PHYSICS 688

ASTRONOMICAL AND ASTROPHYSICAL MEASUREMENTS

Paul Hickson

2002

PHYSICS 688

ASTRONOMICAL AND ASTROPHYSICAL MEASUREMENTS

COURSE INFORMATION

Time and Location:

MWF 1135-1230, Astronomy Conference Room

Instructor:

Paul Hickson
Department of Physics and Astronomy
Room 440 Earth & Space Sciences
email: hickson@physics.ubc.ca

Background assumed:

physical optics, introductory statistics

References:

1. Lena, Lebrun and Mignard, Observational Astrophysics, Springer Verlag 1998
2. Born and Wolfe, Principles of Optics, Pergamon 1980
3. Bracewell, The Fourier Transform and its Applications, McGraw-Hill 1978
4. Lang, Astrophysical Formulae, Springer 1980
5. Tucker, Radiation Processes in Astrophysics, MIT Press 1975
6. Walker, Astronomical Observations, Cambridge 1987
7. Wilson, Reflecting Telescope Optics I, Springer 1996

COURSE OUTLINE

1. **Radiation**

   properties of radiation – photons, coherence
   characterization of radiation – units, photometric systems
   types and sources of radiation – thermal, nonthermal

2. **Detectors**

   fundamentals – coherent and incoherent detection, photon statistics
   measurement – signal-to-noise ratio, DQE, RQE
   single channel detectors – optical, IR, X-ray
   multichannel detectors – CCDs, multiplexed arrays, intensifiers
   coherent detection – radio

3. **Telescopes**

   introduction – principles, types
   image formation – Gaussian imaging, aberrations
   modern telescopes – optical and mechanical design, active optics

4. **Imaging**

   linear theory – OTF, MTF,
   diffraction – Fourier optics
   atmospheric turbulence – theory, propagation effects
   adaptive optics – principles and techniques
   interferometry – Fizeau and Michelson, aperture synthesis

5. **Spectroscopy**

   fundamentals – principles, parameters
   spectrographs – conventional, echelle, MOS, FTS, Fabry Perot

Table of Contents

1 Radiation

This course focuses primarily on the techniques of detection, measurement and interpretation of electromagnetic radiation from astronomical sources. While this is a broad scope, it does not entirely exhaust the opportunities available to observational astronomers. Important results are also obtained by measurements of fluxes of particles such as cosmic rays, neutrinos, and the solar wind. Observational limits on gravitational waves are approaching theoretically interesting levels. However, the great bulk of our knowledge of the universe comes from electromagnetic radiation.

1.1 Properties of radiation

In order to interpret observations and draw conclusions in an astrophysical context, it is necessary to have a basic understanding of the physics of the radiation itself. We therefore begin with a brief review of the fundamental properties of electromagnetic radiation.

1.1.1 Photons

The fundamental quantity of electromagnetic radiation is the ***photon***. A photon is uniquely characterized by its ***energy*** E, ***momentum*** **p** and ***spin*** or ***polarization*** vector **e**. More than one photon can have the same values of these quantities. Such photons are indistinguishable. The energy of a photon is proportional to its ***frequency*** ν (the frequency of oscillation of the associated electric and magnetic fields),

$$E = h\nu \ . \tag{1.1}$$

The constant of proportionality is called ***Planck's constant*** and has the approximate value $h = 6.626\times10^{-34}\,\mathrm{Js}$. Alternatively, we may write

$$E = \hbar\omega \tag{1.2}$$

where $\omega \equiv 2\pi\nu$ is the ***angular frequency*** of the photon and $\hbar \equiv h/2\pi$.

There is also a wavelength λ associated with every photon. The product of wavelength and frequency is a constant c, the speed of light in vacuum.

$$\lambda\nu = c \tag{1.3}$$

The ***propagation vector* k,** also called the ***wave vector*,** is a vector which points in the direction of propagation and has magnitude

$$k = \frac{2\pi}{\lambda} \ . \tag{1.4}$$

The wave vector is related to the momentum of the photon,

$$\mathbf{p} = \hbar\mathbf{k}\ . \tag{1.5}$$

These relations can readily be written in four-dimensional form. The four-dimensional wave vector, defined by

$$\vec{k} \equiv (\omega c, \mathbf{k}) \tag{1.6}$$

is proportional to the four-momentum

$$\vec{p} \equiv (E/c, \mathbf{p}) = \hbar\vec{k}\ . \tag{1.7}$$

The photon has two orthogonal spin states, which we denote by $\left|+\right\rangle$ and $\left|-\right\rangle$, corresponding to clockwise and counterclockwise rotation of the electric vector about the direction of propagation. A photon in either of these two states is said to have ***circular polarization***. A photon may also be in a linear superposition of these two states

$$\left|\mathbf{e}\right\rangle = a\left|+\right\rangle + b\left|-\right\rangle, \quad a^2 + b^2 = 1 \tag{1.8}$$

where a and b are complex numbers.

States with equal weighting ($\left|a\right| = \left|b\right| = 1/\sqrt{2}$) are said to be ***linearly polarized*** because the electric vector oscillates in a fixed plane.

When photons travel through a transmitting medium, interactions with the atoms or ions in the medium cause phase shifts that change the phase velocity of the radiation. In this case (1.3) becomes

$$\lambda\nu = nc \tag{1.9}$$

where n is the index of refraction of the medium. The value of n is frequency-dependent and is generally less than one.

In a medium, the vacuum relation $k = \omega / c$ is replaced by the nonlinear equation $k = \omega / n(\omega)c$ called the ***dispersion relation***. The radiative energy propagates at a speed given by the group velocity

$$\upsilon_g = \left(\frac{dk}{d\omega}\right)^{-1} \tag{1.10}$$

The group velocity never exceeds the speed of light in vacuum.

1.1.2 Coherence

In general, radiation that we observe is not monochromatic, but contains a range of frequencies $\Delta\nu$. The electromagnetic waves will therefore lose phase coherence (the ability to interfere) over a *longitudinal* distance

$$\Delta z \approx \lambda^2 / \Delta\lambda = \lambda R \qquad (1.11)$$

called the ***temporal coherence length***. Interferometers, which combine radiation traveling along two or more different paths, must therefore ensure that the path lengths are equal to within a fraction of this distance – which is just a few wavelengths for broadband radiation.

In general, radiation that we observe does not not originate from a perfect point source, but from a source having some finite angular size $\Delta\theta$. The wavefronts of such radiation are tilted with respect to each other by angles as large as $\Delta\theta$. Therefore, phase coherence is lost over a *transverse* distance

$$\Delta x \approx \lambda / \Delta\theta \qquad (1.12)$$

called the ***spatial coherence length***.

Note that both of these relations can be obtained from the Heisenberg uncertainty relations by considering the photon to be in a superposition of states having an uncertainty in the energy of $\Delta E = h\Delta\nu$ and an uncertainty in the transverse component of the momentum of $\Delta p_x = \hbar k \Delta\theta$.

1.1.3 Flux and intensity

The primary physical quantities of interest to astronomers are summarized in Table 1.1.

Table 1.1 **Fundamental Radiation Measures**

Name	Symbol	Units	Description
Flux	F	$\mathrm{Wm^{-2}}$	Radiant power per unit ⊥ area
Specific Flux	F_ν	$\mathrm{Wm^{-2}Hz^{-1}}$	Flux per unit frequency interval
Intensity	I	$\mathrm{Wm^{-2}sr^{-1}}$	Flux per unit solid angle
Specific Intensity	I_ν	$\mathrm{Wm^{-2}sr^{-1}Hz^{-1}}$	Intensity per unit frequency interval

Flux generally is dependent on position. It usually decreases with distance from the source. ***Intensity***, however, is affected only by absorption or emission of light and by frequency shifts (Doppler effect). If there is no emission or absorption along the line of site, it follows from Liouville's Theorem (the phase space density of photon number is conserved) that the quantity I_ν/ν^3 is constant.

A common unit of specific flux is the ***Jansky*** (Jy). One Jansky is defined to be 10^{-26} $\mathrm{Wm^{-2}Hz^{-1}}$.

Radio astronomers frequently use the terms ***flux density*** (S) and ***brightness*** (B) to refer to specific flux and intensity.

For F_ν and I_ν, we can define corresponding quantities F_λ and I_λ that depend on wavelength rather than frequency. These have units of flux and intensity per unit wavelength.

Because the power contained between two frequencies ν and $\nu + d\nu$ must be the same as the power contained between the corresponding wavelength intervals λ and $\lambda + d\lambda$, we have the relations

$$\begin{aligned} |F_\nu d\nu| &= |F_\lambda d\lambda| \\ |I_\nu d\nu| &= |I_\lambda d\lambda| \end{aligned} \tag{1.13}$$

Using (1.2), we obtain

$$\begin{aligned} F_\lambda &= \frac{\nu^2}{c} F_\nu , \\ I_\lambda &= \frac{\nu^2}{c} I_\nu \end{aligned} \tag{1.14}$$

The quantities $\nu F_\nu = \lambda F_\lambda$ are often used, particularly when comparing the power emitted by a source at different wavelengths.

1.1.4 Radiative transfer

When radiation travels through an emitting or absorbing region, the intensity changes. The rate of change of intensity with distance s is given by the ***equation of radiative transfer***,

$$\frac{dI_\nu}{ds} = -\mu I_\nu + j_\nu \tag{1.15}$$

The first term on the right hand side represents loss by absorption. The energy lost must be proportional to the energy present in the radiation. The constant of proportionality $\mu(\nu)$ is called the ***absorption coefficient***. The second term represents emission of radiation into the beam. This is independent of the intensity and is described by the ***emission coefficient*** j_ν.

The integral of the absorption coefficient along the line of sight is called the ***optical depth***,

$$\tau = \int \mu ds \tag{1.16}$$

If τ is much greater than unity, most of the radiation is absorbed and we say that the absorbing region is ***optically-thick*** (opaque). If If τ is much less than unity, little of the radiation is absorbed and we say that the absorbing region is ***optically-thin*** (transparent).

1.1.5 Astronomical magnitudes

It is convenient to use a logarithmic measure of flux called ***astronomical magnitude***. Magnitudes were first introduced by Hipparchos and Ptolemy to describe the relative brightness of stars (the human eye has a logarithmic response to light). The modern magnitude scale was defined by Pogson in 1854. He found that a logarithmic base equal to the fifth root of 100, which has a value of about 2.51, matched the ancient Greek system quite well. The zero point of the magnitude system is chosen in such a way that the magnitude of a particular ***standard star*** is exactly zero. This leads to the definition

$$m = -2.5\log(F / F^{Vega}) \qquad (1.17)$$

where

$$F \equiv \int F_{\nu} W(\nu) d\nu \qquad (1.18)$$

and the function $W(\nu)$ describes the response of the measuring instrument as a function of the frequency of the incident radiation. Vega is a star (spectral type A0V) defined to have $m = 0$ in all wavelength bands (Vega has now been superceded by a set of several standard stars).

By analogy, on can define a logarithmic measure of intensity called ***surface brightness*** and usually denoted by the symbol μ.

$$\mu = -2.5\log(I / I^{Vega}) \qquad (1.19)$$

where I^{Vega} is the intensity that would result if the light from Vega was spread uniformly over 1 square arcsec. of sky.

1.1.6 Photometric bands

The ***wavelength band*** of a photometric system is the spectral region over which the function $W(\nu)$ is large enough to make a significant contribution to the integral in (1.18) Wavelength bands are often characterized by central wavelength and bandwidth. If the bandwidth is sufficiently small, and the specific flux of the object is a slowly varying function of frequency, we can take F_{ν} outside the integral to get a direct relationship between magnitude and the average specific flux (over the wavelength band).

$$m = -2.5\log F_{\nu} + const \qquad (1.20)$$

where *const* denotes a constant which depends both on the choice of wavelength band and the units of F_V. Clearly, the magnitude defined in this way is meaningful only if the wavelength band is specified.

The corresponding relation for surface brightness is

$$\mu = -2.5 \log I_\nu + 26.57 + const \tag{1.21}$$

where the numerical factor converts steradians to square arcsec and *const* has the same numerical value as in (1.20).

Many standard wavelength bands are in use today. Most common is the ***Johnson*** (or "UBV...") system, summarized in Table 1.2, which divides the optical and near-infrared spectral region into nine relatively broad bands. The ***resolving power***, defined by

$$R = \lambda / \Delta\lambda \tag{1.22}$$

where λ is the central wavelength and $\Delta\lambda$ is the bandwidth, is about 5 for the Johnson system. Figure 1.1 shows the wavelength response of the Johnson UBV filters.

Table 1.2. Johnson Photometric Bands

Band	Central λ (um)	Bandwidth (um)	F_V *(m = 0)* (Wm^{-2}Hz^{-1})	$m_{AB} - m$
U	0.36	0.07	1.88 e-23	0.71
B	0.44	0.10	4.65 e-23	-0.27
V	0.55	0.09	3.95 e-23	-0.09
R	0.70	0.22	2.87 e-23	0.25
I	0.90	0.24	2.24 e-23	0.52
J	1.25	0.38	1.77 e-23	0.78
K	2.2	0.48	6.29 e-24	1.90
L	3.4	0.70	3.12 e-24	2.66
M	5.0	1.2	1.83 e-24	3.24

Differences between magnitudes obtained in different bands, for the same source, are called ***colors***. These are clearly a measure of flux ratios between different wavelengths, and are therefore related to the slope of the spectrum.

In practice, the spectral response of the telescope, detector and filters that one uses is not exactly the same as that used to define the standard system. But by observing a number of standard stars, along with the objects of interest, it is possible to transform the observed ***instrumental magnitudes*** to the standard system. This is a part of the technique of photometric data reduction.

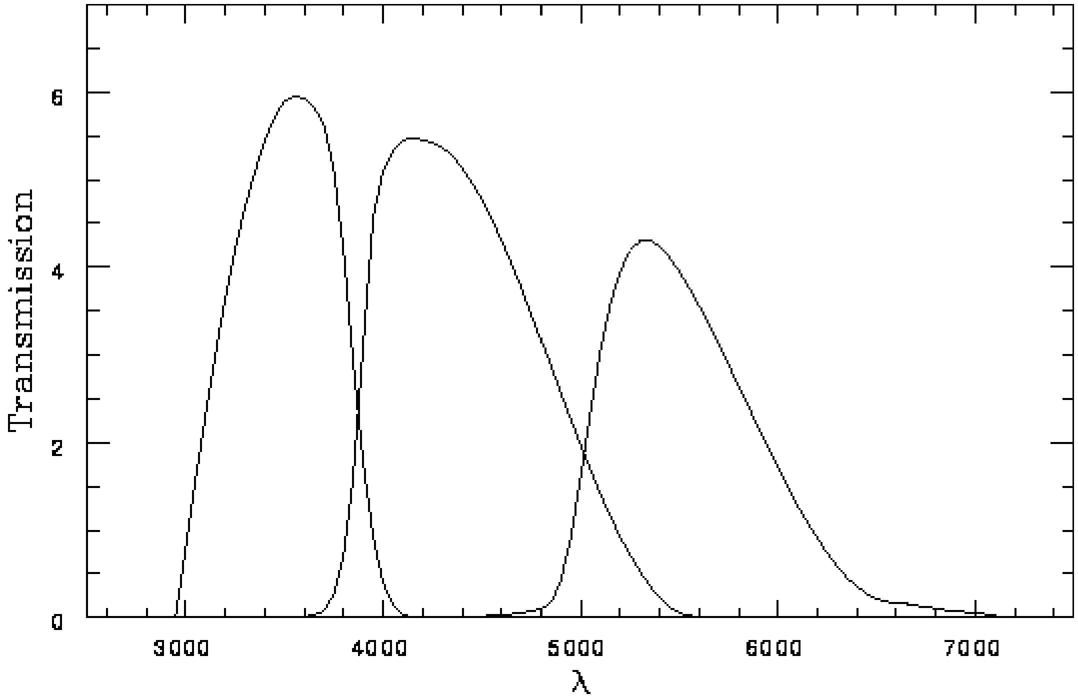


Figure 1.1. Transmission curves of the Johnson U, B and V filters. (Adapted from Johnson, 1952, ApJ 114, 522.)

