been dark current corrected).

2. Windowing and edge extension: There is a tendency for deconvolu-

tion to be complicated by the truncation of the blurred image by the recording frame. For methods that require the Fourier transform of the data, we need to take into account that the Fourier Transform always treats an image as if it were part of a periodically replicated array of identical images extending horizontally and vertically to infinity. Therefore when the recording frame truncates the blurred image signal, we need to ensure that this periodicity requirement is met if we are to minimize artifacts due to 'edge effects'. The simplest approach is to "window" (or "apodize") the data frame by multiplying by a function that smoothly goes to zero at the edge of the frame. There are a large number of window functions that satisfy this condition, each with advantages and disadvantages. The Tukey window, also known as the tapered cosine window, is a common choice. The Tukey window function can be regarded as a unit amplitude cosine lobe of width $\alpha N/2$ that is convolved with a unit height rectangular window of width $(1 - \alpha/2)N$. That is, the data near the edge of the frame are modified by a cosine roll-off such that the signal is zero at the edges of the data frame. Thus, when using a Tukey window to apodize the data we need to bear in mind that the restored data will be corrupted over a distance equivalent to the sum of the cosine lobe width and half of the extent of the PSF from the edge of the frame.

Another approach, which allows for a slightly larger region of the data frame to be accurately restored, is to use edge extension. Simple edge extension involves embedding the data frame in a larger array and then continuing the signal at the edge of the data frame out along straight lines perpendicular to the edge of the frame out to the edge of the larger array and ensuring that the extended signal smoothly tapers to zero at the edge of the larger array. In this case, only the region within a half of the width of the PSF at the edges of the restored image is corrupted.

3. Sectioning: This involves dividing an image with a spatially varying PSF into sections each of which may be treated as having an approximately spatially invariant PSF.

2.2 *Multiplicative deconvolution*

In chapter 1 we saw that images are described by a convolution of an object intensity distribution with a PSF, i.e.

$$I = O * PSF \,, \tag{2.1}$$

which in the Fourier domain can be expressed as

$$\mathscr{F}\{I\} = \mathscr{F}\{O\} \cdot OTF \,. \tag{2.2}$$

To obtain the original object from the blurred image requires knowing the blurring function and performing deconvolution. Given the PSF and the image I it seems trivial to solve for the object O: as long as there are no zeros in the OTF we just use the simple inverse filter $W = 1/OTF$, i.e.,

$$O = \mathscr{F}^{-1}\{\mathscr{F}\{I\} \cdot W\} \,. \tag{2.3}$$

That is, deconvolution can be considered as a filtering operation.

However, from a practical standpoint, treating the imaging equation as being solely due to the convolution of the object with the PSF for the system, is too simplistic: there is always some contamination (e.g. photon shot noise, camera read noise). A better model for an observed image is thus:

$$I = O * PSF + C \tag{2.4}$$

or, equivalently,

$$\mathscr{F}\{I\} = \mathscr{F}\{O\} \cdot OTF + \mathscr{F}\{C\} \,, \tag{2.5}$$

where C represents noise contamination. Now when we use a simple inverse filter we can see that we have a term $\mathscr{F}\{C\}/OTF$ which can cause noise amplification in regions where this ratio swamps $\mathscr{F}\{O\}$.

We can circumvent this issue by defining a modified inverse filter

$$W = \frac{OTF^*}{|OTF|^2 + \Psi} \tag{2.6}$$

which is known as the *Wiener filter*. Here $*$ denotes complex conjugate, and the function Ψ is a measure of the noise-to-signal ratio as a function of spatial frequency. That is, Ψ is an estimate of $|\mathscr{F}\{C\}/\mathscr{F}\{O\}|^2$.

Thus the filter defined by equation (2.6) satisfies

$$\begin{aligned} W &\simeq 1/OTF && \text{when } \ |OTF| \gg \Psi \\ &\simeq OFT^*/\Psi && \text{when } \ |OTF| \ll \Psi \end{aligned} \tag{2.7}$$

We note that the presence of noise contamination in the imaging equation makes it impossible to obtain an exact solution to the deconvolution problem. [1]

When using a Wiener filter approach with a time series of k images of speckle data (i.e. speckle holography), we have

$$O = \mathscr{F}^{-1}\left(\frac{\sum_k \mathscr{F}\{I_k\} \cdot OTF_k^*}{\sum_k \left(|OTF_k|^2 + \Psi_k\right)}\right) \,. \tag{2.8}$$

2.3 *Blind deconvolution*

In the previous section we assumed that we had perfect knowledge of the PSF. This is rarely the case. In some scenarios we may have

[1] IMAGE: NOISY BLURRED IMAGE; SIMPLE INVERSE FILTERED IMAGE; WIENER FILTERED IMAGE

a reasonable estimate of the PSF whilst in others we may have only poor (or no) knowledge. The latter is usually the case when dealing with imagery obtained through atmospheric turbulence. Fortunately, for imagery where we have poor knowledge of the PSF we can turn to the technique of blind deconvolution. Blind deconvolution provides a way to estimate both the original object and the PSF from just the blurred image. There are many algorithms for solving blind image deconvolution but here will focus on iterative blind deconvolution.

In blind deconvolution we are solving for both the object and the PSF given just a single noisy image. This make blind deconvolution an ill-posed inverse problem and in order to make any headway in solving the problem we need to invoke prior information on the physics of the imaging system. There are many forms of prior information that have been invoked in blind deconvolution but probably the most powerful prior information is that for incoherent imaging system we are measuring intensities which are represented by positive functions. That is, both the object and the PSF are positive functions.

The basic iterative blind deconvolution approach we will examine here (there are many variations) has four steps:

1. Define a cost function (also called an objective function in computer science). Here we use

$$\epsilon(O, PSF) = \sum_{x,y} w(x,y)\left(I(x,y) - \hat{I}(x,y)\right)^2 \ , \tag{2.9}$$

   where $\hat{I} = \hat{O} * P\hat{S}F$ is our model for the observed image I, $\hat{O}$ and $P\hat{S}F$ are our estimates for the truth object and PSF, respectively, and w is a weighting function that reflects the noise statistics in I.

2. Model the object and PSF. Here we model them both as 2D pixel arrays in which each pixel is a variable to be determined. We need to ensure that $\hat{O}$ and $P\hat{S}F$ are positive functions. This can be done in a number of ways. One way is to re-parameterize these quantities as the square of the variables we adjust in the minimization (e.g., $O(x,y) = \beta^2(x,y)$ where $\beta(x,y)$ are the variables). We also need to ensure that the PSF has unit volume. We can do this with the re-parameterization $PSF(x,y) = \alpha(x,y) / \sum_{x,y} \alpha(x,y)$.

3. Select an iterative gradient-based optimization algorithm to find the solution. The gradients required for the optimization are computed using: [2]

$$\frac{d\epsilon}{dO}(x,y) = -2\ (R \star PSF)_{x,y} \tag{2.10}$$

   and

$$\frac{d\epsilon}{d\alpha}(x,y) = -2\frac{1 - PSF(x,y)}{\sum_{x,y}\alpha(x,y)}\ (R \star O)_{x,y} \tag{2.11}$$

[2] In 1-D, let $\epsilon = \sum_n r_n^2$ with $r_n = (g_n - \sum_m f_m h_{n-m})$, where we have used the serial product description for convolution. We thus have $\frac{d\epsilon}{df_m} = \sum_n \frac{d\epsilon}{dr_n} \cdot \frac{dr_n}{df_m} = -2\sum_n r_n h_{n-m}$ This is the serial product description for correlation so $\frac{d\epsilon}{df_m} = -2(r \star h)_m$.

where $R(x,y) = w(x,y)\left(I(x,y) - \hat{I}(x,y)\right)$ is the weighted residual, and where $\star$ denotes the correlation, not the convolution. Note, if positivity is enforced by using the square of the variable, equation (2.11) will need to be multiplied by $2\alpha(x,y)$.

4. Provide the algorithm with the best possible initial estimates for $\hat{O}$ and $P\hat{S}F$, and run the optimization.

We can improve on the modeling of the PSF by invoking another prior: we are imaging through a finite aperture and therefore the PSF is a band limited function, i.e. it only has spatial frequencies out to the diffraction limit for the telescope. Implementation of this prior knowledge is straightforward for short-exposure imaging. Instead of a pixel-by-pixel model of the intensity of the PSF, we model the PSF via its pupil function and use a pixel-by-pixel (or Zernike polynomial) representation of the wave front phase in the pupil as the variables (we typically ignore the effects of scintillation). This approach also has the benefit of requiring fewer parameters to model the PSF. For images with exposure times greater than the coherence time for the atmosphere we can impose the prior through an error function on the OTF which penalizes signal outside the cut-off frequency for the imaging system, i.e.

$$\epsilon_{BL} = \sum_{u,v} M(u,v)|OTF(u,v)|^2 \tag{2.12}$$

where $M(uv)$ is a binary mask that is zero inside the cut-off frequency and unity outside. This error function is minimized along with the cost function in equation (2.9).

If a time series of images is available, and the object can be considered stationary over the length of the time series, then we can use multi-frame blind deconvolution (MFBD) and equation (2.9) becomes

$$\epsilon = \sum_{k} \epsilon_k \tag{2.13}$$

where the index k is over frames. The advantage of MFBD is that the object estimate is common to all frames.

3

2.4 *Diversity imaging methods*

Another way to inject *prior* information into the restoration process is by employing the diversity imaging methods:

1. Phase diversity: In this technique we image the target simultaneously in two channels and intentionally apply a known phase perturbation to the incoming wave front in one of the channels. That

[3] IMAGE: TRUTH OBJECT, TRUTH PSF; BLURRED, NOISY IMAGE; REVOVERED OBJECT; RECOVERED PSF

IMAGE: show results for MFBD with different numbers of frames is, the pupil functions in the two channels are given by

$$\begin{aligned} P_1(u,v) &= A(u,v)\exp i\Phi(u,v) \\ P_2(u,v) &= A(u,v)\exp i(\Phi(u,v)+\eta(u,v)) \ , \end{aligned} \tag{2.14}$$

where η is the injected phase perturbation. Typically the perturbation is a simple defocus, which corresponds to a quadratic variation of phase across the exit pupil and is easy to implement (e.g., simultaneously acquire in- and out-of-focus images). However, other phase perturbations can be used. We note that although the traditional application of the technique obtains measurements in two channels, there is no inherent restriction (besides SNR considerations) on the number of channels that can be used.

2. Wavelength diversity: An alternative approach for creating phase diversity is to acquire simultaneous images at different wavelengths. Here we can use the prior knowledge that the wavefront phases in the pupil for the different wavelengths all correspond to the same optical path difference (OPD). That is,the pupil in channel i is given by

$$P_i(u,v) = A(u,v)\exp\left(i\frac{2\pi}{\lambda_i}OPD(u,v)\right) \tag{2.15}$$

By solving for the OPD in the pupil we are able to simultaneously determine the PSFs for the different wavelengths. Similar to phase diversity, although the traditional application of the technique obtains measurements at two wavelengths, there is no inherent restriction (besides SNR considerations) on the number of wavelength channels that can be used.

3. Aperture diversity: In this approach we simultaneously acquire images using aperture functions with different morphology which leads to different PSFs. One example is to partition the full aperture into a series of annuli and to simultaneously form focal plane images for each annular sub-aperture. Here the level of turbulence (as measured by D/r_0) and the spatial resolution are lowest for the smallest annulus data and highest for the largest annulus data. The resulting diversity in the PSFs can be leveraged in a boot-strap approach to the restoration. Basically, we first perform a restoration of the smallest annulus data where the level of turbulence is low and we can expect to obtain a high-quality, low-resolution estimate of the target object. We then use this estimate to seed the restoration of the data for the next sized annulus, and repeat the procedure through to the data for the largest annulus.

   The aperture diversity technique can also be applied to data acquired simultaneously with multiple telescopes with different apertures.

We note that these different approaches to imaging have differing numbers of measurements and unknowns (i.e. estimates of the object and PSFs). In general, for a time-series of N observations with M different measurement channels we have NM measurements and $N + K$ unknowns where K is the number of wavelengths at which the object is to be estimated, and N is also the number of PSFs that need to be estimated. For traditional blind deconvolution, $M = 1$ and $K = 1$. For traditional phase diversity, $M = 2$ and $K = 1$. For traditional wavelength diversity, $M = 2$ and $K = 2$. For the aperture diversity method proposed by Hope et al. (2016), $M = 3$ and $K = 1$.

2.5 *Some practical matters*

Cost functions

Given a noisy measurement of an image I we want to find the most likely model $\hat{I}$. This is known as the maximum *a posteriori* (MAP) solution and mathematically is given by Baye's theorem

$$Pr\{\hat{I}|I\} = \frac{Pr\{I|\hat{I}\}Pr\{\hat{I}\}}{Pr\{I\}} \tag{2.16}$$

where $Pr\{\hat{I}|I\}$ is the probability of obtaining the model $\hat{I}$ given the data I, $Pr\{I|\hat{I}\}$ is the probability of the data given the model, $Pr\{\hat{I}\}$ is the probability of the model, and $Pr\{I\}$ is the probability of the data. Since $Pr\{I\}$ does not depend on the model $\hat{I}$, maximizing $Pr\{\hat{I}|I\}$ with respect to $\hat{O}$ and $P\hat{S}F$ is equivalent to minimizing

$$\epsilon_{MAP} = -\ln\left[Pr\{I|\hat{I}\}\right] - \ln\left[Pr\{\hat{O}\}\right] - \ln\left[Pr\{P\hat{S}F\}\right] \ . \tag{2.17}$$

Minimization of the first term constrains the model $\hat{I}$ to be consistent with the data. The other two terms allow us to incorporate *a priori* knowledge about the object brightness distribution and the PSF. The form of the first term depends on the noise statistics for the image and minimization of just this term represents a maximum likelihood solution.

An observed image is the result of a Poisson process. In this case the $-\ln\left[Pr\{I|\hat{I}\}\right]$ term in equation (2.17) is given by the cost function

$$\epsilon = \sum_{x,y}\hat{I}(x,y) - \sum_{x,y}I(x,y)\log(\hat{I}(x,y)) \ . \tag{2.18}$$

However, for targets with high photon counts the central limit theorem ensures that the the Poisson distribution approaches a Gaussian distribution. In this case we can use the least-squares cost function given in equation (2.9). Practically, the Gaussian approximation can be invoked

when there are more than about 10 counts in a detector pixel.
We note that equation (2.9) is the zero-lag term of the maximum residual likelihood cost function [4]

$$\epsilon = \sum_{x,y} (R \star R)^2_{x,y} \quad . \tag{2.19}$$

Minimization of the non-zero lag terms in equation (2.19) has the added benefit that it is robust against spatial correlation in the residuals. That is, it helps to prevent over fitting to the noise in the data.

Optimization routines

In this class we will be using a constrained semi-Newton method (constrained: because we enforce positivity, semi-Newton: because it estimates the second derivative to improve convergence speed). The user needs to provide the objective function and its gradient with respect to the minimization parameters. The semi-Newton method estimates the best step to take based on the current gradient, the constraints, and the history of the previous steps. Note that there are other optimization methods possible (all iterative), including gradient-less methods such as Markov Chain Monte-Carlo optimizers (simulated annealing, parallel tempering) and splitting/proximal methods (Half Quadratic Split, ADMM).

Analytic derivatives for pupil plane model of PSF

To model PSFs at different wavelengths on the same detector grid we need to scale the wave front phase for the wavelength under consideration and then adjust the pupil size to ensure the correct cut-off frequency.

We define the discrete forward and inverse 1-D Fourier Transforms as

$$A_m = \frac{1}{N} \sum_n a_n e^{-i2\pi nm/N} \tag{2.20}$$

and

$$a_n = \sum_m A_m e^{i2\pi nm/N} \tag{2.21}$$

respectively, where N is the number of samples.
If $A_m = |A_m| e^{i\phi}$ and $h_n = a_n a_n^*$, then

$$\frac{dh_n}{d\phi_m} = a_n^* \frac{da_n}{d\phi_m} + a_n \frac{da_n^*}{d\phi_m} \tag{2.22}$$

[4] Pina & Puetter, 1992

and, for a function ϵ that depends on ϕ_m we have

$$\begin{aligned}\frac{d\epsilon}{d\phi_m} &= \sum_n \frac{d\epsilon}{dh_n}\frac{dh_n}{d\phi_m} \\ &= \sum_n \frac{d\epsilon}{dh_n}\left[iA_m a_n^* e^{i2\pi nm/N} - iA_m^* a_n e^{-i2\pi nm/N}\right] \\ &= -2\, Im\{A_m \sum_n \frac{d\epsilon}{dh_n} a_n^* e^{i2\pi nm/N}\}\end{aligned} \tag{2.23}$$

The summation term $\sum_n \frac{d\epsilon}{dh_n} a_n^* e^{i2\pi nm/N}$ is just the discrete inverse Fourier Transform of $(\frac{d\epsilon}{dh_n} a_n^*)$.