1.1.7 AB magnitudes

While standard photometric systems provide a convenient way to measure broadband fluxes and colours of objects, they are not well suited to observations at higher spectral resolution. The measurement of flux as a function of frequency or wavelength is called ***spectrophotometry***. The result can be described by specifying F_ν or I_ν as a function of frequency, or F_λ *or* I_λ as a function of wavelength. However, it is often more convenient to use a logarithmic measure. For this reason, the AB magnitude system is defined by

$$m_{AB} = -2.5 \log F_\nu - 56.1 \tag{1.23}$$

The constant is chosen so that the AB magnitude of Vega is zero at a wavelength of 550 nm. It therefore corresponds approximately to the Johnson V magnitude at that wavelength. Like F_ν, the AB magnitude is a continuous function of frequency or wavelength. It deviates from the broadband magnitudes at other wavelengths.

One milli-Jansky (mJy) corresponds to $m_{AB} = 16.4$, one micro-Jansky (uJy) to $m_{AB} = 23.9$ and one nano-Jansky (nJy) to $m_{AB} = 31.4$. The approximate differences between AB magnitudes and Johnson magnitudes are listed in the last column of Table 1.2.

1.2 Types and sources of radiation

The spectrum (the flux as a function of frequency or wavelength) of radiation is an important diagnostic of physical conditions in the source. The following types of radiation are most important in astrophysics. A detailed description can be found in Tucker, *Radiation Processes in Astrophysics.*

1.2.1 Line radiation

Electronic transitions within atoms result in the emission or absorption of radiation at specific frequencies, primarily in the visible, UV and near infrared (NIR) regions of the spectrum. ***Emission lines*** are produced by a hot gas and ***absorption lines*** occur when radiation with a continuous spectrum passes through a gas. The frequency of the line is related to the transition energy *E* by (1.1). Examples of emission and absorption lines in stellar spectra are shown in Figure 1.2.

The lines have an intrinsic width Δν that is inversely related to the lifetime Δ*t* of the excited state according to the Heisenberg uncertainty principle:

$$\Delta E \Delta t \approx h \tag{1.24}$$

therefore,

$$\Delta \nu = \frac{\Delta E}{h} \approx \frac{1}{\Delta t} \tag{1.25}$$

Further broadening occurs from atomic motions, electrostatic perturbations, Doppler shifts, etc. Analysis of the resulting spectra may provide information about the chemical composition, temperature, pressure, velocities, and excitation state of the emitting or absorbing region.

Rotational and vibrational transitions in molecules produce lines in the infrared region of the spectrum. These provide an important diagnostic of the densities and temperatures of molecules in interstellar space.

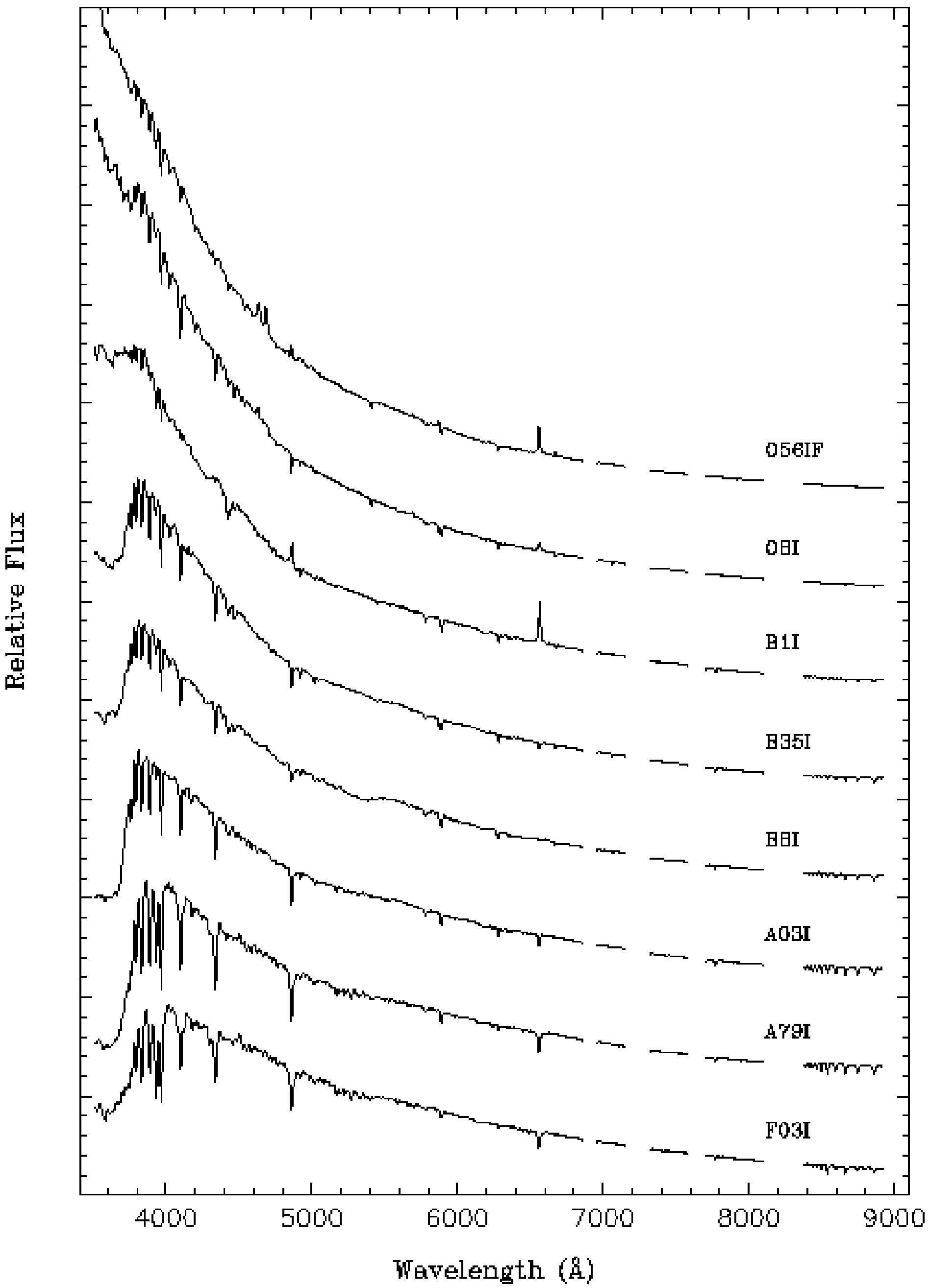


Figure 1.2 Spectra of hot stars, showing absorption and emission lines. The Balmer series of Hydrogen is particularly prominent. (Silva & Cornell, 1992, ApJS, 81, 865).

1.2.2 Black-body radiation

Black body radiation is emitted by any perfectly absorbing body or gas cloud that is in thermal equilibrium at some temperature *T*. The specific intensity of this radiation is given by the ***Planck formula***

$$I_\nu = \frac{2h\nu^3}{c^2}\left[\exp(h\nu/kT)-1\right]^{-1} \tag{1.26}$$

where *k* is Boltzmann's constant ($k \approx 1.381\times10^{-23}\,\mathrm{JK}^{-1}$).

By integrating this expression over frequency and solid angle, one obtains the flux passing through a surface,

$$\begin{aligned} F &= \int d\nu \int_{2\pi} d\Omega\, I_\nu \cos\theta \\ &= \pi I \\ &= \sigma T^4 \end{aligned} \tag{1.27}$$

where $\sigma = 2\pi^5k^4/15h^3c^2 \approx 5.670\times10^{-8}\,\mathrm{Wm^{-2}K^{-4}}$ is the ***Stefan-Boltzman constant***. The cosine factor is needed because I_ν is energy per steradian per unit area *perpendicular* to the direction of propagation.

For isotropic radiation, the energy density is simply related to the intensity. I_ν is just the energy per square metre per second contained in photons of frequency ν for any particular direction. In one second, these photons travel a distance of *c* metres and therefore contribute I_ν/c to the energy density. Integrating over all directions (4π steradians) gives the energy density per unit frequency

$$\rho_\nu = \frac{4\pi}{c} I_\nu. \tag{1.28}$$

The total energy density follows by integration over frequency,

$$\rho = \frac{4\pi}{c} I = \frac{4\sigma}{c} T^4 \tag{1.29}$$

The most celebrated astrophysical example of black-body radiation is provided by the cosmic microwave background (CMB), which is a relic from a time when the universe was both hot and opaque. This radiation is observed to have a perfect black-body spectrum (to within the measurement errors) with a characteristic temperature of 2.73K.

1.2.3 Bremsstrahlung

In an ionized gas, electrons move freely, but are subject to electrostatic deflection when they pass near ions. Such an encounter results in a net acceleration of the electron in a direction perpendicular to its motion. As a result, electromagnetic radiation is produced, known as ***bremsstrahlung*** (German for "braking radiation") or ***free-free*** emission.

Bremsstrahlung has a flat spectrum (nearly constant F_ν) over a wide frequency range. At high frequencies, $\nu \sim kT/h$, there is an exponential cutoff because the photon energies are too high for the photons to be easily produced. At low frequencies, the gas becomes ***optically thick*** (opaque) due to the inverse process (in which a photon is absorbed by an electron in the electric field of an ion). At these frequencies the radiation has the same spectrum and intensity as a black body.

The X-rays emitted by clusters of galaxies is due to bremsstrahlung, and indicates the presence of intracluster gas at a temperature of order 10^6 K.

1.2.4 Synchrotron radiation

Synchrotron radiation is produced by relativistic electrons (electrons moving with speed $v \approx c$) in a strong magnetic field. The Lorentz force causes the electrons to move in a helical orbit along magnetic field lines. The constant acceleration produces radiation. If the magnetic field is orderly, the resulting radiation is often highly polarized.

Synchrotron radiation is found to have a power-law spectrum over a wide range of frequencies. The ***spectral index*** α is defined by the equation

$$F_\nu \propto \nu^{-\alpha} . \tag{1.30}$$

The power law spectrum is a reflection of the distribution of electron energies, which is generally well represented by a power law $f(E) \propto E^{-\beta}$. If β is the index of the electron energy distribution, then it can be shown that $\alpha = (\beta - 1)/2$. The highest energy electrons produce the highest energy photons and vice versa. They also radiate the most power and therefore cool faster. The power radiated by an electron of energy $\gamma m_e c^2$ is

$$P = \frac{32\pi}{9} r_0^2 c \rho_B \gamma^2 , \tag{1.31}$$

where m_e is the electron mass, γ is the Lorentz factor, $r_0 = e^2 / m_e^2 c^2$ is the classical electron radius, e is the electron charge and $\rho_B = B^2 / 8\pi$ is the energy density in the magnetic field B.

At low frequencies the radiation is self absorbed (by the inverse process) and the cloud is optically thick. In this region the spectrum has the form

$$F_\nu \propto \nu^{5/2} \tag{1.32}$$

This differs from the black-body spectrum because the radiation is ***non-thermal*** (the electron gas is not in thermal equilibrium). Of course ions also contribute to synchrotron radiation, but because their masses are much greater than those of the electrons their acceleration is much smaller so they radiate comparatively little power.

Synchrotron radiation is important in many astrophysical objects. It forms the bulk of the radiation emitted by radio galaxies, relativistic jets, and some supernova remnants, such as the Crab Nebula.

1.2.5 Inverse Compton radiation

Inverse Compton radiation is produced when energetic electrons collide with low-energy photons, thereby boosting them to higher energy. The CMB provides a copious source of target photons (about 400 per cubic centimeter in the Universe today). Like synchrotron radiation inverse Compton radiation has a power-law spectrum with a turndown at high energies (the quantum cutoff) and at low energies (self absorption due to the inverse process – the Compton effect). The power radiated by a single electron of energy $\gamma m_e c^2$ is

$$P = \frac{32\pi}{9} r_0^2 c \rho_\gamma \gamma^2 , \tag{1.33}$$

where ρ_γ is the energy density of target (CMB) photons.

Note the similarity of this formula with that for synchrotron radiation. This is not accidental – the electromagnetic force results from the exchange of virtual photons, so synchrotron radiation can also be regarded as resulting from collisions between electrons and (virtual) photons.

Inverse Compton radiation is important in the most energetic astrophysical phenomena. Quasars, active galactic nuclei, gamma ray bursts may radiate much of their energy by this process.

Exercises

1.1 A perfect polarizing filter transmits only linearly polarized light with electric vector parallel to the direction of polarization of the filter. Suppose such a filter is inserted in a beam of unpolarized light (light having no net polarization) of intensity *I*.

a) What is the intensity of the light transmitted by the polarizer.
b) Suppose that a second polarizer is inserted with polarization direction oriented 90 degrees to that of the first polarizer. What is the intensity of light transmitted by the second polarizer?
c) Suppose now that a third polarizer is inserted between the two, with polarization direction oriented 45 degrees to each of the others. What is the intensity of light transmitted by the entire system?

1.2 Derive equations (1.11) and (1.12) using the Heisenberg uncertainty principle

1.3 Liouville's theorem states that during collisionless evolution, the phase space density of particles in a gas remains constant. Show that the quantity I_ν / ν^3 is proportional to the number density of photons in phase space and is therefore conserved by all processes that neither create nor destroy photons.

1.4 Using (1.26), determine the frequency and wavelength that correspond to the maximum intensity of the CMB. Estimate the number of CMB photons per cubic centimeter in the Universe today.

1.5 Light propagating to an observer on Earth is partially absorbed by dust particles in the atmosphere, a phenomena called ***extinction*** The extinction is wavelength dependent, being greatest at short wavelengths (which is why sunsets are red). Ignore the curvature of the Earth, and assume that the dust density in the atmosphere is a function only of altitude. Using the equation of transfer, show that the extinction can be represented by the formula $m = m_0 + k \sec Z$, where m is the observed magnitude, m_0 is the true magnitude (ie. in the absence of extinction), k is a wavelength-dependent constant called the ***primary extinction coefficient***, and Z is the ***zenith angle*** (the angle between the line of sight to the object and the vertical direction).

2 Detectors

Detectors are devices used by astronomers to convert radiation (light, radio waves, etc) into more-readily measurable quantities such as electric charge or current. Detectors invariably add noise to the signal. A good detector is one that responds to radiation with high efficiency and adds very little noise. As we shall see, there is a wide variety of types of detectors. Successful observations rely on selecting the appropriate detector for the type of measurement. This chapter begins with a general analysis of detector performance and characterization and then proceeds to a discussion of the main types of detectors in common use.

2.1 Fundamentals

We begin with a consideration of what a detector does and how it should be characterized. This requires some discussion of the fundamentals of the theory of measurement.

2.1.1 Signals, detectors and noise

A ***signal*** is a quantity or set of quantities containing information. It may take the form of a function of one or more variables, such as the flux of radio waves from a pulsar as a function of time, or the angular distribution of intensity of light arriving from a distant galaxy. Alternatively, it might be represented as a discrete set of one or more numbers, such as the number of photons received from a source within specified time intervals.

Very generally, a ***detector*** is a device that converts information from one form to another. There is therefore an input signal and an output signal associated with every detector (Figure 2.1). As an example, consider a photocell, which is a device that produces an electrical response when light falls upon it. In this case the input signal is the light flux and the output signal is the voltage or current produced by the device.

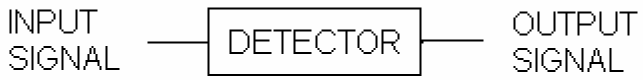


Figure 2.1. A detector converts a signal from one form to another.

Noise is a quantity whose value cannot be predicted in advance. This may result from either random or chaotic behavior. Invariably, signals contain some degree of noise, implying that they consist of a systematic component and a fluctuating noise component. The ratio of the systematic component to the noise component (usually the root-mean-square value) is called the ***signal-to-noise ratio*** (s/n ratio).

2.1.2 Characterization of noise

Statistically, a fluctuating quantity *x* can be completely described by its ***frequency function*** $f(x)$ (also called the ***probability density function*** or ***probability distribution***). The quantity $f(x)dx$ gives the probability that *x* will be found in the range $(x, x+dx)$. Since x must have some value, the integral of the frequency function over the entire range of *x* must give unity,

$$\int_{-\infty}^{\infty} f(x)dx = 1\,. \tag{2.1}$$

Here we have assumed that the range of *x* is the set *R* of all real numbers, but of course other ranges (such as the set of positive real numbers) are also possible.

If x takes only discreet values $(x_1, x_2, ..., x_n)$, then its frequency function is discreet: f_i is the probability that $x = x_i$. The normalization condition becomes

$$\sum_{i=1}^{n} f_i = 1\,. \tag{2.2}$$

In what follows, we shall generally assume that *x* is continuous. To arrive at the discreet just make the substitutions $x \to x_i$, $f(x)dx \to f_i$ and replace the integral with a summation.

The ***expectation value*** of any function $g(x)$ is defined by

$$\langle g(x) \rangle = \int_{-\infty}^{\infty} g(x) f(x) dx \tag{2.3}$$

which clearly obeys the linearity conditions

$$\begin{gathered} \langle g(x) + h(x) \rangle = \langle g(x) \rangle + \langle h(x) \rangle \\ \langle a g(x) \rangle = a \langle g(x) \rangle \end{gathered} \tag{2.4}$$

where *a* is any constant.

A fluctuating quantity can also be described by its moments. The k-th moment of x is just $\left\langle x^k \right\rangle$. Of particular interest is the first and second moments which define the ***mean value*** ($\overline{x}$) and the ***variance*** of x,

$$\begin{aligned} \overline{x} &= \langle x \rangle \\ \mathrm{Var}(x) &= \left\langle (x-\overline{x})^2 \right\rangle = \left\langle x^2 \right\rangle - \overline{x}^2 . \end{aligned} \tag{2.5}$$

The last equality follows from the linearity of the expectation value.

The square root of the variance is called the ***standard deviation*** of x. It is just the root-mean-square (RMS) value of x.

$$\sigma_x = \left(\mathrm{Var}(x)\right)^{1/2} . \tag{2.6}$$

2.1.3 The Gaussian distribution

Many fluctuating quantities are well-represented by the ***Gaussian*** or ***Normal distribution***

$$f(x) = \frac{1}{\sqrt{2\pi\sigma}} \exp\left(-\frac{1}{2\sigma^2}\left(x-\overline{x}\right)^2 \right) \tag{2.7}$$

The Gaussian distribution is completely characterized by the constants $\overline{x}$ and σ which, as is easily confirmed, are equal to the mean value and standard deviation of x.

It can be shown that if a fluctuating quantity results from the cumulative effect of a number of statistically-independent fluctuating quantities, its frequency function is well-approximated by a Gaussian distribution. This is a consequence of the ***central limit theorem*** which states that the frequency function of the sum of a number of random variables converges to a Gaussian distribution in the limit as the number of random variables increases (more details, and a proof, can be found in many texts such as Matthews & Walker 1970, Methods of Mathematical Physics). In practice the Gaussian distribution is a very good approximation even if the number of component variables is relatively small (of order 10) as long as their frequency functions are reasonably smooth.

2.1.4 The Poisson distribution

A useful discreet frequency function is the ***Poisson distribution***

$$P_n(z) = \frac{z^n}{n!} \exp\left(-z\right) \tag{2.8}$$

where $n = x_i$ is a non-negative integer and z is a free parameter.

It is not hard to show that, for the Poisson distribution,

$$\bar{n} = \mathrm{Var}(n) = z\ . \tag{2.9}$$

Therefore

$$\sigma_n = \sqrt{\bar{n}} \tag{2.10}$$

As an example of the use of this distribution, consider a random distribution of objects, galaxies on the sky, for example. (The distribution of galaxies is actually not completely random, but suppose for the sake of illustration that it is.) If the mean number of galaxies per square degree is z, then, in a randomly chosen area of one square degree, the probability of finding exactly n galaxies is $P_n(z)$. For example, the probability of finding no galaxies, when on average z galaxies are expected, is $\exp(-z)$.

2.1.5 Signal-to-noise ratio

The signal-to-noise ratio is a measure of the quality, or information content, of a signal. For example, suppose we have a spectrum (a measure of the specific flux or intensity over a finite frequency or wavelength range) of a distant galaxy and see what appears to be an emission feature (see Figure 2.2). What is the probability that this feature is real (ie not due to noise)? If the amplitude of the feature is a and σ is the RMS noise in the spectrum, then the signal-to-noise ratio is $s = a/\sigma$. We can determine the probability $P(s)$ of finding a "feature" of signal-to-noise ratio greater or equal to s, at that location, by integrating the frequency function of the noise from s to infinity. If the noise has a Gaussian distribution, we obtain, using (2.7),

$$P(s) = \frac{1}{\sqrt{2\pi}} \int_s^\infty \exp(-x^2/2)dx = \frac{1}{2}\mathrm{erfc}\left(\frac{s}{\sqrt{2}}\right) \tag{2.11}$$

$P(s)$, which is the probability that the feature is spurious, is listed in Table 2.1 for several different values of s. Normally, a signal-to-noise ratio of at least 3 is required before one can say with any confidence that a feature is not the result of a noise fluctuation.

Table 2.1. Chance probability vs signal-to-noise ratio

s	$P(s)$
1	0.1587
2	0.0228
3	0.0013
4	0.00003

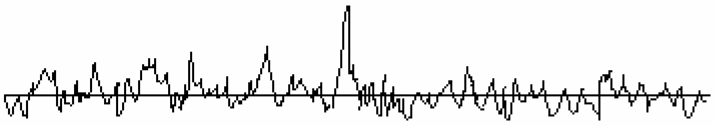


Figure 2.2. Is the peak near the center of the spectrum real or just a noise fluctuation? The answer depends on the signal-to-noise ratio.

2.1.6 Addition of noise

We are often faced with the situation where the noise is the sum of the effects of two or more noise sources, for example $z = x + y$. The properties z can be determined by standard rules for the manipulation of probabilities. For example, probability of finding z in the interval $(z, z + dz)$ is equal to the probability of finding *x* in the interval $(x, x + dx)$ multiplied by the probability of finding *y* in the interval $(z - x, z - x + dz)$, integrated over all possible values of *x*. Thus,

$$f_z(z) = \int_{-\infty}^{\infty} f_x(x) f_y(z - x) dx \qquad (2.12)$$

where the subscripts denote the frequency functions of the individual variables. From this we see that the frequency function of the sum of two random variables is just the convolution of the individual frequency functions. The extension to several variables should be evident – one convolves all of the frequency functions.

Now consider the mean and variance of the sum of several independent fluctuations. From the linearity of the expectation value (2.4) we have

$$\overline{x_1 + x_2 + ...} = \overline{x_1} + \overline{x_2} + ... \qquad (2.13)$$

so the mean value of a sum is just the sum of the mean values. Apply this now to the variance. Using (2.4) and (2.5),

$$\begin{aligned}\mathrm{Var}(x_1+x_2+...) &= \left\langle \left(x_1+x_2+...-\overline{x_1+x_2+...}\right)^2 \right\rangle \\ &= \left\langle \left(x_1+x_2+...-\overline{x_1}-\overline{x_2}-...\right)^2 \right\rangle \\ &= \left\langle \left((x_1-\overline{x_1})+(x_2-\overline{x_2})+...\right)^2 \right\rangle \\ &= \left\langle (x_1-\overline{x_1})^2+(x_2-\overline{x_2})^2+... \right\rangle \\ &= \mathrm{Var}(x_1)+\mathrm{Var}(x_2)+... \end{aligned} \tag{2.14}$$

where the second last step follows because the variables are independent so the expectation values of all the cross terms are zero. Thus we see that when adding independent fluctuating quantities, the *variances* add. This result can be expressed in terms of the standard deviations,

$$\sigma_{x+y+z+...} = \sqrt{\sigma_x^2+\sigma_y^2+\sigma_z^2+...} \tag{2.15}$$

Standard deviations combined in this way are said to be ***added in quadrature***.

2.1.7 Functions of noise

It is frequently necessary to determine the properties of some function of the noise, such as amplification, smoothing, etc. If x is a fluctuating quantity and y is some function such that $y = y(x)$ and $y+dy = y(x+dx)$, then the probability of finding y in the interval $(y, y+dy)$ must equal the probability of finding x in the interval $(x, x+dx)$. Therefore,

$$\begin{gathered}\left| f_y(y)dy \right| = \left| f_x(x)dx \right| \\ f_y(y) = \left| \frac{dx}{dy} \right| f_x(x)\end{gathered} \tag{2.16}$$

which is the law of transformation of probability distributions. The absolute value signs are required because probabilities cannot be negative.

In the special case of a linear function $y = ax+b$, it is easily shown from (2.4) and (2.5) that

$$\begin{gathered}\overline{y} = a\overline{x}+b \\ \mathrm{Var}(y) = a^2\mathrm{Var}(x).\end{gathered} \tag{2.17}$$

2.1.8 Detective quantum efficiency

We are now in a position to discuss quantitatively the “quality” of a detector. Recall that a detector is a device which converts information from one form to another. The information content of the signal is never increased by this, but it can be degraded if the detector adds noise in the process. A good detector is one that adds very little noise. To quantify this, we define the ***detective quantum efficiency*** (DQE, or just *D*) as the squared ratio of the output to input signal-to-noise ratios,

$$D = \left(\frac{(s/n)_{out}}{(s/n)_{in}} \right)^2 . \tag{2.18}$$

The reason for the square will become evident soon. The DQE is the most fundamental measure of the quality of the detection process. It describes the amount by which the signal-to-noise ratio is degraded by the detector. Because the detector can only add noise, not remove it, the DQE is never greater than unity (and in practice is always less),

$$D \le 1 . \tag{2.19}$$

2.1.9 Quantum Efficiency

For detectors which respond to individual quanta of radiation, *ie*. photons, one defines the ***quantum efficiency*** (*Q*) to be the probability that a photon incident on the detector will contribute to the output signal. In other words, *Q* is the fraction of incident photons that are actually detected. (Not all photons are necessarily detected – they can be absorbed, scattered, reflected and not reach the sensitive area of the detector. Also, the charge carriers that they produce may not all reach the output.)

Many detectors respond to incident photons by producing discreet charges, such as an electron or hole in a semiconductor. If the photon energy is sufficiently high, more than one charge carrier may be produced. The average number of charge carriers produced per photon is called the ***quantum yield*** (η) and the product of the quantum yield and quantum efficiency is called the ***responsive quantum efficiency*** (RQE).

$$RQE = \eta Q . \tag{2.20}$$

2.2 Sources of Noise

The information content of a signal is ultimately limited by noise. It is therefore important to examine the primary sources and properties of noise that affects astrophysical measurements.

2.2.1 Photon noise

Because electromagnetic radiation comes in the form of discreet quanta, photons, the flux and intensity of this radiation is quantized. To a good approximation, the arrival times of photons are random. (Under certain conditions, discussed more fully later, photons having nearly the same wave vectors and times are correlated because photons are indistinguishable particles.) They are therefore governed by Poisson statistics. For example, suppose that light from a star is measured, using a perfect detector, and the number of photons detected in a specified time interval is recorded. This observation is repeated many times and the results are compared. The number of photons received each time will vary. This variation is an unpredictable fluctuating quantity and is therefore a source of noise. If x is the average number of photons received, then the probability that exactly n photons will be received in any individual measurement is $P_n(x)$, given by (2.8). The rms variation in the number of photons received in an individual measurement is given by (2.10), $\sigma_n = \sqrt{n}$. Now, if we have only a single measurement, we don't know the mean value $\bar{n}$ of very many measurements, but we can estimate it. The best estimate that we have for $\bar{n}$ is n, the number of photons that were detected in our single measurement. Therefore, our best estimate of the RMS uncertainty in a single measurement of n photons is

$$\sigma_n = \sqrt{n} \,. \tag{2.21}$$

2.2.2 Thermal noise

Associated with every electrical resistance is a fluctuating electrical signal that arises from the thermal motions of electrons in the resistive material. This is referred to as ***Johnson noise*** or ***thermal noise***. If the resistor is coupled to an external load, power will be transferred from the resistor to that load. It is easy to show that the power transferred will be maximum when the impedence (resistance in this case) of the load equals that of the resistor. For simplicity, assume that the load has zero temperature and therefore generates no noise itself. Figure 2.3 shows the equivalent circuit for this arrangement. The power spectral density (electrical power at frequency ν, per unit frequency interval) transferred to the load can be shown to be

$$P_\nu = kTp(\nu, T) \tag{2.22}$$

where

$$p(\nu, T) = \frac{h\nu/kT}{\exp(h\nu/kT) - 1} \tag{2.23}$$

and k is Boltzmann's constant. This result is called ***Nyquist's theorem***. The function $p(\nu, T)$ is very close to unity at low frequencies but approaches zero exponentially when the energy $h\nu$ of a fluctuation exceeds the typical thermal energy kT. Thus, below this cutoff the power spectral density of the fluctuations is practically independent of frequency, a characteristic that is referred to as ***white noise***. In practice, we are only sensitive to fluctuations within a certain frequency range. If $W(\nu)$ represents the relative response of the measuring system at frequency ν (a function whose maximum possible value is unity), then

$$\begin{aligned} P &= \int_0^\infty P_\nu W(\nu)d\nu \\ &= kTB \end{aligned} \tag{2.24}$$

where the bandwidth B is defined by

$$B = \int_0^\infty W(\nu)d\nu \tag{2.25}$$

Now, the mean power can also be written as $P = \left\langle V^2 \right\rangle R/4$, where $\left\langle V^2 \right\rangle$ is the mean square voltage generated by the resistor. The factor of 4 occurs because only half of the voltage appears across the load. This gives the familiar result

$$\left\langle V^2 \right\rangle = 4kTRB\,. \tag{2.26}$$

A common source of noise in electronic amplifiers is the Johnson noise associated with their input resistance.

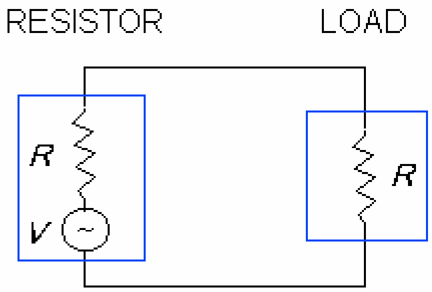


Figure 2.3 Equivalent circuit of a resistor connected to a load. The circle indicates the fluctuating voltage generated by thermal motion of electrons in the resistor.

2.2.3 KTC noise

A closely related form of thermal noise occurs in the measurement of electrical charge. This charge, such as that produced by the action of light on a semiconductor, is stored in some form of capacitor. This produces a voltage which is amplified and measured, as in Figure 2.4. The voltage produced by a charge Q stored on a capacitance C is

$$V = Q/C \tag{2.27}$$

Johnson noise in the amplifier will produce an RMS voltage fluctuation

$$\sigma_V = 2\sqrt{kTRB}\ , \tag{2.28}$$

according to (2.26). Now the capacitor has a reactance $-i/2\pi\nu C$ that decreases with increasing frequency and will attenuate fluctuations above the characteristic frequency ($1/RC$) of the RC circuit. Integration of the power response function gives the effective bandwidth

$$\begin{aligned} B &= \int_0^\infty \left((2\pi\nu RC)^2+1\right)^{-1} d\nu \\ &= \frac{1}{4RC} \end{aligned} \quad . \tag{2.29}$$

Combining (2.27), (2.28) and (2.29) gives the RMS error in the measurement of the charge

$$\begin{aligned} \sigma_Q &= C\sigma_V \\ &= \sqrt{kTC} \end{aligned} \tag{2.30}$$

which is independent of the value of the resistance! This noise is called ***KTC noise***, for obvious reasons.