Similarly, for a function ϵ that also depends on $|A_m|$ we have

$$\frac{d\epsilon}{|A_m|} = 2\, Re\{\frac{A_m}{|A_m|}\sum_n \frac{d\epsilon}{dh_n} a_n^* e^{i2\pi nm/N}\} \ . \tag{2.24}$$

However, if the function ϵ depends on both $|A_m|$ and ϕ_m then it is better to work in the real and imaginary parts of the complex number $A_m = A_m^R + iA_m^I$ as opposed to the amplitude and phase. That is,

$$\frac{d\epsilon}{dA_m^R} = -4\Re\left[FT^{-1}\left(r_n a_n^*\right)\right] \tag{2.25}$$

and

$$\frac{d\epsilon}{dA_m^I} = -4\Re\left[FT^{-1}\left(ir_n a_n^*\right)\right] \tag{2.26}$$

The extension of the above equations from 1-D to 2-D is straightforward.

If ϕ_m can be represented by a polynomial basis set

$$\phi_m = \left(\sum_k a_k Z_k\right)_m \ , \tag{2.27}$$

then

$$\frac{d\epsilon}{da_k} = \sum_m \frac{d\epsilon}{d\phi_m}\frac{d\phi_m}{da_k} \tag{2.28}$$

Modeling PSFs at different wavelengths

To model PSFs at different wavelengths on the same image grid we need to scale the wave front phase for the wavelength under consideration and then adjust the size of the pupil aperture (and the phase it samples) to ensure the correct cut-off frequency.

3
Adaptive Optics Compensation

The basic goal of adaptive optics (AO) compensation is to measure the aberrations of an incident wave front and then cancel these out by applying compensating aberrations in real time.

An AO system basically consists of three components, a wave front sensor, a wave front corrector, and a control system. These components operate together in a closed feedback loop. In an imaging system equipped for AO compensation, light from the object of interest is captured by the telescope and a part of the light beam is sampled by a wave front sensor (WFS): the rest of the light goes to an imaging sensor. The signal from the WFS is sent to a control computer which then calculates the necessary changes to the optical path needed to flatten the wave front and subsequently sends correction signals to a wave front corrector that is located upstream of the WFS. All of this happens in a time much shorter than the atmospheric coherence time. A key requirement for AO is the availability of a bright light source that can be used to sense the wave front (a "beacon"). If the ambient light from the object is not bright enough to allow the wave front to be determined, then we need to find another beacon: this is usually either a nearby natural guide star or an artificial (laser) guide star.

3.1 Wavefront sensing

Wave front sensing is a technique by which an arbitrary wave front surface is converted into a uniquely defined intensity distribution, which in turn is inverted to yield an estimate of the original phase. There are two main approaches to wave front sensing: "direct" in the pupil plane, and "indirect" in the focal plane. In the former we split the pupil up into sub-apertures, bring the light from each sub-aperture to a focus, and use the resultant intensity distribution in each sub-aperture to deduce the phase of the wave front. In the latter, the wave front properties are deduced from intensity measurements made at or near the focal plane using the full aperture (i.e. the pupil is not divided before measurement). There are numerous wave front sensing techniques and the difference between them, as we will see in the rest of this section, is the way in which phase differences are turned into intensity differences.

Distortion in a wave front results in local variations in the gradient of the wave front, and after the wave front propagates for some distance these variations result in interference. The interference leads to variations in the intensity which can then be used to reveal the shape of the wave front surface. The process of converting a wave front into measurable intensity variations is described using the irradiance transport equation. [1]

$$\frac{\partial I}{\partial z} = -\left(\nabla I \cdot \nabla W + I \nabla^2 W\right) \; . \qquad (3.1)$$

Here ∇ is the gradient operator $\left(\frac{\partial}{\partial x} + \frac{\partial}{\partial y}\right)$ in the x, y plane, and we model the complex amplitude of a beam with irradiance $I(x, y, z)$ propagating along the z-axis of an optical system as

$$A(x, y, z) = [I(x, y, z]^{1/2} \exp\left[i \frac{2\pi}{\lambda} W(x, y, z)\right] \; , \qquad (3.2)$$

where $W(x, y, z)$ is the wave front surface at distance z.

Equation (3.1) shows that by measuring the (x, y) intensity distribution at two locations of z we can determine information on the local wave front slope ∇W and curvature $\nabla^2 W$: either of which can be used to reconstruct the wave front through spatial integration. Hardy[2] notes that for small wave front gradients, the intensity changes can be greatly enhanced by placing a mask with a known transmission function in one of the z planes, and then measuring the resulting intensity distribution in the other plane. The mask is made to match the spatial coherence length of the wave front.

Desirable characteristics for a wave front sensor include a linear relationship between the wave front and the intensity measurements, and the ability operate over a wide range of wavelengths.

Gradient-Based Sensors

Here we will consider two types of gradient-based wave-front sensor (WFS): the Shack-Hartmann WFS and the Pyramid WFS.

Shack-Hartmann WFS

In a Shack-Hartmann WFS the aperture is subdivided using a microlens array. For an incoming plane wave-front the images formed by the microlenses are located in a regular grid defined by the microlens array geometry. However, as soon as the wave-front is distorted, the images become displaced from their nominal positions by an amount

[1] Hardy 1998

[2] Hardy 1998

that is proportional to the wave-front slopes in the corresponding sub-apertures. That is,

$$\Delta\vec{x} = \frac{\lambda}{2\pi} M f \nabla W(\vec{x}) \tag{3.3}$$

where $\Delta\vec{x}$ is the displacement in the (x, y) plane, f is the focal length of the microlens array, and M is the (de)magnification of the imaging system ($M = f_{telescope}/f_{collimator}$). For a point source target, the magnification is often set such that the image from a microlens is a "spot" that is contained within a 2x2 pixel "quad cell" on the detector array. Quad cells are faster to read and to compute the centroid, and are less sensitive to noise than for a signal covering a larger number of pixels. The displacements of image centroids in the two orthogonal directions are determined using the "quad cell" formulas

$$\Delta_x \simeq \frac{b}{2}\left[\frac{(I_2 + I_1) - (I_1 + I_4)}{\sum_{i=1}^{4} I_i}\right] \quad \Delta_y \simeq \frac{b}{2}\left[\frac{(I_3 + I_2) - (I_3 + I_4)}{\sum_{i=1}^{4} I_i}\right] \tag{3.4}$$

where b is the diameter of the "spot" image on the quad cell. For point sources

$$\begin{aligned} b &= \lambda/d \quad \text{for } d < r_0 \\ b &= \lambda/r_0 \quad \text{for } d > r_0\ , \end{aligned} \tag{3.5}$$

For extended sources b is equal to the source size.

The disadvantage of the quad cell approach is that the sensitivity depends on the spot size which can vary as the turbulence conditions change. In addition, the quad cell signal becomes non-linear and saturates for large angular deviations (e.g., when the spot is not on all four quadrants of the quad cell): i.e. the sensor has a small dynamic range.

For images that extend beyond 2x2 pixels we calculate the image displacement via the x and y centroids (center-of-masses), i.e.,

$$\Delta_x = \frac{\sum_{i,j} x_{i,j} I_{i,j}}{\sum_{i,j} I_{i,j}}, \qquad \Delta_y = \frac{\sum_{i,j} y_{i,j} I_{i,j}}{\sum_{i,j} I_{i,j}} \tag{3.6}$$

The disadvantage here is that the signal-to-noise ratio (SNR) is decreased as we are spreading a given amount of light over more pixels. However, this is the optimal centroid estimator for the case where the spot is Gaussian distributed and the noise is Poisson. We note that the centroid gives the mean slope of the wave front (G-tilt) whereas what we usually want is the least-mean-squares slope (Z-tilt).

For WFS on extended targets with low contrast, such as the Sun, we modify the Shack-Hartmann sensor so that each sub-aperture image extends over a large enough field-of-view that the image displacement can be determined through via the peak of the cross-correlation function generated from cross-correlating the image in the sub-aperture with a the image in another sub-aperture that is defined to be the reference.

Pyramid WFS

In the pyramid sensor light from the pupil is focused at the tip of a glass pyramid which splits the light into four parts. Each beam is reimaged at a CCD to form four conjugate images of the pupil. Using the intensity at equivalent pixels in each pupil image along with Equations (3.4) gives a signal which is proportional to the x and y wave front gradients in the pupil. One advantage of the pyramid technique over a Shack-Hartmann sensor is that the spatial resolution of the sensor is the size of the detector pixel in contrast to the larger sub-aperture size of a micro-lens in the Shack-Hartmann. Another advantages include: (a) the pupil sampling and sensor sensitivity can both be dynamically varied to match changing observing conditions, (b) a linear response over a large dynamic range (requires dithering the pyramid with respect to the image in a time that is short compared to τ_0), and (c) it can be used with faint guide stars. The disadvantage of a pyramid WFS is that it is fundamentally dynamic (i.e. it has moving parts). However, use of a diffusing plate or a spatial light modulator to mimic the dithering provide solutions with no moving parts.