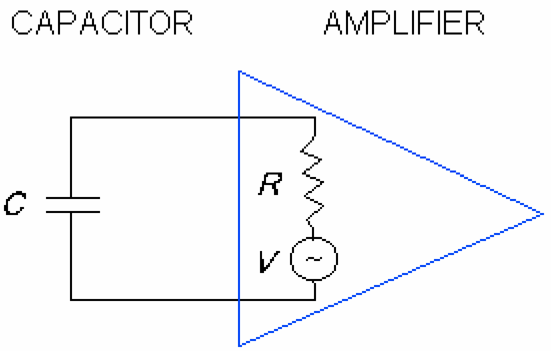


Figure 2.4 Equivalent circuit for the measurement of charge stored on a capacitor. Johnson noise generated in the input resistance of the amplifier causes charge fluctuations on the capacitor with RMS value of $\sqrt{kTC}$.

2.2.4 1/*f* noise

Electronic systems are affected by a type of noise whose power spectral density has the form

$$P_\nu \propto \nu^{-1} \tag{2.31}$$

This noise is called **1/*f*** or ***flicker noise***. It predominates at low frequencies and its origin is unknown. Normally, some form of low-frequency cutoff filter is required to reduce this noise.

2.2.5 Shot noise

When a signal is represented by an electric current, the noise arises because of the discrete nature of the charge carriers. The arrival times of electrons moving in a circuit are random and therefore governed by Poisson statistics. For a steady current *I*, the average number of electrons passing a given point in time *t* is

$$N = It / e \tag{2.32}$$

where *e* is the charge of the electron. The RMS fluctuation in the current will therefore be

$$\begin{aligned} \sigma_I &= \frac{e}{t}\sigma_N \\ &= \frac{e}{t}\sqrt{\frac{It}{e}} \\ &= \sqrt{eIB} \end{aligned} \tag{2.33}$$

where we have used (2.10) and taken $B = 1/t$.

2.2.6 Recombination noise

In semiconductors, the signal is represented by charge carriers (electrons or holes) which move in the conduction band. Occasionally, some of these recombine with atoms in the semiconductor lattice. This random process can sometimes be a significant source of noise in such devices.

2.3 Detectors

There are two main classes of detectors: ***incoherent detectors*** respond only to the energy of the incident radiation; ***coherent detectors*** also provide information about the phase of the radiation. Coherent detectors are available for low-frequency radiation but not for high-frequencies. The boundary between coherent and incoherent detection presently occurs at a wavelength of approximately 1 mm. At higher frequencies, a wavelength of about 10 um or less, individual photons have sufficient energy to create a measurable response in the detector, either by ejection of excited electrons (the ***photoelectric effect***), or by ionization or deposition of energy, as in a scintillator. As coherent detection will be discussed in Section 2.4, the remainder of this section will discuss only incoherent detectors. Even here, the discussion will be limited to those types of detectors most commonly used by astronomers.

2.3.1 Bolometers, pyroelectric detectors and thermocouples

A ***bolometer*** is a device that responds to the total power that it absorbs from the radiation field. Typically, the absorbed radiation causes a change in temperature of the device which results in a change of electrical resistance. Bolometers are frequently used at infrared and sub-millimetre wavelengths, from about 1 mm to 10 um. Figure 2.5 and Figure 2.6 show the bolometer feedhorn array, and an individual germanium bolometer of the SCUBA detector used with the James Clerk Maxwell Telescope (JCMT).

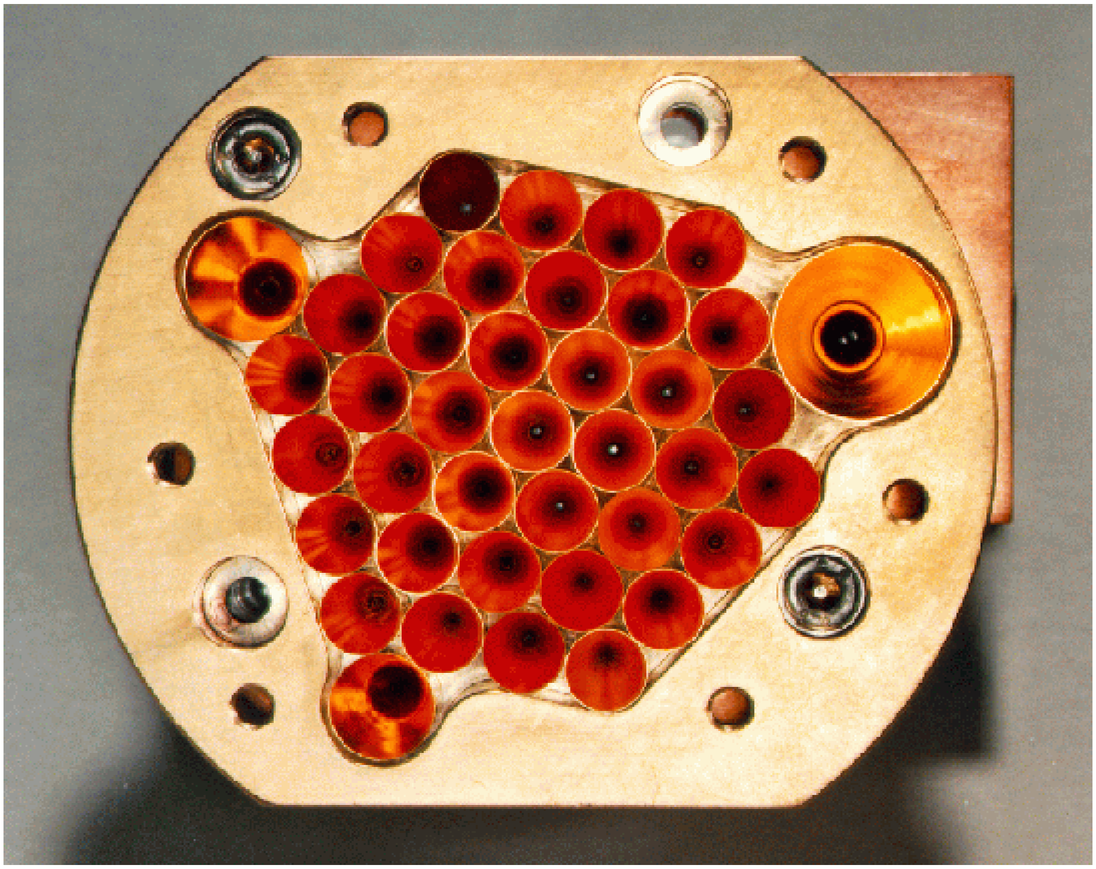


Figure 2.5. The feedhorn array of the SCUBA imaging detector. Each horn channels submillimetre radiation to a germanium bolometer.

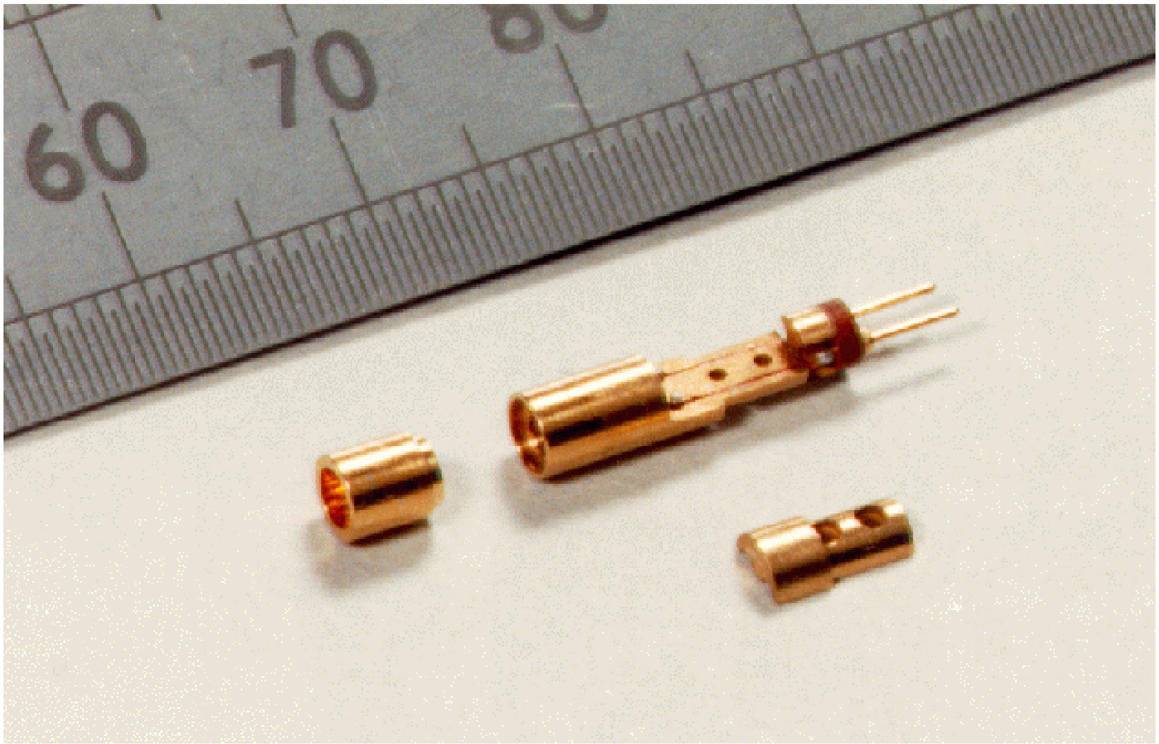


Figure 2.6. Exploded view of a germanium bolometer of the SCUBA array.

Figure 2.7 shows a simple circuit diagram illustrating the operation of a bolometer. The variation in electrical resistance causes a change in current flow in the circuit, which is then amplified and measured. In order to achieve high sensitivity, one needs a material in which a small change in temperature produces a relatively large change in electrical resistance. Semiconductor materials, such as germanium, are often used. In a semiconductor, most electrons are bound to atoms in the lattice and are not able to carry an electric current. If sufficient energy is provided, an electron can be freed from the lattice and move about the material. In terms of energy, such electrons occupy a ***conduction band***. The vacancy left by the electron leaves the atom with a net positive charge. This vacancy can be filled by an electron from a neighboring atom, and thereby propagate through the crystal. It therefore represents a freely-moving positive charge called a ***hole***. The minimum energy E_g required to promote an electron from a bound state to the conduction band is called the ***bandgap energy***. In thermal equilibrium at temperature *T*, the number of thermally excited electrons in the conduction band is proportional to $T^3 \exp(-E_g / k_B T)$. Since the conductivity *G* of the material is proportional to the number of free charge carriers, we have

$$G \propto T^3 \exp(-E_g / k_B T) \tag{2.34}$$

The resistance of the device is inversely proportional to the conductivity. It is this exponential dependence of the conductivity on temperature that gives semiconductor bolometers their high sensitivity.

The primary source of noise in a bolometer is the thermal noise associated with its resistance. For this reason, bolometers are commonly cooled to about 4K, using liquid helium, to reduce this noise.

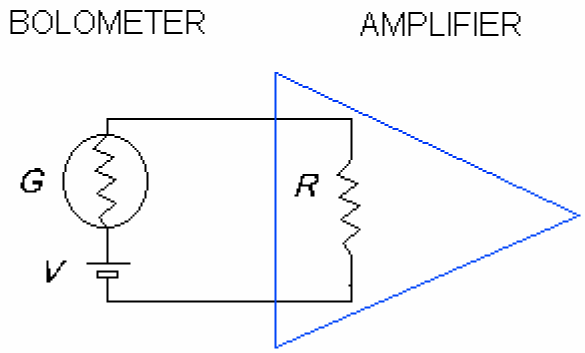


Figure 2.7. Operation of a bolometer. Light falling on the bolometer, illustrated by an ellipse, heats the device which results in a change of conductivity. This changes the current flow in the circuit causing a change in the output voltage of the amplifier.

Two related devices, used primarily for the detection of infrared radiation, are ***pyroelectic detectors*** and ***thermocouples***. Pyroelectric material has a temperature-dependent electrical polarization. When placed between the plates of a capacitor, variations in the polarization create an electrical signal, proportional to the energy absorbed from the radiation field. These devices are operated with a chopping mechanism that interrupts the radiation coming from the source in order to eliminate slow variations (ie less than the chopping frequency, which is typically about 10 Hz) in the polarization and drift in the electronic circuit.

A thermocouple is a junction between dissimilar metals. The differing ***contact potential*** (the energy required to free an electron from the metal, divided by the charge of the electron) for the two metals results in a voltage that can drive an electrical current, a phenomena referred to as the ***Peltier effect***. The amount of current flow depends on the thermal energy provided to the junction. Normally the circuit includes two junctions in series, with opposite polarity, so that the net current is proportional to the temperature difference between the two junctions. One junction is held at a reference temperature, using a liquid helium bath, and the other is exposed to the incident radiation. Like a bolometer, the thermocouple responds to the total absorbed energy. However, thermocouples have not achieved a comparable level of sensitivity.

It is interesting to note that the Peltier effect also works in reverse. If current is supplied to the junction, the temperature will increase or decrease, depending on the direction of current. This effect is exploited by ***thermoelectric coolers*** that are sometimes used to cool detectors.

2.3.2 Photomultipliers

A photomultiplier is a vacuum-tube device containing a ***photocathode***, ***anode***, and several intermediate structures called ***dynodes***, as shown in

Figure 2.8. Electrical potentials applied to these create a strong electric field which accelerates electrons in the direction of the anode. Light incident on the photocathode will cause the ejection of electrons by the photoelectric effect, provided that the photons have energy at least as great as that required to free an electron from the photocathode material. This energy, known as the ***work function*** depends on the composition of the photocathode. Once ejected, an electron is accelerated by the electric field towards the first dynode. It strikes the dynode with sufficient energy to cause the ejection of several secondary electrons. These in turn are accelerated toward the second dynode and the process continues. In this manner, a single photoelectron results in a shower of as many as a million electrons reaching the anode. Because of this electron multiplication, photomultipliers produce a measurable signal even for a single photon (provided that it produces a photoelectron). They are frequently used in photon counting applications.

Photomultipliers come in many shapes and sizes. Some examples are shown in Figure 2.9. The response of a photomultiplier depends on the wavelength of the incident light and the composition and thickness of the photocathode. Examples of quantum-efficiency curves for common photocathode materials is shown in Figure 2.10. In astronomy, photomultipliers were used to conduct photoelectric photometry, which superceded less-accurate photographic photometry for most observations. While photomultipliers can provide high accuracy and dynamic range, they have been almost entirely replaced by modern CCD detectors for optical astronomy. CCDs have typically four times higher peak quantum efficiency than photomultipliers, and also have the advantage of being imaging detectors.

2.3.3 Image intensifiers

Photomultipliers can provide a measure of the total flux, but not the intensity since they provide no information about the location of a photon which strikes the photocathode or the direction from which it came. They are therefore rarely used for imaging applications. (The exception is large arrays of photomultipliers where each corresponds to one pixel in the image, such as the Kameokande detector.) However, there are a number of photoelectric devices that do preserve image information.

A simple ***electrostatic image intensifier*** (Figure 2.11) uses an electron “lens” to direct photoelectrons to a phosphor screen according to their position, much like a pinhole camera. The electrons strike the phosphor with enough energy to cause the emission of several hundred photons. For proper focusing, the photocathode surface must be curved, so a fiber-optic plate is used to channel light from the (flat) input image to the curved photocathode. It is comprised of a large number of parallel optical fibers which conduct the light with little loss of resolution. Two or three such devices and be connected together to yield a higher gain. However, they suffer from strong field distortion and vignetting (discussed in Section 3) and do not produce a high-quality image.