The wave front error variance associated with a pyramid sensor is given by

$$\sigma^2 = \frac{1}{2N} + \frac{2e_n^2}{N} \quad \mathrm{rad}^2 \qquad (3.7)$$

where N is the number of photoelectrons per pixel and e_n are the read noise electrons.

Curvature-Based Sensors

The curvature sensor uses a point-by-point subtraction of the out-of-focus images obtained at equal distances either side of the focal plane, to provide a measurement of the local wavefront curvature $\nabla^2 W$ through the relationship

$$\frac{I_1(\vec{r}) - I_2(\vec{r})}{I_1(\vec{r}) + I_2(\vec{r})} = \frac{f(f-s)}{s}\left[\nabla^2 W(\frac{f}{s}\vec{r}) - \frac{\partial W(\frac{f}{s}\vec{r})}{\partial \vec{n}}\vec{\delta}_r\right] . \qquad (3.8)$$

[3] Here f is the focal length of the imaging lens, s is the distance from focus on either side of the focal plane, and the Dirac delta $\vec{\delta}_r$ represents the outward pointing normal derivatives on the edge of the signal. The first term on the right-hand-side of Equation (3.8) represents the curvature of the wave front, and the second term the gradient at the edge of the aperture. Equation (3.8), which can be derived from Equation (3.1), shows that the sensitivity of a curvature sensor is inversely proportional to the amount of defocus s. However, as s increases the

[3] IMAGE: CARTOON OF HOW A CURVATURE SENSOR WORKS

spatial resolution decreases, diffraction effects increase, and the linear approximation breaks down (i.e. the relationship between the normalized intensity difference measurement and the wave front curvature eventually becomes non-linear). Note, in order to provide an accurate measurement of wave front curvature, we need to ensure that the blur from the turbulence is small compared to the area where the curvature measurement is taken.

The variance of a single curvature measurement is given by

$$\sigma^2_{curvature} = \frac{s^2}{f^4 N_p} \tag{3.9}$$

where N_p is the photon count.

We note that a Shack-Hartmann sensor can also be used as a curvature sensor. In fact it can be used to measure both the wave front curvature and gradient simultaneously, by sampling the incoming wave front at two axial locations along the optical path. [4] From Equation (3.1) we see that if the spatial intensity across the surface of a microlens is approximately constant (i.e. away from the edge of the aperture) then

$$\nabla^2 W \simeq -\left(\frac{\partial I}{\partial z}\right) / I \, . \tag{3.10}$$

Hence we can estimate the wave front curvature by using the intensity difference between corresponding spots in the two Shack-Hartmann "spot" images to approximate the partial derivative of the intensity with respect to the z axis (optical axis), and the intensity of each spot to approximate the total intensity.

[5]

Phase Retrieval

Phase retrieval is a non-linear iterative technique that is used to recover the phase in the aperture plane from intensity measurements at the focal plane and prior information about the object and noise.

A number of iterative phase retrieval algorithms have been proposed. These include the Gerchberg-Saxton routine, phase diversity, and multi-aperture phase retrieval (MAPR). The performance of the latter being particularly impressive in strong turbulence conditions. However, achieving real-time performance for these algorithms is difficult and they have yet to make a real entrance in wave front sensing for AO.

3.2 Wave front reconstruction

For each of the wave front sensors described in the previous section, we need to be able to reconstruct the wave front phase from a set of intensity measurements: either a map of its measured slopes (gradients) in two directions or its Laplacians and edge slopes. There are two main approaches for solving this inverse problem: zonal and modal.

---

[4] Barbero et. al. 2006

[5] TABLE OF PROS/CONS of different WF sensors: include cost, complexity, wavebands, real examples

Zonal reconstruction

In the zonal method the phase is determined on a discrete set of points distributed over the telescope aperture. The measurements S are linked to the wave front phase ϕ through the matrix equation

$$S = A\phi \tag{3.11}$$

where A is known as the geometry matrix[6] as its elements are often composed of sums and differences based on the geometrical configuration of wave front sensor positions and phase determination positions. If the geometry matrix includes the effects of a corrector, such as a deformable mirror, it is called the influence matrix.

We can invert Equation (3.11) to reconstruct the phases from the measured slopes using least-squares methods, i.e. minimize

$$\epsilon = ||S - A\phi||^2 , \tag{3.12}$$

which has the solution

$$\phi = \left[\left(A^T A \right)^{-1} A^T \right] S \ = BS, \tag{3.13}$$

where B is known as the reconstructor matrix (as it allows us to reconstruct the phases from the measurements). However, in almost all cases the matrix inversion presents problems because the matrix $A^T A$ is singular. This is because some parameters (or combinations of parameters) are not constrained by the data. For example, we can not determine piston from the slope measurements. In practice the matrix inversion is done by removing the indetermined (or poorly determined) parameters with the help of a Singular Value Decomposition algorithm.

An example of the elements of a simple A matrix is given by a Shack-Hartmann WFS with square sub-apertures. Here Fried (1977) proposed the model

$$\begin{aligned} S^x_{i,j} &= \left[(\phi_{i+1,j+1} + \phi_{i+1,j}) - (\phi_{i,j} + \phi_{i,j+1}) \right] /2d \\ S^y_{i,j} &= \left[(\phi_{i+1,j+1} + \phi_{i,j+1}) - (\phi_{i,j} + \phi_{i+1,j}) \right] /2d \end{aligned} \tag{3.14}$$

where the $\phi_{i,j}$'s are the phase values at the four corners of the sub-aperture, $S^x_{i,j}$ and $S^y_{i,j}$ are the measured slopes in the $x-$ and $y-$directions, respectively, and d is the width of the sub-aperture. The elements of A are $\pm(2d)^{-1}$ or 0.

---

[6] Tyson 1998

Modal reconstruction

In the modal method the phase is represented by the coefficients of expansion in a set of basis functions, called modes. In this case, for a set of wave front slope measurements at various positions i, j in the aperture, we have the set of linear equations

$$
\begin{aligned}
S^x_{i,j} &= \sum_{k=1}^{K} a_k \left(\frac{\partial Z_k(x,y)}{\partial x} \right)_{i,j} \\
S^y_{i,j} &= \sum_{k=1}^{K} a_k \left(\frac{\partial Z_k(x,y)}{\partial y} \right)_{i.j} .
\end{aligned}
\tag{3.15}
$$

where $Z_k(x,y)$ represents the basis functions, and a_k are the coefficients.

This set of equations can be represented by Equation (3.11) with the wave front phase vector ϕ replaced by the mode coefficient vector a, and the elements of the matrix A are now the derivatives of the polynomial basis functions evaluated at the wave front sensor sub-aperture positions. A number of basis functions can be used (e.g., Zernike polynomials, disk harmonics) and although there is a requirement that they be linearly independent, it is not necessary that they be orthogonal.

3.3 *Deformable mirrors*

There are several types of deformable mirror (DM), each has its own characteristic fitting error

$$\sigma^2_{fit} = \mu \left(d/r_0 \right)^{5/3} \quad \text{[radians]} \tag{3.16}$$

where d is the diameter of the sub-aperture and μ depends on the type of DM (its influence functions) and on the actuator geometry.

Exactly how large d is relative to r_0 is a design decision. One choice is to match d the median r_0 value at a site. This means that the DM needs to have $\sim (D/r_0)^2$ sub-apertures. Another design decision is the required dynamic range (or "stroke"). The typical stroke (total up and down range) for astronomy depends on telescope diameter: ± several microns for a 10 m class telescope and ± 10-15 microns for a 30 m class telescope. The last design requirement is low hysteresis. That is, we want the actuators to go to same position when we apply the same voltage.

Now the different types of DMs have different numbers of actuators. Since the total number of controlled actuators (for one actuator per sub-aperture) is $N = (D/d)^2$, then the fitting error becomes

$$\sigma^2_{fit} = \mu (D/r_0)^{5/3} N^{-5/6} \quad \text{[radians].} \tag{3.17}$$

Thus, for two types of DM to provide the same level of correction we have

$$\left(\frac{N_1}{N_2}\right) = \left(\frac{\mu_1}{\mu_2}\right)^{6/5} . \tag{3.18}$$

Segmented DMs

The segmented DM is made of separate segments with small gaps. Each segment can have piston only control (typical $\mu = 1.26$) using a single actuator or tip-tilt/piston control using 3 actuators per mirror segment (typical $\mu = 0.18$).