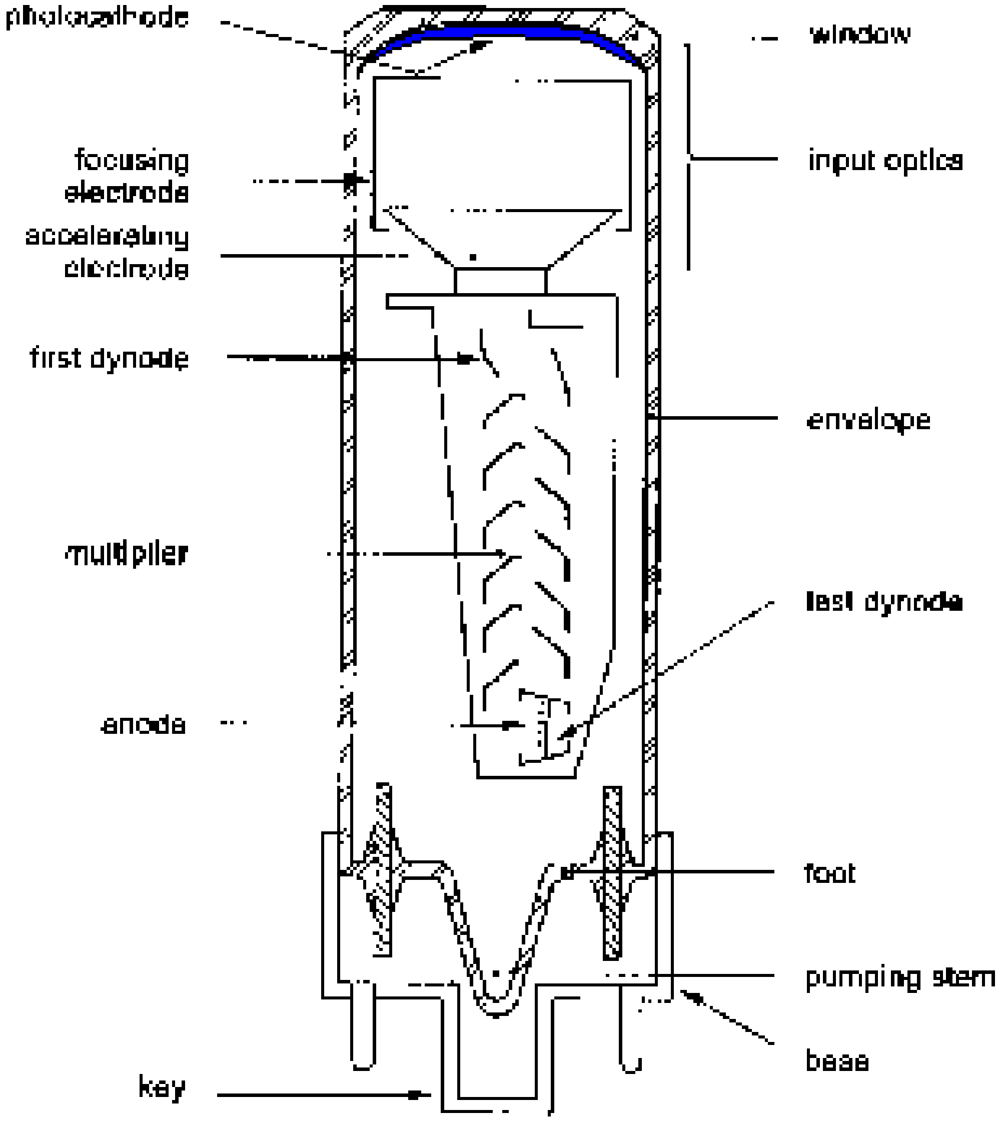


Figure 2.8. Internal components of a typical photomultiplier tube.

Figure 2.9. Examples of different photomultiplier types manufactured by Hammamatsu Corp. The large tube in the picture on the right is one of the large photomultipliers used to detect Cerenkov photons in the Kamiokande neutrino detector.

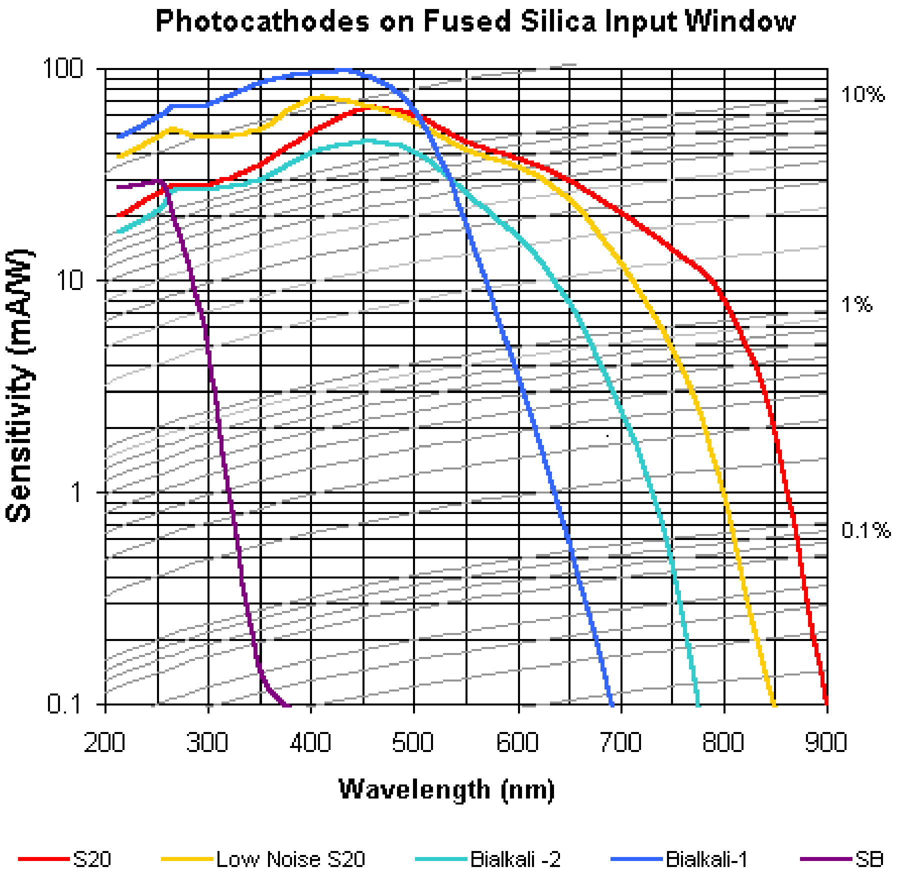


Figure 2.10. Quantum efficiency curves (right hand scale) for common photocathodes.

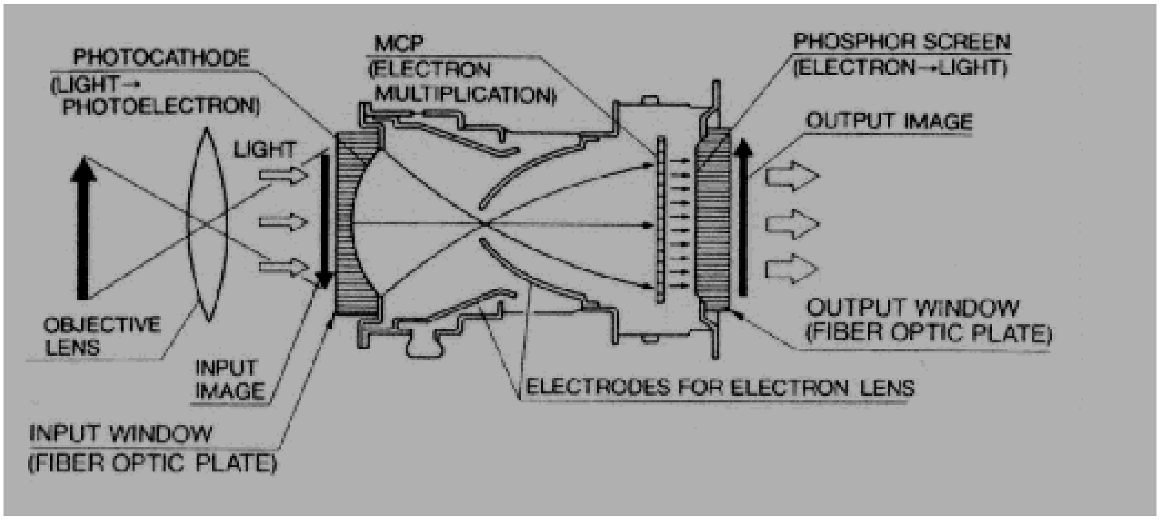


Figure 2.11. Cross section of a single-stage electrostatic image intensifier. The simplest such devices produce an inverted image that is about 10 times brighter than the incident image. The intensifier shown in the figure also contains a microchannel plate, just before the phosphor screen, which provides additional gain.

The ***microchannel intensifier*** provides much better performance in a compact package (Figure 2.12). In this device, a ***microchannel plate*** (MCP) is located in close proximity to the photocathode. This plate consists of a semiconductor material permeated by a large number of channels, closely packed in a hexagonal array (Figure 2.13). A large potential is applied across the plate and photoelectrons that enter a channel and strike the side eject secondary electrons, producing a cascade of electrons in a similar manner to a photomultiplier (Figure 2.14). These electrons strike a phosphor, creating a bright spot. Because the electron cascade is confined to the channel in which the photoelectron entered, the positional information is preserved. Figure 2.15 shows a magnified view of a microchannel plate.

As the figures suggest, electron cascades can be initiated directly by other high-energy particles such as X-ray photons, without the need for a photocathode. Because of this, MCPs are effective detectors of X-rays. It is also possible to eliminate the phosphor screen and detect the electron cascade directly. Devices such as the MAMA detector employ a grid of electrodes that collect the electrons and, with appropriate electronics, can record the position of each electron cascade.

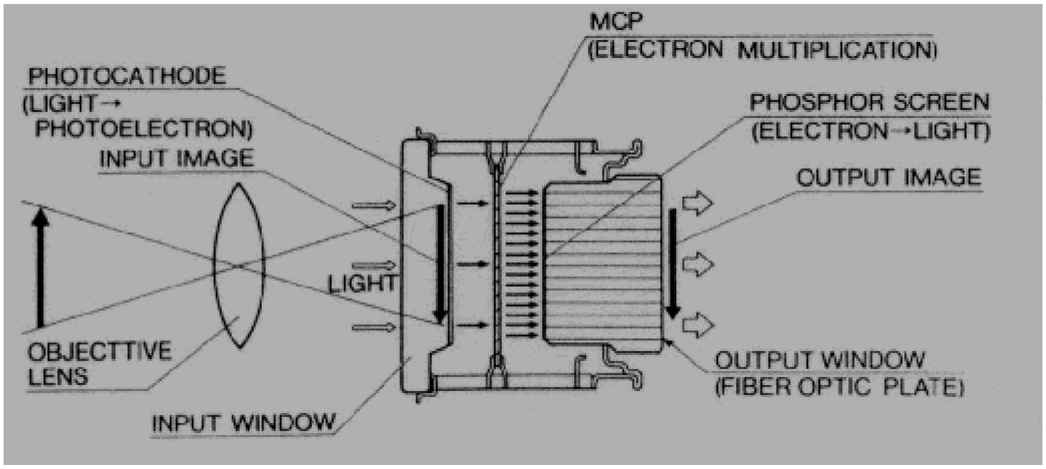


Figure 2.12 Cross section of a microchannel intensifier. Electrons from the photocathode create electron showers in a microchannel plate which then strike a phosphor. The optical gain of such devices rangers from 10^3 to 10^6.

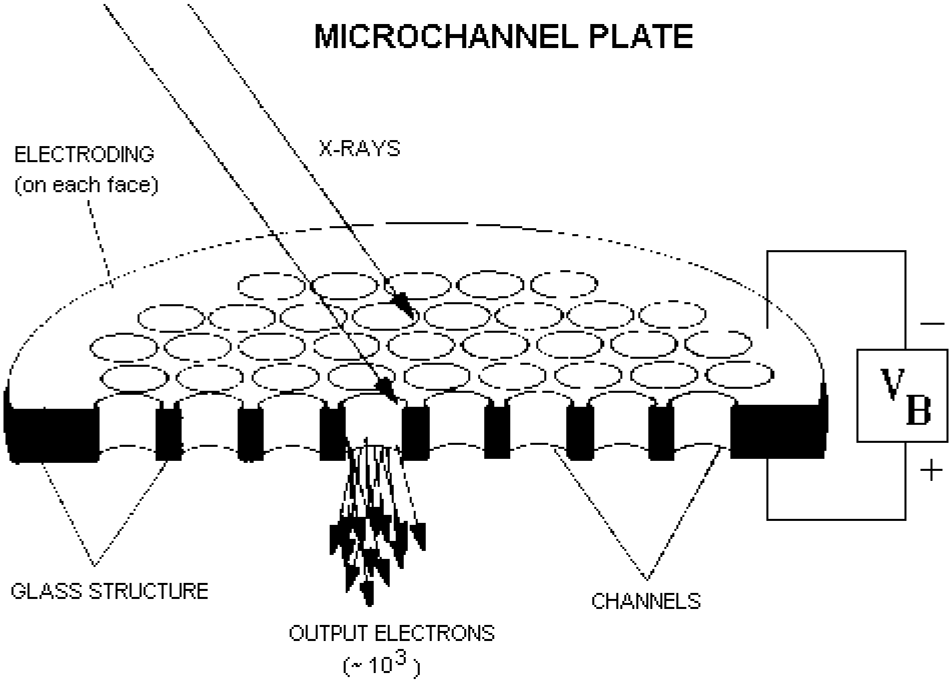


Figure 2.13. Schematic view of a microchannel plate. Electrons, or high-energy photons, striking the front face eject electrons that are amplified by secondary emission in the channels.

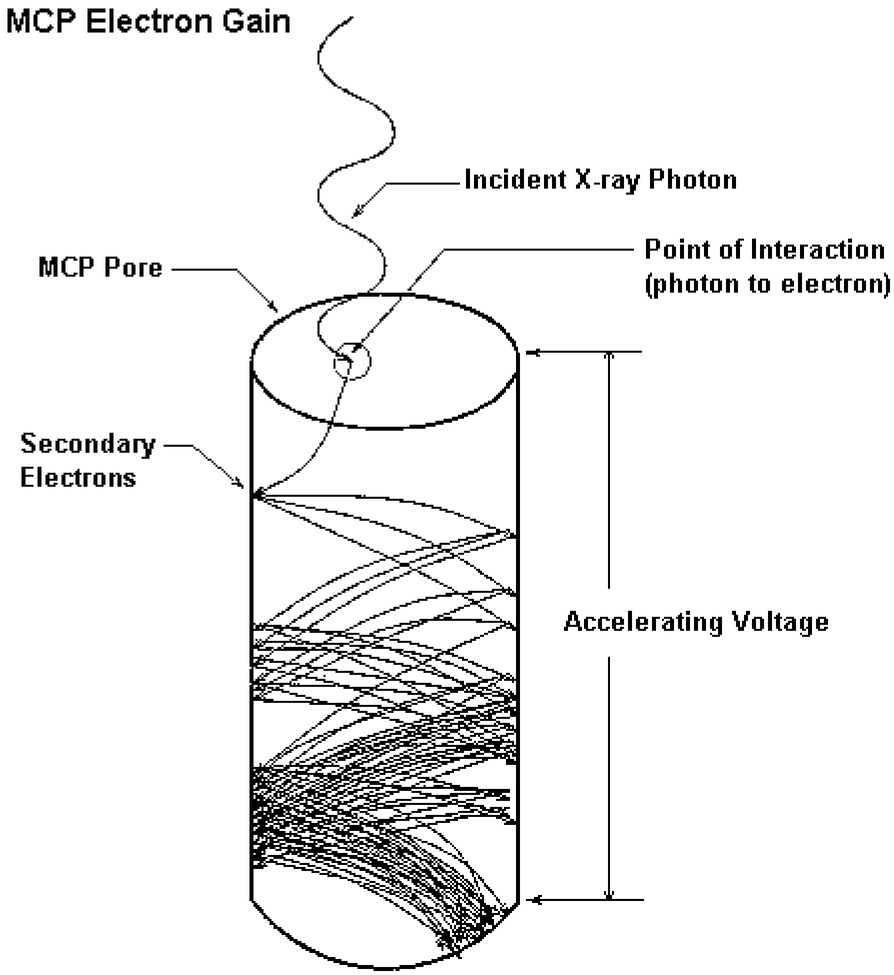


Figure 2.14. Illustration of the electron cascade within a single channel of a microchannel plate. Secondary electrons are produced in collisions with the channel walls.

2.3.4 Photodiodes

The photodiode is perhaps the simplest solid-state photoelectric detector. Semiconductors, such as crystalline silicon, contain a conduction band which normally contains only a few thermally excited electrons. If a small quantity of phosphorus is added to the crystal, a process called ***doping***, phosphorus atoms replace some of the silicon atoms in the lattice. Since phosphorus has a valence of +5, one greater than silicon, it has an extra electron in its outer shell. This electron is weakly bound, requiring only 0.04 eV of energy to enter the conduction band. At normal temperatures almost of these electrons are thermally excited, providing free negative charge carriers. Silicon doped with phosphorus is called ***n-type*** silicon because of these negative charge carriers. The phosphorus atom, on the other hand, now contains a net positive charge that is fixed in the lattice. This charge offsets the negative charge in the conduction band so that the material is electrically neutral overall.

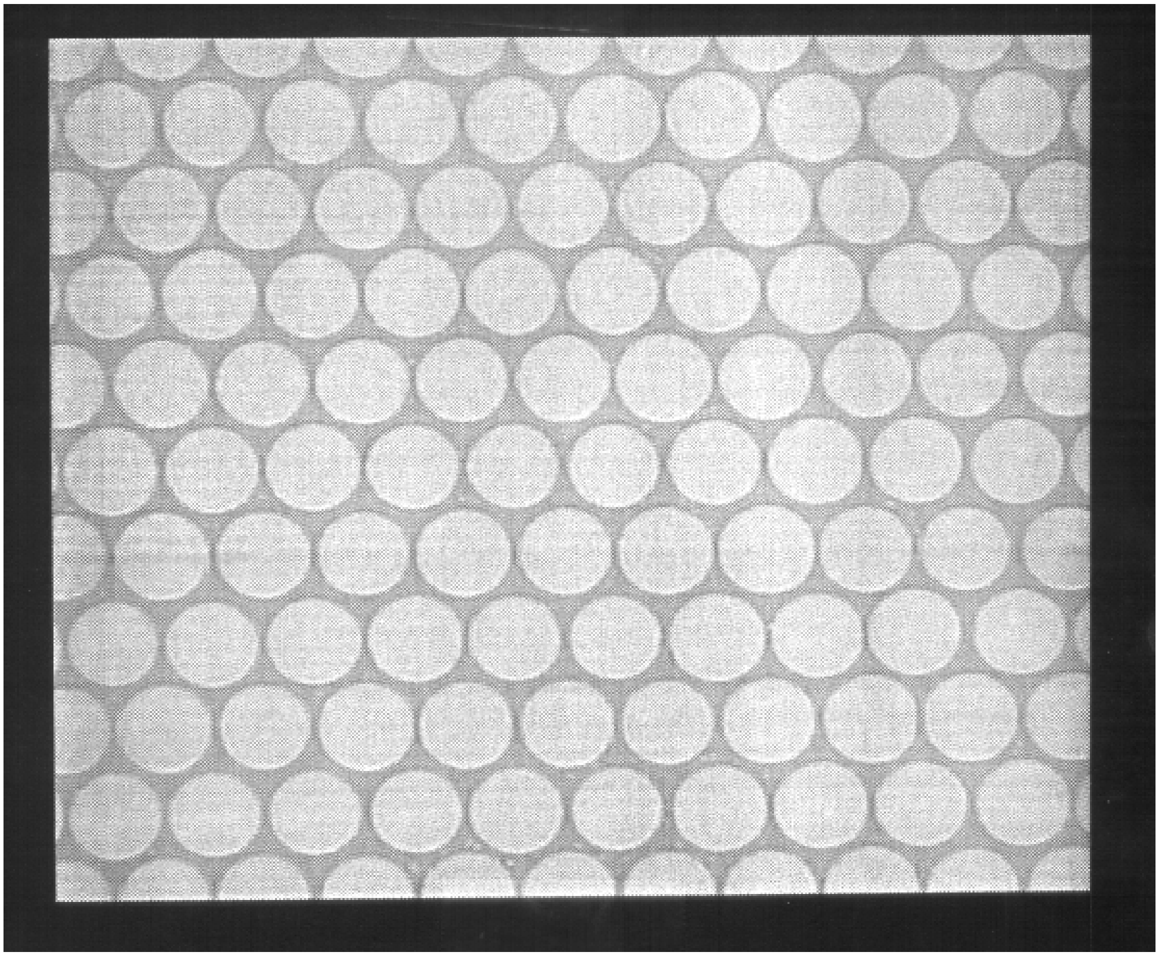


Figure 2.15. Micrograph of the surface of a microchannel plate. The individual channels are seen as light circles that have a diameter of about 12 um.

If boron atoms are used instead of phosphorus, the opposite effect occurs. Boron has a valence of +3, so it provides one less electron than does silicon. An electron from a neighboring atom can be captured here, so the deficit produces a hole, ie. a free positive charge carrier. Boron-doped silicon is therefore said to be ***p-type***. The captured electron provides a negative charge that is not balanced by the nucleus, so each boron atom effectively contributes a fixed negative charge.

If *p*-type and *n*-type silicon (or other semiconductor) are brought together in contact, they form a ***pn junction***. In the absence of any external electrical potential, the charge carriers move about the semiconductor by thermal diffusion. Conduction-band electrons in the n-type material and holes in the *p*-type material diffuse across the junction and recombine. Because of this there is a region surrounding the junction that is depleted of free charge carriers (electrons and holes). It is naturally called the ***depletion region***. Because there are no, or very few, charge carriers in this region, the fixed charges attached to the dopant atoms (phosphorus and boron) are no longer neutralized and the region is electrically charged (see Figure 2.16), with a positive charge in the *n*-type region and a negative charge in the *p*-type region. This charge creates an electric field in the depletion region that opposes the diffusion of charge carriers, thereby creating an equilibrium that sets the size of the depletion region. Because of the electric field, the electrical potential on the *p* side of the depletion region is higher than that on the *n* side. This potential difference is called the ***junction potential***.

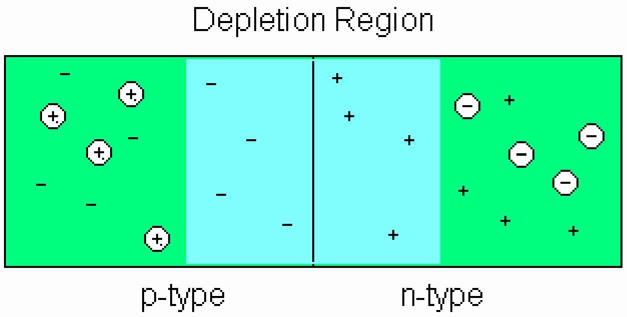


Figure 2.16. A *pn* junction. Circled charges represent free charge carriers (electrons and holes). Un-circled plus and minus signs denote fixed charges associated with dopant atoms. In the depletion region the free carriers have recombined. The remaining fixed charges are now un-neutralized and create an electric field. This gives rise to a potential difference across the junction.

If an external voltage is applied to the junction, by connecting wires to the *p* and *n* sides to a battery for example, the potential difference across the junction can be increased or decreased. If the polarity is such that a positive voltage is applied to the *p*-type material, the junction potential is reduced, or even reversed and current flow across the junction increases. The junction is said to be ***forward-biased***. If the polarity is reversed, the junction potential increases and current flow across the junction decreases. The junction is said to be ***reverse-biased***. This is the principle of the semiconductor diode, which allows current to pass in one direction (forward biased) but not the other (reverse biased).

A photodiode is made by reverse biasing a *pn* junction. In this situation, the current flow is small, being comprised mainly of thermally-excited electrons in the depletion region. If light is allowed to fall on the depletion region, and the photons have sufficient energy, electrons can be photoexcited into the conduction band. These electrons are swept across the junction by the electric field and constitute a current which is proportional to the number of photons per second entering the depletion region. For silicon, the bandgap energy is 1.1 eV, which corresponds to a wavelength of about 1 um. A silicon photodiode can therefore detect optical photons, to a long-wavelength cutoff of about 1 um. In the near-infrared, InSb can be used, as this material has a smaller bandgap energy corresponding to a wavelength cutoff of about 5 um. Photodiodes generally have very high quantum efficiencies, typically 90% or more in the most sensitive wavelength region.

2.3.5 Photodiode detectors

In astronomy, the light that we need to detect is usually very faint and would not produce a significant current. Rather, we are interested in measuring the total charge accumulated during an exposure, since this is directly related to the total number of photons received. In order to do this, the photodiode is connected in a manner shown schematically in Figure 2.17. When an external voltage is applied (reverse bias), current flows momentarily into the diode. This charge combines with electrons or holes, increasing the size of the depletion region. There is therefore an increase of fixed charge, equal to the total charge that entered and recombined, and the junction potential increases to equal the applied voltage. Thus, the diode acts as a capacitor, which stores charge applied to it with a resulting increase in voltage $\Delta V = \Delta q / C$. To make use of this effect to measure light, the photodiode is first charged by applying a potential V by momentarily closing a switch (usually an electronic one). The switch is then opened so that no charge can enter or leave the diode. The diode is now exposed to light. Photons illuminating the semiconductor produce free electrons and holes. These diffuse to the depletion region where they are swept across the junction to the by the electric field, electrons to the n side and holes to the p side. The size of the depletion region thereby shrinks, and the junction potential is reduced in proportion to the charge generated by the photons. At the end of the exposure, the charge is measured by closing the switch and measuring the current that flows to recharge the diode. The amount of charge that must flow to bring the junction potential back to V is exactly that which was produced by the photons. In practice, there is also charge due to thermally excited electrons. This charge depends on temperature and exposure time, but not light intensity. The resulting current is called ***dark current***. The limiting noise in a photodiode is the KTC noise associated with the junction capacitance. This is typically on the order of a few electrons RMS.

One-dimensional arrays of photodiodes can be fabricated on a single crystal, by the lithographic process used to make integrated circuits. Switches, made from metal-oxide semiconductor field effect transistors (MOSFETs), link all the diodes to a common output amplifier. The arrangement is shown in Figure 2.18. The gates of the MOSFETs are connected to a shift register. The diodes can be set and reset sequentially by loading a logical “1” into the shift register, then applying a series of pulses to the clock input. Each pulse causes the 1 to move ahead one position in the shift register, which causes the active output to move in sequence, closing and opening one switch after another. In this manner, an array of N photodiodes can be sequentially read out by a sequence of N pulses. The noise for a photodiode array is greater than that of an individual photodiode because of the large number of MOSFET switches all connected to the ouput line. Each contributes a small capacitance that adds to the total. The output capacitance is proportional to N and the KTC noise is therefore proportional to $N^{1/2}$.

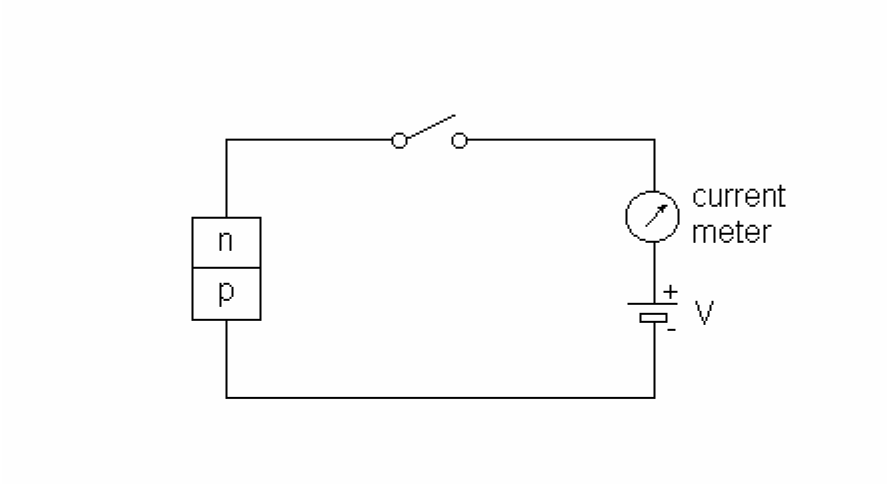


Figure 2.17. Simple photodiode detector. Closing the switch charges the reverse-biased diode to potential V. The switch is then opened and the diode is exposed to light. Photoelectrons neutralize some of the charge on the diode, reducing the potential. At the end of the exposure, the switch is again closed and the current that flows to recharge the diode is measured.

Two-dimensional photodiode arrays can be made by replicating one-dimensional arrays, which then form the rows of an $M \times N$ array. The outputs are all connected to a common amplifier. These have not found much use in astronomy because of their relatively small number of pixels, and high read noise.

One-dimensional, and also $2 \times N$, photodiode arrays have been used effectively by astronomers for spectroscopic applications. Their chief advantages are high quantum efficiency, dynamic range, and linearity. They have now largely been superseded by large-format CCDs, which have lower noise.

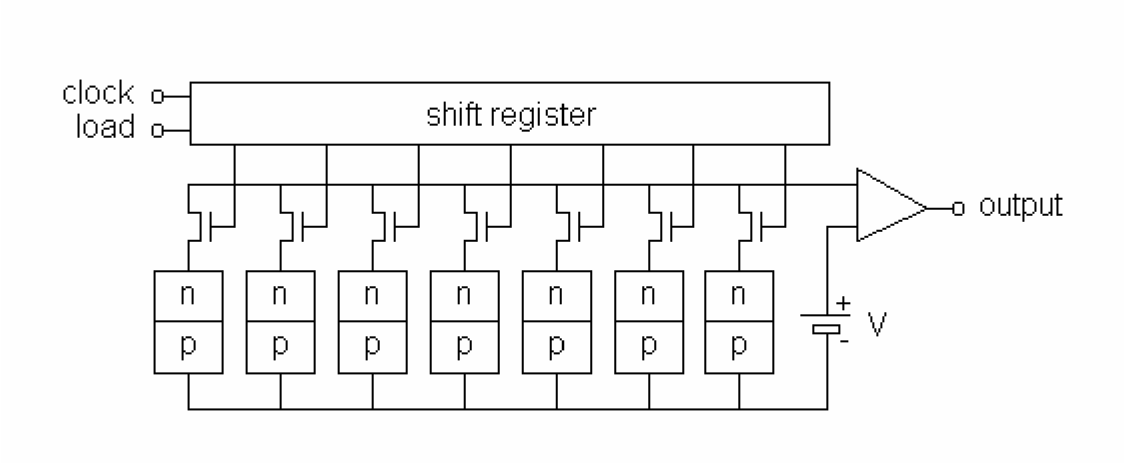


Figure 2.18. Architecture of a 1-dimensional photodiode array. A logical 1, loaded into the shift register, is shifted, activating one output after another, by pulses applied to the

clock input. The FET switch connected to the active output closes, allowing one *pn* diode after another to be reset and the accumulated charge measured.