Continuous face-sheet DMs

The continuous face-sheet mirror has an array of actuators glued to the back of thin mirror. These actuators push and pull on the mirror by applying a force perpendicular to the surface. This is done by applying a voltage to the actuator causing it to expand or contract in length. The face sheet thickness must be large enough to maintain flatness during polishing, but thin enough to deflect when pushed or pulled by the actuators. The thickness also determines the mirror's "influence function". This is the response of the mirror shape to a "push" or "pull" by a single actuator. Thick face sheets have a broad influence function, while thin face sheets have a more peaked influence function.

We note that with a value of $\mu = 0.28$, a continuous face-sheet DM requires 6.2 times less actuators than a piston-only segmented DM, and 1.8 times less actuators than a segmented DM with piston and tip/tilt correction, for the same level of fitting error.

Bimorph DMs

A imorph mirror consists of two piezoelectric wafers which are bonded together and are oppositely polarized (parallel to their axes). An array of electrodes is deposited between the two wafers. The front and back surfaces are connected to ground. The front surface acts as a mirror.

When a voltage is applied to an electrode, one wafer contracts and the opposite wafer expands, which produces a local bending. The local curvature being proportional to voltage, as such bimorph mirrors are well matched to curvature sensing AO systems and the electrode pattern is normally shaped to match the sub-apertures in the curvature sensor.

MEMS DMs

Micro-electro-mechanical systems (MEMS) DMs are fabricated using micro-fabrication methods from the integrated circuit industry (using semiconductor batch processing technology). They use a continuous membrane mirror and the actuation takes place through electrostatic charged plates. Early MEMS devices had a limited amount of stroke of $\sim 2\mu m$, however, modern devices have demonstrated a larger stroke of up to 8 μm. MEMS DMs are an attractive prospect as they have the potential to be inexpensive and to be made with large formats. Unfortunately, this potential is yet to be realized as it is currently difficult to produce a large format mirror that is free of defects.

Adaptive secondary mirrors

Adaptive secondary mirrors were pioneered by the University of Arizona and Arcetri Observatory in Italy.

The main advantages to making the secondary mirror into the "deformable mirror" are that no additional mirror surfaces are required and the DM can have high stroke. The main disadvantage is that they are hard to build and handle.

3.4 *Close loop operation*

The bandwidth error is due to the limited correcting bandwidth of the AO system. The residual phase variance due to the bandwidth error is given by

$$\sigma^2_{BW} = (\tau_0 f_S)^{-5/3} \ . \tag{3.19}$$

where the inverse of the atmospheric coherence time (τ_0) is called the Greenwood frequency and f_S is the bandwidth of the AO system. In the special case of a single turbulent layer moving at a speed v, the Greenwood frequency can be written as (Hardy, 1998; Tyson, 2011):

$$f_G = 0.427 \frac{v}{r_0} \ . \tag{3.20}$$

Equation (3.19) shows that to keep the residual wave front error due to the finite bandwidth of the servo system of the AO, at a low level, the AO system should correct the wave front on a time scale significantly shorter than τ_0.

3.5 *Artificial guide stars*

Astronomical AO systems typically use stars as "natural guide stars (NGSs)" to measure the wave-fronts. However, as we will now see, this is quite restrictive.

First, a guide star should ideally be located within the isoplanatic patch θ_0 of the target. This is because for a given distance θ between the guide star and target, the residual wave-front error due to anisoplanatism is given by

$$\langle \sigma^2_{\rm iso} \rangle = \left(\frac{\theta}{\theta_0} \right)^{5/3} . \tag{3.21}$$

Second, the photon noise error is inversely proportional to the photon flux, which is related to the stellar magnitude m:

$$\langle \sigma^2_{\rm phot} \rangle \propto \lambda^{-3.6} 10^{-0.4m} . \tag{3.22}$$

With these two restrictions it is clear that the number of targets available on the sky that will have a suitable NGS nearby is extremely small.

Fortunately, we can get some relief by using an artificial laser guide star (LGS), also called a laser beacon. The existing two types of LGS use either the Rayleigh scattering from air molecules or the fluorecence of sodium atoms in the mesosphere. In both cases the laser spot is formed at some finite altitude H above the telescope: H=10-20 km for a Rayleigh LGS or 90 km for a sodium LGS.

A consequence of this is that a turbulent layer at altitude h will be sampled differently by the laser and stellar beams. This gives rise to the so-called "cone effect" (or "focal anisoplanatism") which affects us in three ways: (1) The turbulence above H is not sensed by the LGS. (2) The outer portions of the stellar wave-front are not sensed. (3) The laser and stellar wave-fronts are scaled differently: the laser beam diameter is reduced by (1 - h/H) times. Hence, there is a differential "stretching" between the LGS and target wave-fronts. It turns out that the last factor is the most important one. While the laser wave-front is compensated by the AO, the stellar wave-front has a residual error due to the cone effect of

$$\sigma^2_{\rm cone} = \left(\frac{D}{d_0} \right)^{5/3} . \tag{3.23}$$

Here d_0 is a new parameter characterizing the cone effect. To the first approximation, it is given by

$$d_0 \approx 2.91 \theta_0 H; \tag{3.24}$$

and has a typical size of a few meters to 20 meters. Equation(3.23) shows that the cone effect can be a serious limitation on AO performance. Sodium LGSs have a clear advantage over Rayleigh LGSs with respect to the cone effect.

Another limitation of LGSs is that they do not sense the tip and tilt aberrations. The laser beam is deflected by the atmosphere twice, on its way upwards and downwards, whereas the stellar beam experiences only one deflection. If the LGS is projected from the main telescope, the upward and downward tilts compensate completely and the LGS image is stable in the telescope focal plane.

The standard solution for long-exposure imaging with LGS consists in using an additional NGS for tip-tilt compensation. If the angular distance between the target and the NGS is θ then the residual phase variance caused by the tip-tilt anisoplanatism is given by

$$\sigma^2_{\text{tilt}} \approx 0.1 \left(\frac{\theta}{\theta_0}\right)^2 \left(\frac{D}{r_0}\right)^{-1/3} .$$

The "tilt problem" is considered as a major obstacle for an LGS-assisted AO system. One approach is to sense tilt at the IR wavelength, where the NGS image falls within the isoplanatic patch.
Lastly,the laser spot is larger than a NGS which results in an increase in the residual phase variance due to the increased uncertainty in the measurement of the spot,

$$\sigma^2_{measure} \sim (6.3/SNR)^2 \tag{3.25}$$

Rayleigh beacons

The beam of a pulsed laser is focused at altitudes H between 10 and 20 km above ground, the return signal is obtained as a light back-scattered by the air density fluctuations. This Rayleigh scattering is more efficient at short wavelength (the cross-section is proportional to λ^{-4}), which explains the blue color of the clear sky and the increased extinction of blue stellar light. For a constant air density the return flux of a Rayleigh beacon would be proportional to H^{-2}, however, it decreases with altitude more steeply than this because the air gets more rarefied with altitude.

Most of the light is scattered at low altitudes. Beam focusing is not sufficient to define the desired LGS altitude. To do this, the return signal is gated in time (range gating). The length of the exposure typically corresponds to a 1-2 km range in altitude. Gating is achieved by a fast electro-optical shutter in front of the WFS.

Sodium beacons

The sodium layer at an altitude of about 90 km and with a thickness of about 10 km surrounds the Earth. It is most likely formed by micrometeorite ablation. Parameters of the layer (total number of atoms, mean altitude, profile) change seasonally, but also on time scales of days, hours and even minutes. Strong sporadic layers appear sometimes and then dissipate in few hours. On the average, there are some 10^{13} sodium atoms per square meter.

The sodium atoms can be excited by a laser beam tuned to the D2 line (wavelength 0.5890 microns) and radiate at the same wavelength. The natural width of the sodium D2 line is determined by the thermal motion of the atoms in the mesosphere and by the hyper-fine structure of the D2 line itself, it is about 3 GHz.