2.3.6 CCDs

Charge-coupled detectors (CCDs) have revolutionized optical astronomy, and have proven to be very effective also for X-ray astronomy. A CCD is an array of light sensitive sites connected in such a way that charge can be transferred from one site to another. A simple 1-d CCD is illustrated in Figure 2.19. It consists of a substrate of p-type silicon upon which is built an insulated gate (electrode) structure. The simplest arrangement is a 3-phase system in which there are three gates per pixel, each connected to the corresponding gate of all the other pixels. When a positive potential is applied to one set of gates, phase 1 say, a potential well (a region of low potential energy) is created around each gate in the set, ie one per pixel. The positive potential on the gate repels the free charge carriers (holes) in the silicon, creating a depletion region below the gate. Electrons, either photoexcited or thermally excited, entering one of these regions are attracted toward the gate but cannot reach it because of the insulating layer. When exposed to light, each pixel accumulates a number of electrons proportional to the light intensity at that point. The unique aspect of the CCD is the manner in which these charges are measured.

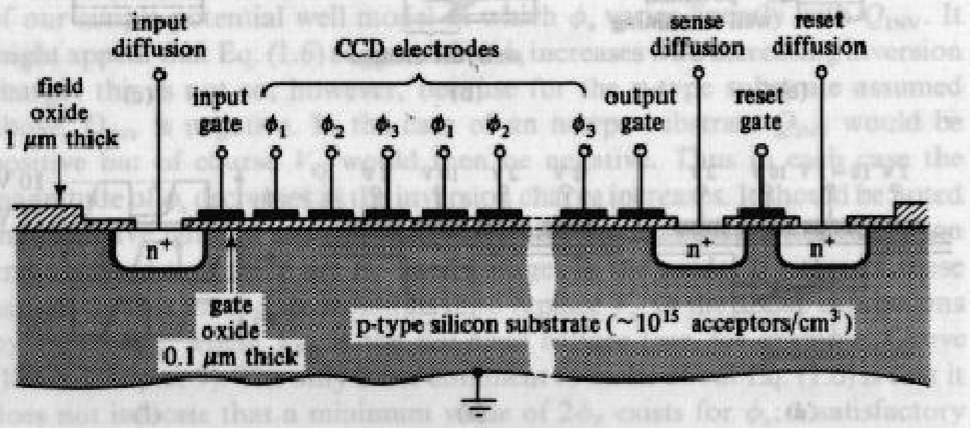


Fig. 1.8 Diagrammatic representation of a complete CCD system, including input and output circuitry (n^+ denotes heavily doped n-region).

14

Figure 2.19 Cross section of a 1-dimensional CCD. The input diffusion is normally not used for astronomical applications. The sense diffusion is connected to the output amplifier.

If the potential on phase 1 is now reduced to zero, while simultaneously, the potential on phase 2 is increased, the potential well moves to the region below the phase 2 gate, carrying with it the electrons it contains. This happens for all the pixels simultaneously. Thus, the electrons have all been moved slightly towards the end of the array. The process repeats now with the potential on phase 2 being reduced to zero while a positive potential is applied to phase 3. This moves the charges further, to the region below the phase 3 gate. Now the potential is switched again, this time from phase 3 back to phase 1. Again the electrons move, but now those in the first pixel enter the second, those in the second enter the third, and so on. The charge on the last pixel is transferred to an amplifier where it is measured. In this manner, the charges on an array of *N* pixels will all have been transferred sequentially to the output amplifier after *N* cycles of each of the three phases.

The key advantage of a CCD, compared to the diode array, is that although all of the charges are ultimately transferred to the amplifier and measured, the amplifier is connected only to a single pixel, the last one in the array. The capacitance is therefore low, and is independent of the number of pixels. As a result, the KTC noise associated with the measurement is also low, and does not increase with larger array sizes.

Two-dimensional CCDs (eg Figure 2.20) are made by arranging M one-dimensional arrays in the form of columns, as illustrated in Figure 2.21. Their gates are all connected in parallel, so that a cycle of these parallel gates moves the charges in all columns downwards together. The charges at the bottom of each column are transferred to a horizontal CCD, called the serial or horizontal register, which has its own set of gates, called the serial gates. To read out the array, the parallel gates are cycled once to move the bottom row of charge into the serial register. The serial gates are then cycled M times, to move the M charges in the serial register to the amplifier where they are measured. The parallel gates are then cycled again to load the next row of charges into the serial register and the process is repeated until all the charges have been read out.

Figure 2.20 A commercial CCD array.

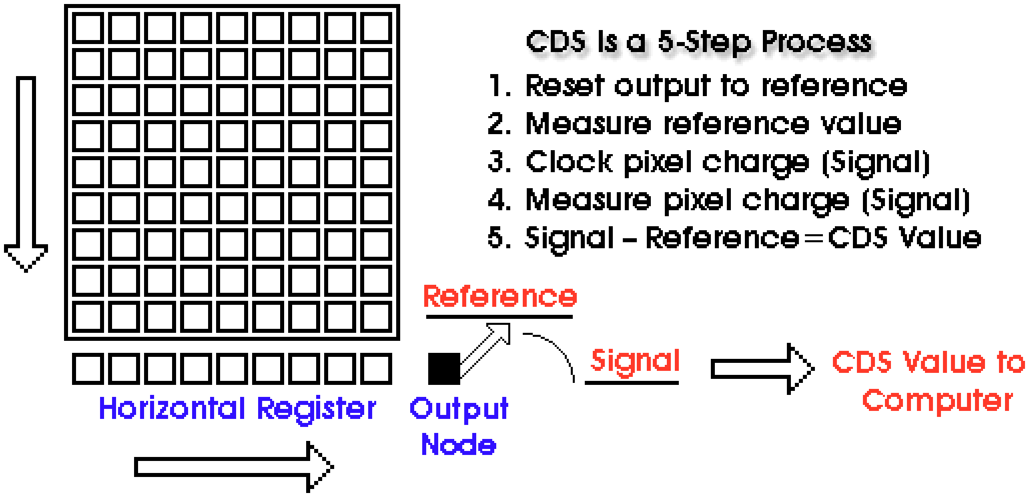

CDS Is a 5-Step Process
1. Reset output to reference
2. Measure reference value
3. Clock pixel charge (Signal)
4. Measure pixel charge (Signal)
5. Signal – Reference = CDS Value
Reference
Signal
CDS Value to Computer
Horizontal Register
Output Node

Figure 2.21 Two-dimensional CCD read out geometry.

For CCDs to work, the efficiency with which the electrons are transferred from pixel to pixel must be very high. In a 2048×2048 pixel CCD, the charge at the top of the first column must undergo 4096 transfers to reach the output amplifier. Even a small loss of charge during each shift cannot be tolerated as it would be compounded over and over. The charge transfer efficiency of modern CCDs is very good, typically 99.999% or better. Most CCDs are ***front illuminated***, which means that the light must pass through the overlying gate structure in order to reach the silicon below. The gates are made from a semitransparent conductive material (polysilicon), but inevitably some photons are absorbed or reflected in these layers. Because of this, the peak quantum efficiency of front illuminated devices rarely exceeds 40%.

This problem can be overcome by reversing the CCD and allowing light to enter from the back side (ie the silicon itself). For this to work, the silicon has to be thinned to a thickness of some tens of microns otherwise the photons are absorbed too far from the gates to be efficiently collected. Thinning is a difficult and delicate process that sometimes results in the destruction of the device. These ***back-illuminated*** CCDs often have peak quantum efficiencies over 80%, but are much more expensive and difficult to obtain than are front-illuminated devices. Figure 2.22 shows quantum efficiency curves for several front and back-illuminated CCDs.

CCDs can be used to detect energetic particles or photons such as X-rays. These particles deposit energy in the silicon creating multiple electrons. The number of electrons created by a single particle is proportional to the energy of the particle. If the flux of particles is not too high, they can be detected individually in the CCD image. In this manner one obtains not only an image in X-rays of the source, but also a measure of its energy spectrum.

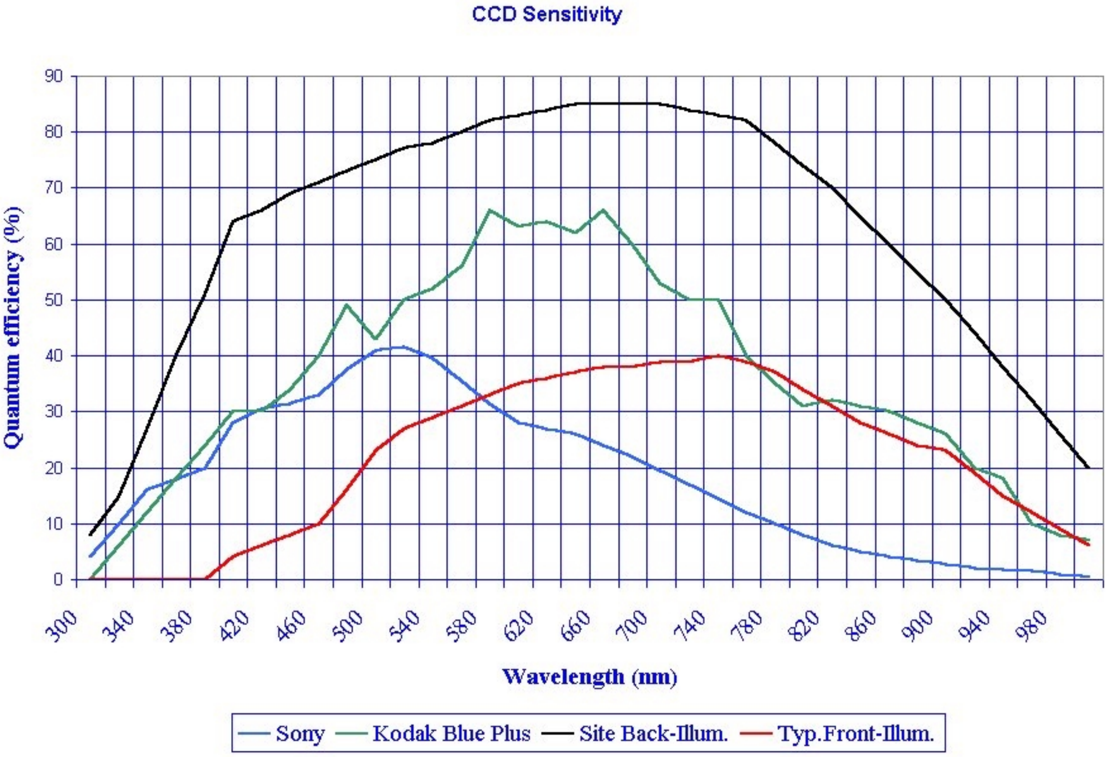


Figure 2.22 Comparative quantum efficiencies of typical front and back-illuminated CCDs.

Cosmic rays are a problematic source of background in CCDs. Each cosmic ray excites many electrons, leaving a visible track or spot in the image. The problem is less severe in back-illuminated CCDs as these are thinner and therefore have a lower cross section for cosmic ray absorption. Also, the cosmic ray images are smaller and can be more easily distinguished from stars. For observations at faint light levels, the cosmic ray background often is the main factor limiting the exposure time that can be used.

Another problem affecting CCDs is ***fringing***. Light is subject to optical interference from multiple reflections in the gate layers and the silicon itself, which produces optical fringes in the image. In imaging observations these usually arise from light from the night sky that has several strong emission lines. Fringing is most serious in thinned CCDs used for spectroscopy, where it can result in a modulation of the intensity of the image by as much as 30%.

Despite their problems, CCDs are far more sensitive than most alternative detectors for optical radiation. They can also be made in very large formats (4096×4096 and even larger), or assembled to form mosaics (Figure 2.23). The largest mosaic CCD cameras today have over a billion pixels! They are currently the dominant detector type for optical astronomy.

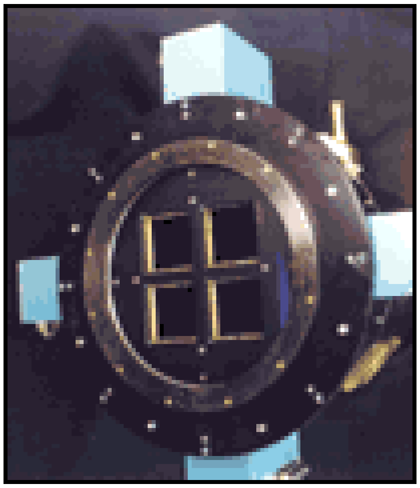


Figure 2.23 The Big Throughput Camera (BTC) – a mosaic of four 2048×2048 pixel CCDs. The CCDs are housed in a vacuum dewar and cooled with liquid-nitrogen to approximately 170 K.

2.3.7 Active-pixel sensors

Although CCDs have proven to be very effective detectors for optical and X-ray astronomy, they have certain limitations. At low temperatures charge transfer efficiency decreases because the electrons have less thermal energy and therefore less mobility. At temperatures below about 100 K they effectively stop working. This makes CCDs unsuitable for very low-temperature environments such as the Next Generation Space Telescope (NGST) which will operate at an ambient temperature of only 30 K.

Active pixel sensors avoid the problem of charge transfer efficiency by using multiplexed addressing. Unlike CCDs, active-pixel sensors are diode arrays. They differ from those already discussed in that every pixel has its own amplifier. Each amplifier sees only the capacitance of a single diode so the KTC noise is low. The amplified signals are then multiplexed to a single output, much as in the original diode array. Because the multiplexed signal has already been amplified, at each pixel, the noise contributed by the multiplexer and final amplifier is negligible. For silicon, the photodiodes, amplifiers and multiplexer can be made on a single integrated circuit. Each pixel then contains a photodiode, amplifier, and multiplexer switches and lines. Because of this, the active area (the region sensitive to light) is smaller because of the space occupied by the amplifier. These devices therefore have lower quantum efficiency, typically less than 50%. Because complimentary MOSFET technology is used for the amplifiers and switches, these devices are also called ***CMOS imagers***. Current devices are intended primarily for the television/video market and contain fewer than 10^6 pixels.

2.3.8 Hybrid arrays

Another important limitation is wavelength sensitivity. Large-format astronomical CCDs and active pixel sensors have been made only from silicon and are therefore unresponsive to wavelengths greater than approximately 1 um. A different kind of detector is needed for infrared observations. Hybrid arrays combine a silicon multiplexer and amplifier array with an array of photodiodes that can be made from a different material. They are similar in architecture to active pixel sensors, but eliminate problem of low quantum efficiency by locating the amplifier and multiplexer circuitry *underneath* the photodiode, so the latter can occupy all of the pixel area. This is done by combining two integrated circuits (ICs), made on separate wafers. (A ***wafer*** is a slice of crystalline silicon on which is manufactured several integrated circuits.) One IC, called the read-out integrated circuit (ROIC), contains the multiplexer and the amplifiers for each pixel. The other contains the photodiodes. In order to connect the two electrically, a small bead of indium, called a ***bump bond***, is attached to each photodiode (the other side of the photodiode junctions are all connected together and connected to one or more bump bonds at the side of the array. The diode array is then flipped over, aligned, and glued to the ROIC. The indium bumps contact pads on the ROIC, located at each pixel, and at the edge of the array. Heat and pressure are applied to ensure a good electrical connection.

Not only do hybrid arrays provide the best of both worlds – high quantum efficiency and low noise – but they also allow diode materials that differ from the silicon used to make the ROIC. ***Indium antimonide*** (InSb) photodiodes can be used to provide infrared sensitivity, to a cutoff wavelength of about 5 um. InSb hybrid arrays of 256×256 pixels are used for the infrared imagers on the Space Infrared Telescope Facility (SIRTF). Larger arrays are under development. ***Mercury cadmium telluride*** (HgCdTe) hybrid arrays have been fabricated in formats as large as 2048×2048 pixels. The band-gap energy of these devices depends on the relative concentrations of the component materials and can be chosen to give wavelength cutoffs in the range 2.5 – 10 um. As always, there are compromises to be made in this choice. Longer wavelength cutoffs imply smaller band-gap energies. This results in greater dark current as there are then more thermally excited electrons reaching the conduction band. At wavelengths beyond about 2.2 um, thermal black-body radiation from the instrument, telescope and atmosphere is very strong and becomes the dominant source of noise. While the detector itself can be cooled to a low temperature, this is not possible for the telescope and atmosphere.