4

Optical interferometry: theory

4.1 Coherent and incoherent light

Correlation functions are used to characterize the statistical and coherence properties of an electromagnetic field. The degree of coherence is the normalized correlation of electric fields. Interferometers are merely devices to measure the coherence functions. The first order correlation is actually the amplitude-amplitude correlation, and is measured through amplitude interferometers. The normalized first order correlation function $\gamma^{(1)}$ quantifies the coherence between two electric fields:

$$\gamma^{(1)}(\boldsymbol{r}_1, t_1; \boldsymbol{r}_2, t_2) = \frac{\langle E^*(\boldsymbol{r}_1, t_1) E(\boldsymbol{r}_2, t_2) \rangle}{\left[\left\langle |E(\boldsymbol{r}_1, t_1)|^2 \right\rangle \left\langle |E(\boldsymbol{r}_2, t_2)|^2 \right\rangle \right]^{1/2}} \tag{4.1}$$

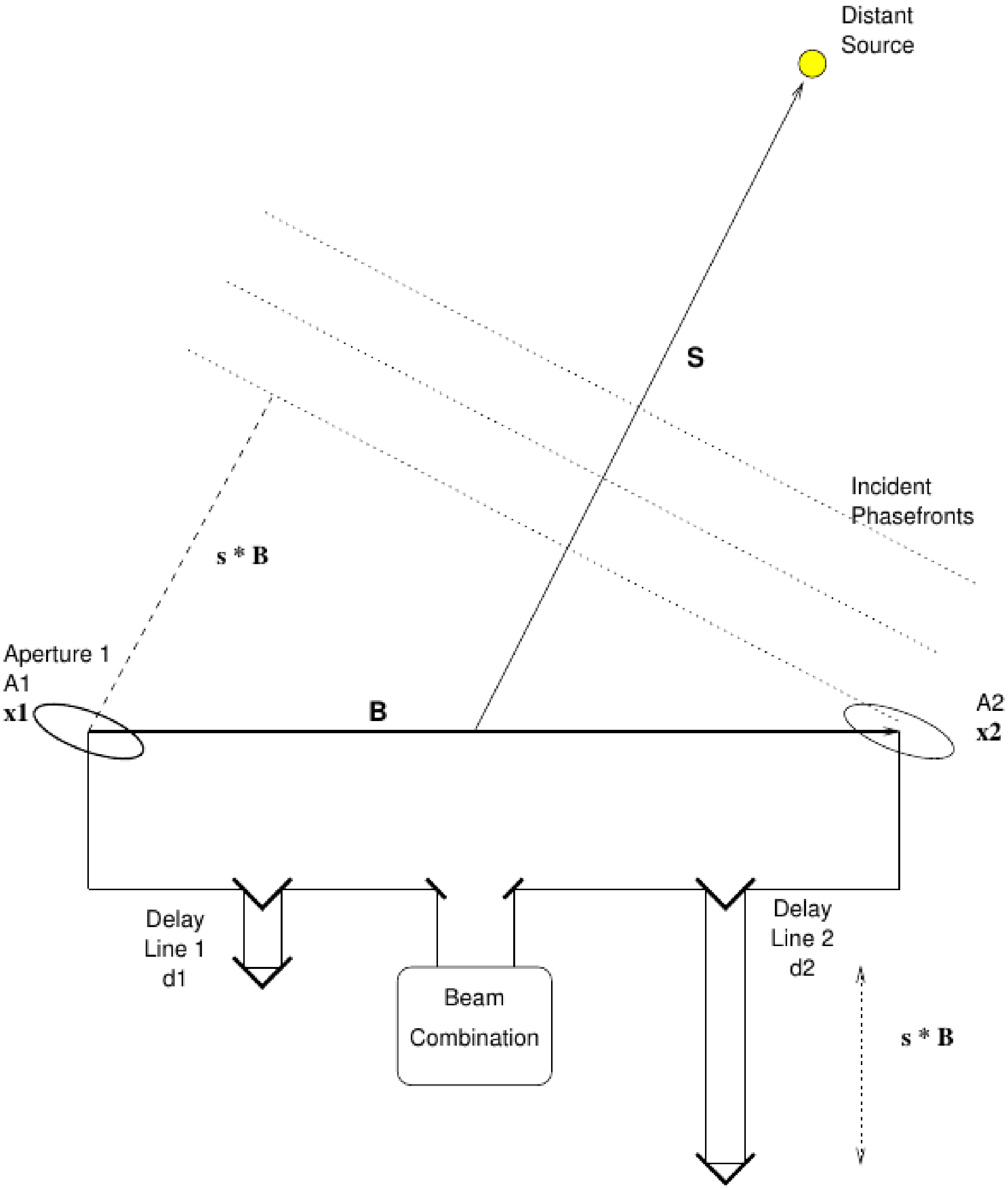
where $\langle . \rangle$ denotes an ensemble (statistical) average. Amplitude interferometers generally split an electric field into two components, introduces a time delay to one of the components, and then recombines them before measuring the resulting correlation $\gamma^{(1)}$, called the *complex degree of coherence*. Note that $|\gamma^{(1)}|$ ranges from zero, for incoherent electric fields, to one, for coherent electric fields. Anything in between is described as partially coherent.

The normalized second order correlation function $\gamma^{(2)}$ is used to find the statistical character of intensity fluctuations:

$$\gamma^{(2)}(\boldsymbol{r}_1, t_1; \boldsymbol{r}_2, t_2) = \frac{\langle E^*(\boldsymbol{r}_1, t_1) E^*(\boldsymbol{r}_2, t_2) E(\boldsymbol{r}_1, t_1) E(\boldsymbol{r}_2, t_2) \rangle}{\left\langle |E(\boldsymbol{r}_1, t_1)|^2 \right\rangle \left\langle |E(\boldsymbol{r}_2, t_2)|^2 \right\rangle} \tag{4.2}$$

and is measured by intensity interferometers.

Figure 4.1: A two telescope interferometer.


Distant
Source
S
Incident
Phasefronts
s * B
Aperture 1
A1
x1
B
A2
x2
Delay
Line 1
d1
Beam
Combination
Delay
Line 2
d2
s * B

4.2 *A two-element interferometer*

Monochromatic interference: fringes

Let's consider interference fringes produced in a two telescope stellar interferometer. The direction of the target is given by vector **s**, while telescopes are separated by vector **B**, called a *baseline*. An optical delay of d_1 (resp. d_2) is introduced by the delay line system for the light arriving at telescope 1 (resp. telescope 2). Observations are taking place at wavelength λ.

For monochromatic light, considering only a single polarization:

$$E_1 = A_1(t)\exp\left(ik[\mathbf{s}.\mathbf{B} + d_1]\right) \tag{4.3}$$

$$E_2 = A_2(t)\exp\left(ikd_2\right) \tag{4.4}$$

with $k = \frac{2\pi}{\lambda}$. A_1 and A_2 are complex fields that include information about the target and vary temporally.

At the combination point $E = E_1 + E_2$, and the time averaged intensity is:

$$\begin{aligned} I = \langle EE^* \rangle &\propto \langle \{A_1\exp(ik[\mathbf{s}.\mathbf{B} + d_1]) + A_2\exp(ikd_2)\} \times \{A_1^*\exp(-ik[\mathbf{s}.\mathbf{B} + d_1]) + A_2^*\exp(-ikd_2)\}\rangle \\ &\propto \left\langle |A_1|^2 \right\rangle + \left\langle |A_2|^2 \right\rangle + \langle 2\Re e\left\{A_1^* A_2 \exp\left(ik[\mathbf{s}.\mathbf{B} + d_1 - d_2]\right)\right\}\rangle \\ &\propto \left\langle |A_1|^2 \right\rangle + \left\langle |A_2|^2 \right\rangle + 2\sqrt{I_1 I_2}\Re e\left\{\gamma_{12}\exp\left(ik[\mathbf{s}.\mathbf{B} + d_1 - d_2]\right)\right\} \\ &\propto I_1 + I_2 + 2\sqrt{I_1 I_2}|\gamma_{12}|\cos\left(\phi_{12} + k\delta\right) \\ &\propto I_1 + I_2 + |V_{12}|\cos\left(\phi_{12} + k\delta\right) \end{aligned}$$

where we pose $\delta = \mathbf{s}.\mathbf{B} + d_1 - d_2$ as the OPD difference. The *incoherent flux* is represented by $I_1 + I_2$, while the coherent flux depends on the complex degree of coherence $\gamma_{12} = |\gamma_{12}|e^{i\phi_{12}} = \frac{\langle A_1^* A_2 \rangle}{\sqrt{I_1 I_2}}$. The *complex visibility* of the fringes is the normalized degree of coherence, i.e. a complex number. The *visibility amplitude* of the fringes is the ratio of the fringe amplitude to the total background illumination:

$$|V_{12}| = \frac{2\sqrt{I_1 I_2}|\gamma_{12}|}{I_1 + I_2} \tag{4.5}$$

and is also called the *fringe contrast*. The *visibility phase* ϕ_{12} carries the rest of the information about the observed object; but as we will see later, from the ground it is too corrupted by the atmosphere to be straightforwardly usable. The intensity varies a function of $k\delta$, which itself depends on the wavevector, $k = 2\pi/\lambda$, the baseline B, the pointing direction s, and the optical path difference $d_1 - d_2$ between the two interferometer arms. Adjacent fringe peaks are separated by Δd_1 or $\Delta d_2 = \lambda$, or $\Delta(\boldsymbol{s}.\boldsymbol{B}) = \lambda$ or $\Delta(1/\lambda) = 1/\delta$.

[1]

[1] IMAGE: Simple fringes with phase and amplitude

Polychromatic light, Coherence Envelope

Let see what happens in the simple case where $I_1 = I_2 = I_0$ (in which case $V_{12} = \gamma_{12}$, by the way), for a uniform bandpass in $[\lambda_0 - \frac{\Delta\lambda}{2}, \lambda_0 + \frac{\Delta\lambda}{2}]$:

$$I \propto \int_{\lambda_0 - \frac{\Delta\lambda}{2}}^{\lambda_0 + \frac{\Delta\lambda}{2}} 2I_0 \left[1 + V_{12} \cos\left(\frac{2\pi\delta}{\lambda}\right)\right] d\lambda \tag{4.6}$$

$$= 2I_0 \Delta\lambda \left[1 + V_{12} \frac{\sin\left(\pi\delta\frac{\Delta\lambda}{\lambda_0^2}\right)}{\pi\delta\frac{\Delta\lambda}{\lambda_0^2}} \cos\left(\frac{2\pi\delta}{\lambda_0}\right)\right] \tag{4.7}$$

$$= 2I_0 \Delta\lambda \left[1 + V_{12} \,\mathrm{sinc}\left(\frac{\pi\delta}{\Lambda_{\mathrm{coh}}}\right) \cos\left(\frac{2\pi\delta}{\lambda_0}\right)\right] \tag{4.8}$$

where $\Lambda_{\mathrm{coh}} = \frac{\lambda_0^2}{\Delta\lambda}$ is the coherence length and $\mathrm{sinc}(x) = \sin(x)/x$. Polychromatic fringes are modulated by a coherence envelope with a characteristic width equal to the coherence length. Here we supposed that V_{12} stays constant withing the bandpass, but in practice this is not always the case. In fact now we need to understand what V_{12} truly is, which is expressed by the Van Cittert-Zernike theorem. [2]

4.3 *Amplitude interferometry*

Amplitude interferometry is based on the measurements of the first order correlation function; in contrast, intensity interferometry measures the second order correlation function.

Spatial coherence: the Van Cittert-Zernike theorem

For measurements made at the same time in different location, the visibility only depends on the vector separation $\boldsymbol{\rho}$ between the two measurement locations:

$$\langle E(\mathbf{r_1}, t_1) \times E(\mathbf{r_1}, t_1)\rangle = V(\mathbf{r_1} - \mathbf{r_2}) = V(\boldsymbol{\rho}). \tag{4.9}$$

The Van Cittert-Zernike theorem, named after physicists Pieter Hendrik van Cittert and Frits Zernike, is a formula in coherence theory that states that for spatially incoherent sources in the far field, the normalized spatial coherence is equal to the normalized Fourier transform of the source's angular brightness distribution $I(\boldsymbol{\alpha})$:

$$V(\boldsymbol{\rho}) = \frac{\int I(\boldsymbol{\alpha}) e^{-i\boldsymbol{\alpha}.\boldsymbol{\rho}} d\boldsymbol{\alpha}}{\int I(\boldsymbol{\alpha}) d\boldsymbol{\alpha}}. \tag{4.10}$$

The spatial coherence function is what a Young's slit experiments measures: the apertures define the two regions where the wavefront is measured. This is also what optical interferometers (CHARA, VLTI) measure, and in this case $\boldsymbol{\rho} = \boldsymbol{B}$.

---

[2] IMAGE: How fringes at multiple wavelengths blend in to create a fringe envelope. Show what the coherence length corresponds to.

We can now tie the Van Cittert-Zernike to our fringe equation. Consider two very distant parallel planes, both perpendicular to the line of sight, and let's call them source plane and observation plane. Let's also assume we have the two-telescope system from figure 4.2. If we keep d_1 and d_2 equal using the delay lines, we see the fringe pattern corresponding to the current baseline orientation with visibility:

$$\begin{aligned} V_{12}(k, \boldsymbol{B}) &= \int I(\boldsymbol{\Delta s}) e^{-ik(\boldsymbol{\Delta s}.\boldsymbol{B})} d\Omega \\ V_{12}(u, v, 0) &= \iint I(\alpha, \beta) e^{-ik(\alpha B_x + \beta B_y)}\, d\alpha\, d\beta \\ V_{12}(u, v, 0) &= \iint I(\alpha, \beta) e^{-2\pi i(\alpha u + \beta v)}\, d\alpha\, d\beta \end{aligned} \tag{4.11}$$

where I_{source} is the intensity of the source, α and β are the direction cosines of a point on a distant source in the source plane, $u = B_x/\lambda$ and $v = B_y/\lambda$ are respectively the x-distance and the y-distance between the two observation points on the observation plane in unit of wavelength. u and v can also be understood as the coordinates of the spatial frequencies associated with the physical baselines of the interferometer, i.e. coordinates in the OTF. The *uv-coverage* is the location of all the sampling points in the (u, v) plane. The more points, the better the frequency sampling of the object and thus the better the observed object can be recovered.

Equation 4.11 implies that the wavefront from an incoherent source will appear mostly coherent at large distances. Intuitively, this can be understood by considering the wavefronts created by two incoherent sources. If we measure the wavefront immediately in front of one of the sources, our measurement will be dominated by the nearby source. If we make the same measurement far from the sources, our measurement will no longer be dominated by a single source; both sources will contribute almost equally to the wavefront at large distances.

The van Cittert-Zernike theorem is at the core of astronomical radio and optical interferometry. By measuring the visibilities at different points in the imaging plane of an astronomical object (i.e. by changing u and v), an interferometrist samples the Fourier transform of the source image. With enough of these samples, modeling and even imaging can be attempted. In practice we use the Discrete Fourier Transform (DFT) or the faster Nonequispaced Fast Fourier Transform (NFFT) to compute visibilities from a source image.

Finally should note some of the assumptions in the Van Cittert-Zernike theorem:

- A spatially coherent source does not obey the van Cittert-Zernike theorem. With the exception of pulsars and masers, all astronomical sources are spatially incoherent. Because they are observed at distances large enough to satisfy the van Cittert-Zernike theorem, these objects exhibit a non-zero degree of coherence at different points in the imaging plane.

- The van Cittert-Zernike theorem requires that we observe the source in the so-called far field. Depending on the wavelength, this may not be verified for some very close objects (GEO satellites, solar system bodies).

Temporal coherence: the Wiener-Khintchine theorem

For measurements made at the same location over time, the visibility depends on the time difference τ only:

$$\langle E(r_1, t_1) \times E(r_1, t_1) \rangle = V(t_1 - t_2) = V(\tau), \tag{4.12}$$

so we are quantifying the temporal coherence the extent to which the fields along a given wave train are correlated. This is what a laboratory Michelson interferometer measures. The Wiener-Khinchin theorem states that the normalized value of the temporal coherence function $V(\tau)$ is equal to the normalized Fourier transform of the spectral energy distribution of the source $B(\omega)$:

$$V(\tau) = \frac{\int B(\omega) e^{-i\omega\tau} d\omega}{\int B(\omega) d\omega} \tag{4.13}$$

This is spectroscopy without a dispersing element: Fourier spectroscopy. A broad spectral energy distribution leads to a coherence function that decays rapidly since τ and ω are reciprocal coordinates. We can define a temporal coherence time: $\tau_{\mathrm{coh}} \sim \frac{1}{\Delta\nu}$, with $\Delta\nu = \frac{\Delta\omega}{2\pi}$ the spectral bandwidth of the radiation.

The birth of stellar optical interferometry

One of the first uses of optical interferometry was applied by the Michelson stellar interferometer on the Mount Wilson Observatory's reflector telescope to measure the diameters of stars. The red giant star Betelgeuse was the first to have its diameter determined in this way on December 13, 1920.

In the 1940s radio interferometry was used to perform the first high resolution radio astronomy observations. For the next three decades astronomical interferometry research was dominated by research at radio wavelengths, leading to the development of large instruments such as the Very Large Array and the Atacama Large Millimeter Array. Optical/infrared interferometry was extended to measurements using separated telescopes by Johnson, Betz and Townes (1974) in the infrared and by Labeyrie (1975) in the visible.

In the 1980s the aperture synthesis interferometric imaging technique was extended to visible light and infrared astronomy by the Cavendish Astrophysics Group, providing the first very high resolution images of nearby stars. Other astronomical optical interferometers soon followed (IOTA, VLTI, CHARA, NPOI, MROI...), expanding the range of baselines (8 to 400 meters) and wavelengths (blue to mid-infrared).

4.4 *Hanbury Brown and Twiss effect: intensity interferometry*

In 1956, Robert Hanbury Brown and Richard Q. Twiss published *"A test of a new type of stellar interferometer on Sirius"*, in which two photomultiplier tubes (PMTs), separated by about 6 meters, were aimed at the star Sirius. Light collected into the PMTs using mirrors from searchlights. An interference effect was observed between the two intensities, revealing a positive correlation between the two signals, despite the fact that no phase information was collected. Hanbury Brown and Twiss used the interference signal to determine the apparent angular size of Sirius, claiming excellent resolution. The original HBT result met with much skepticism in the physics community. Although intensity interferometry had been widely used in radio astronomy where Maxwell's equations are valid, at optical wavelengths the light would be quantized into a relatively small number of photons. Many physicists worried that the correlation was inconsistent with the laws of thermodynamics and some even claimed that the effect violated the uncertainty principle. Hanbury Brown and Twiss resolved the dispute in a neat series of papers which demonstrated, first, that wave transmission in quantum optics had exactly the same mathematical form as Maxwell's equations albeit with an additional noise term due to quantization at the detector, and secondly, that according to Maxwell's equations, intensity interferometry should work. After a number of experiments, the whole physics community agreed that the observed effect was real.

The HBT effect

The Hanbury Brown and Twiss effect can actually be predicted solely by treating the incident electromagnetic radiation as a classical wave. Suppose we have a single incident wave with frequency ω on two detectors. Since the detectors are separated, say the second detector gets the signal delayed by a phase of ϕ. Since the intensity at a single detector is just the square of the wave amplitude, we have for the intensities at the two detectors:

$$\begin{aligned}
I_1 &= E^2 \sin^2(\omega t) \\
I_2 &= E^2 \sin^2(\omega t + \phi) = E^2(\sin(\omega t)\cos(\phi) + \sin(\phi)\cos(\omega t))^2
\end{aligned}$$

which makes the correlation

$$\begin{aligned}
\langle I_1 I_2 \rangle &= \lim_{T\to\infty} \frac{E^4}{T} \int_0^T \sin^2(\omega t)(\sin(\omega t)\cos(\phi) + \sin(\phi)\cos(\omega t))^2\, dt \\
&= \frac{E^4}{4} + \frac{E^4}{8}\cos(2\phi), \qquad (4.14)
\end{aligned}$$

a constant plus a phase dependent component. Most modern schemes actually measure the correlation in intensity fluctuations at the two detectors, but it is not too difficult to see that if the intensities are correlated then the fluctuations $\Delta I = I - \langle I \rangle$ ought to be correlated. In general:

$$\begin{aligned}
\langle \Delta I_1 \Delta I_2 \rangle &= \langle (I_1 - \langle I_1 \rangle)(I_2 - \langle I_2 \rangle) \rangle = \langle I_1 I_2 \rangle - \langle I_1 \langle I_2 \rangle \rangle - \langle I_2 \langle I_1 \rangle \rangle + \langle I_1 \rangle \langle I_2 \rangle \\
&= \langle I_1 I_2 \rangle - \langle I_1 \rangle \langle I_2 \rangle, \qquad (4.15)
\end{aligned}$$

and since the average intensity at both detectors in this example is $E^2/2$,

$$\langle \Delta I_1 \Delta I_2 \rangle = \frac{E^4}{8}\cos(2\phi), \qquad (4.16)$$

so our constant vanishes. The average intensity is $E^2/2$ because the time average of $\sin^2(\omega t)$ is $1/2$.

The Hanbury Brown and Twiss effect uses the relation $\gamma^{(2)}(\tau) = 1 + |\gamma^{(1)}(\tau)|^2$ to find $|\gamma^{(1)}(\tau)|$ from a measurement of $\gamma^{(2)}(\tau)$. An intensity interferometer can thus be built from two light detectors, typically either radio antennas or optical telescopes with photomultiplier tubes (PMTs), separated by some distance (the baseline). Both detectors are pointed at the same astronomical source, and intensity measurements are then transmitted to a central correlator facility. A major advantage of intensity interferometers is that only the measured intensity observed by each detector must be sent to the central correlator facility, rather than the amplitude and phase of the signal.

The Narrabri Stellar Intensity Interferometer

After initial experiments were carried out in radio, the Narrabri Stellar Intensity Interferometer (NSII) was the first astronomical interferometer to measure the diameters of a large number of stars at visible wavelengths (Michelson and Pease's 1920 experiment only measured five stars, including Betelgeuse). It was designed by (amongst others) Robert Hanbury Brown. It was built by University of Sydney School of Physics and was located near the town of Narrabri in north-central

New South Wales, Australia. The design was based on the earlier optical intensity interferometer built by Hanbury Brown and Richard Q. Twiss at Jodrell Bank in the UK. The NSII operated from 1963 until 1974. With telescope diamters of 6.7m and baselines ranging from 10 to 188m, it was used to measure the angular diameters of 32 stars. These were in fact all stars of sufficient signal-to-noise, since intensity interferometry is much less sensitive than amplitude interferometry.

A resurgence of interest

Despite the low sensitivity, in recent years there has been a resurgence of interest for intensity interferometry.

The reason is that amplitude interferometry has many shortcomings: it requires good optical quality in the beam train and enough stability to control the time-dependent OPD between the arms of the interferometer at sub-millisecond time scales with accuracies better than the wavelength. While this is done routinely in near-infrared and visible (although still largely untested at blue/ultraviolet or kilometer-long baselines), this make the arrays static (non-mobile) and prevents space applications.

Intensity interferometry only requires OPD control with given by the bandwidth of the bandpass filter (equivalently, by the time resolution of the acquisition chain), rather than by the wavelength of the radiation itself. For example, an intensity acquisition chain with 100 ps time-resolution requires an OPD control with millimetric accuracy only. Linking telescopes does not involve delay lines, but only simple signal cables. In addition, the role of the telescopes is only to collect photons onto the sensitive surfaces of the sensors. This requires correct guiding and if possible fast tip-tilt correction, but the optical quality of the telescopes can be less-than-ideal. The current popular idea is using Gamma ray telescopes – which are essentially large light buckets (> 12 meter diameters) – as intensity interferometry collectors.

4.5 From visibilities to models and images: the inverse problem

The van Cittert-Zernike theorem is crucial to the measurement of the brightness distribution of a source. With two telescopes, a radio astronomer (or an infrared or submillimeter astronomer) can measure the correlation between the electric field at the two dishes due to some point from the source. By measuring this correlation for many points on the source, the astronomer can reconstruct the visibility function of the source. By applying the van Cittert-Zernike theorem, the astronomer can then take the inverse Fourier transform of the visibility function to discover the brightness distribution of the source. This technique is known as aperture synthesis or synthesis imaging.

In practice, radio astronomers rarely recover the brightness distribution of a source by directly taking the inverse Fourier transform of a measured visibility function. Such a process would require a sufficient number of samples to satisfy the Nyquist sampling theorem; this is many more observations than are needed to approximately reconstruct the brightness distribution of the source. Astronomers therefore take advantage of physical constraints on the brightness distribution of astronomical sources to reduce the number of observations which must be made. Because the brightness distribution must be real and positive everywhere, the visibility function cannot take on arbitrary values in unsampled regions. Deconvolution algorithms for interferometry are called image reconstruction algorithms, they attempt tot reconstruct the brightness distribution of the source from a limited number of observations.

List of Figures

1.1 2

1.2 The propagation of light between two planes is regarded as a linear system. 3

1.3 Fourier Transform by a lens. L1 is the collimating lens, L2 is the Fourier transform lens, u and v are normalized coordinates in the transform plane. 5

1.4 A Golay array with three sub-apertures. 6

1.5 The first 21 Zernike polynomials, ordered vertically by radial degree and horizontally by azimuthal degree. Each represent a possible mode of the pupil phase. 8

1.6 How to build "by hand" the autocorrelation of a single aperture pupil P. For simplicity we assume here the pupil is uniform and real-valued (zero phase). We apply Equation 1.50 by sliding the pupil on top of itself (here from left to right), and computing the integral of the overlapping surface (hashed). Note that since the pupil is point-symmetric, so is the MTF. We obtain the well-known result that the frequency cutoff is $\frac{D}{\lambda}$. 15

1.7 How to build "by hand" the autocorrelation of a pupil with two sub-apertures. The same principles as for Figure 1.3 apply. This pupil is not point-symmetric, so the shape of the MTF slide depends on the sliding direction: this is obvious since this pupil will provide better resolution along the horizontal axis than the vertical one. 15

4.1 A two telescope interferometer. 56

List of Tables

1.1 The first few Zernike modes, ordered by Noll index j, normalized so that $\int_0^{2\pi} \int_0^1 Z_j^2 \, \rho \, d\rho \, d\theta = \pi$. 8

1.2 Some characteristics of convolution and correlation of two real functions 28

1.3 Some characteristics of autocorrelation of a real function 28

List of Exercises

1.1 Exercices on Zernikes 9
1.2 Exercises on PSF 12
1.3 Exercices on image formation 31