The read noise for current hybrid arrays is in the range 15-60 electrons RMS, which is higher than the noise that can be achieved with the best CCDs. However, unlike diode arrays or CCDs, the charge stored in the pixels of hybrid arrays can be measured non-destructively. It is therefore possible to measure the charge on each pixel repeatedly and average the result to reduce the noise. Multiple reads are made before and again after integration and the difference of the mean values is used. This technique, (called ***Fowler sampling***), reduces read noise and cancels drift.

2.4 Coherent Detection

At radio frequencies, and increasingly at higher frequencies too, it is possible to detect radiation coherently. This means that the phase is preserved, at least for some stages of the process. The most common detector employed by radio astronomers is the ***heterodyne receiver***.

2.4.1 The heterodyne receiver

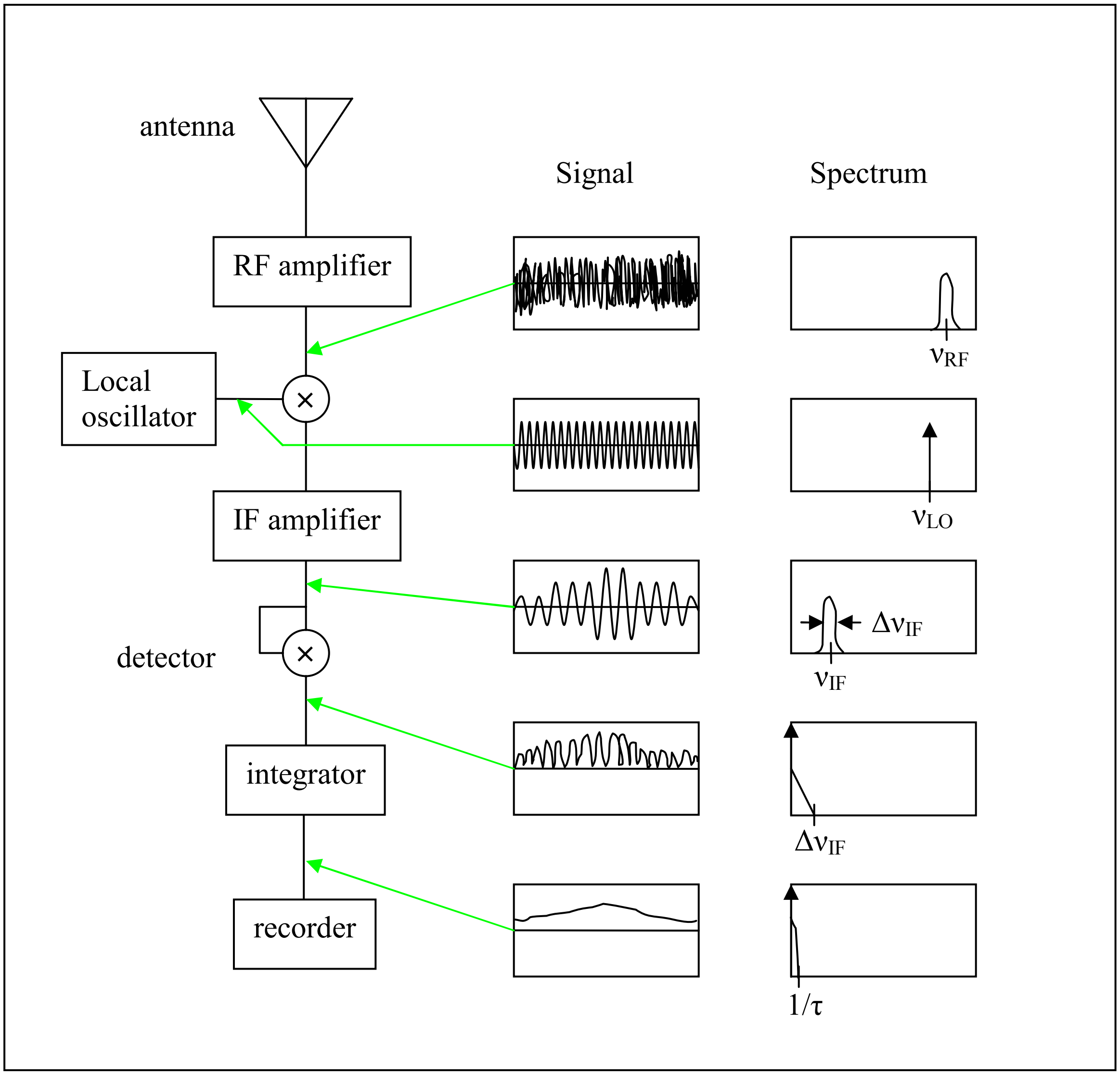


Figure 2.24. Block diagram of a heterodyne receiver. The sketches show typical electrical signals and power spectral densities at intermediate points in the receiver.

The fundamental components of a heterodyne receiver are shown in Figure 2.24; an example of a heterodyne receiver, operating at sub-millimeter wavelengths, is shown in Figure 2.25. Electromagnetic radiation induces an electrical signal in an antenna which then enters the receiver. There it is amplified by a low-noise ***radio-frequency*** (RF) amplifier. This amplifier passes frequencies in a range $\Delta\nu_{RF}$ centered at frequency ν_{RF}. This frequency is generally too high and the bandwidth $\Delta\nu_{RF}$ too large to allow a direct measurement of the signal. Normally we are interested in measuring the power as a function of frequency, or over a set of narrow frequency bands (to select particular emission lines). So, the signal must be coherently reduced in frequency.

This is done by combining the RF signal with a sinusoidal signal produced by a ***local oscillator*** (LO). The LO signal is spectrally pure (monochromatic) with frequency ν_{LO}. The RF and LO signals are fed to a nonlinear device called a ***mixer*** which produces a signal proportional to the product of these two. Since

$$\cos(2\pi\nu t)\cos(2\pi\nu_{LO}\,t) = \frac{1}{2}\{\cos[2\pi(\nu-\nu_{LO})t] + \cos[2\pi(\nu+\nu_{LO})t]\} \qquad (2.35)$$

The resulting signal contains two frequency bands. One is centered at a frequency $\nu_{IF} = \nu_{RF} - \nu_{LO}$ called the ***intermediate frequency*** (IF). The IF is normally much smaller than the RF, so the IF signal can be easily and effectively amplified and measured with conventional electronic components. This is done by the narrowly-tuned high-gain ***IF amplifier***. The range of frequencies $\Delta\nu_{IF}$ passed by the IF amplifier is normally much smaller than that of the RF amplifier. It is the IF amplifier that determines the bandwidth of the receiver $B = \Delta\nu_{IF}$.

The second frequency band, centered at $\nu_{RF} + \nu_{LO}$ is a much higher frequency (called the *image frequency*) and is blocked by the IF amplifier, falling outside its pass-band.

The signal from the IF amplifier is then sent to a nonlinear device (called a ***detector*** – not to be confused with our more-general use of the term) which produces an output signal that is proportional to the *square* of the input signal. Rather than oscillating about zero, the signal coming from the detector has no negative values. Its average RMS value is proportional to the square of the flux $F_\nu B$ being received by the antenna (integrated over a frequency band of width $B = \Delta\nu_{IF}$ centered at frequency ν_{RF}).

One can show that when one squares a random signal, the power contained in the constant (DC) component and fluctuating (AC) component of the spectrum are equal.

In order to remove the rapid IF oscillations and to reduce noise, the signal from the detector is passed through an ***integrator***. This is effectively a low-pass filter that substantially attenuates fluctuations having frequencies greater than $1/\tau$, where τ is the ***time constant*** of the integrator (for a simple RC low-pass filter, $\tau = RC$).

The result of the low pass filter is that the power in the fluctuating component is reduced. Because the effective bandwidth is reduced from $B/2$ to $1/\tau$, the AC power is reduced by a factor $2/\tau B$.

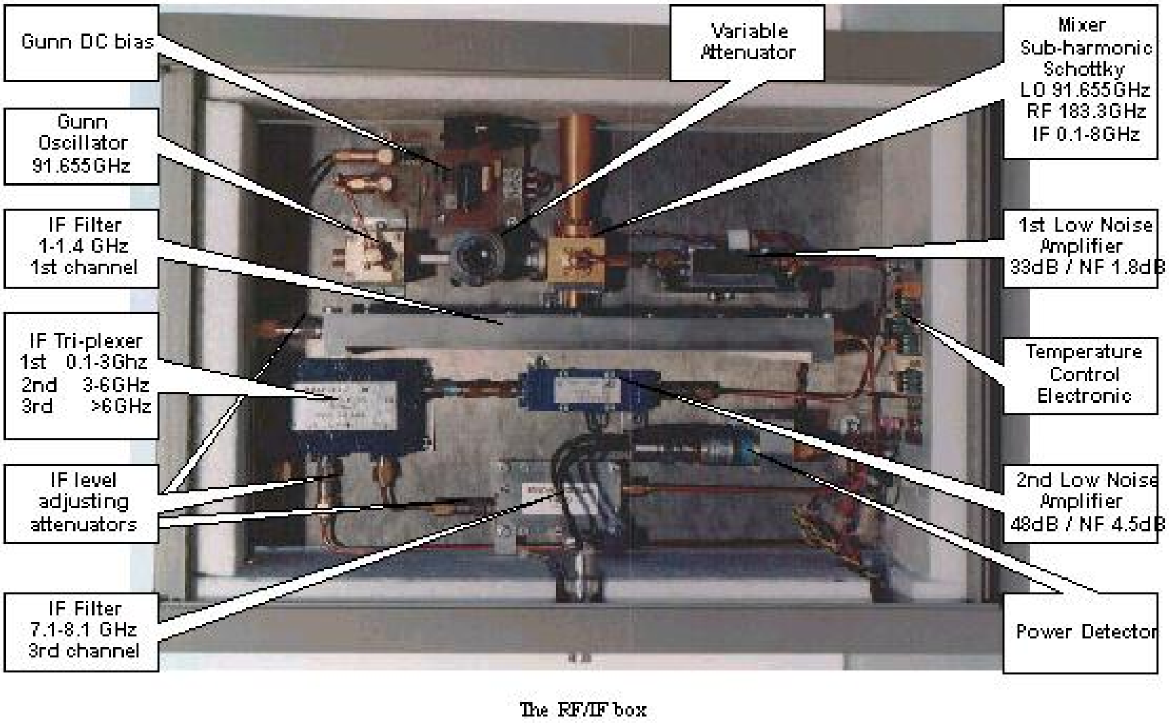

Gunn DC bias
Gunn Oscillator 91.655GHz
IF Filter 1-1.4 GHz 1st channel
IF Tri-plexer 1st 0.1-3Ghz 2nd 3-6GHz 3rd >6GHz
IF level adjusting attenuators
IF Filter 7.1-8.1 GHz 3rd channel
Variable Attenuator
Mixer Sub-harmonic Schottky LO 91.655GHz RF 183.3GHz IF 0.1-8GHz
1st Low Noise Amplifier 33dB / NF 1.8dB
Temperature Control Electronic
2nd Low Noise Amplifier 48dB / NF 4.5dB
Power Detector
The RF/IF box

Figure 2.25. Photograph of a 183 GHz sub-millimeter heterodyne receiver. Radiation enters through the copper waveguide at the top center of the box and is mixed with the LO signal produced by a Gunn diode oscillator. The IF amplifier has two amplifier stages and three simultaneous IF bands (courtesy of Chalmers University of Technology).

2.4.2 Receiver sensitivity

It is common in radio astronomy to describe power or flux levels in terms of temperature. By Nyquist's theorem, a resistor at temperature T produces kT WHz^{-1} of noise power, independent of its resistance. The ***system temperature*** is defined to be the temperature of a resistor which, if connected to the input of the receiver in place of the antenna, would produce the observed output power from the detector. The system temperature includes the effect of all sources of noise within the receiver as well as the power received by the antenna.

When the antenna is pointed at a source of radio waves, the system temperature increases. The amount of increase is directly proportional to the power received from the source. The sensitivity of the receiver is ultimately limited by the fluctuating (AC) component of the signal coming out of the detector. The RMS value of this fluctuating power can similarly be represented by a temperature ΔT, called the ***noise temperature***. The ratio of the noise temperature to the system temperature is given by the square root (because the detector squares the signal) of the ratio of the AC and DC components coming from the integrator.

$$\begin{aligned}\frac{\Delta T}{T} &= \left[\frac{\text{AC power}}{\text{DC power}}\right]^{1/2} \\ &= \left[\frac{(\text{DC power})(2/\tau B)}{\text{DC power}}\right]^{1/2} \\ &= \sqrt{2/\tau B}\end{aligned} \tag{2.36}$$

where we have used the fact that the AC and DC power are equal at the output of the detector, and that the effect of the integrator is to reduce the AC component by a factor $2/B\tau$.

From this we see that the faintest power level that can be detected (at the 1σ level, ie. a signal-to-noise ratio of 1) is

$$\Delta T = \sqrt{\frac{2}{B\tau}}T \tag{2.37}$$

Where T is the system temperature. We will use this formula again, in the next chapter, to derived an expression for the *flux limit* of a radio telescope.

Exercises

2.1 Prove Equation (2.9).

2.2 A source is observed in the infrared as follows: First, a small area of sky is observed which contains the object of interest. Then, an equal area of sky, in nearly the same direction but not containing the object (or any other), is observed for an equal length of time. The signals from these two observations are then subtracted to determine the flux from the source alone. If n_1 photons are detected in the first field and n_2 in the second, estimate the signal-to-noise ratio that will be obtained for the flux.

2.3 Suppose that during an observation, n photons are incident on a detector that has quantum efficiency Q. Suppose also that the detector adds noise which has an RMS value equal to the signal produced by r photons. Derive an expression for the DQE D of this observation, in terms of the above quantities. Show that in the limit when the detector noise can be neglected that $D = Q$.

2.4 When photomultipliers are used to count photons, there is a brief period of time, immediately after a photon is detected, when the photomultiplier, amplifier, counter, etc, is unable to respond to any other photons. This is called the ***dead time*** of the system. Therefore, the measured count rate will be smaller than the true rate (the rate that would be measured if there was no dead time), because some photons will arrive during the dead time and will be missed. Because the dead time is usually a constant, independent of the count rate, a larger fraction of photons will be missed if the count rate is high. Derive an expression relating the observed count rate R , the true rate R_0 , and the dead time t_d

3 Telescopes

In order to measure flux or intensity, one needs not only to detect the radiation, but also to control and measure the direction and solid angle. A ***collimator*** is a device that restricts and defines the solid angle of the radiation reaching the detector. For example, a lead plate with a hole drilled through it is an effective way of collimating certain kinds of radiation. More sophisticated instruments employ telescopes. A telescope generally has several functions:

- It defines the direction and solid angle of the radiation reaching the detector
- It increases the effective area of radiation that is intercepted and measured
- It increases the solid angle of the radiation reaching the detector

The last of these functions imply that the source has been magnified by the focusing effect of the telescope, because the rays coming from it now subtend a larger angle on the sky.

Telescopes are used for radiation ranging all the way from the radio to X-ray regions of the spectra. There is a great variety of telescope types, design and construction for these different wavelength regions, but they share the same fundamental properties. Outside this region, there are fundamental differences in detection techniques. Low frequency radio waves are detected by antennas, which may be no more than wires strung on poles. Very high-energy radiation such as gamma rays, or cosmic ray particles, are more effectively detected by particle detectors that respond to ionization, or indirectly by detecting photons produced in the interaction of these particles with the Earth's atmosphere. This chapter emphasizes telescopes used for optical and infrared radiation, however the fundamentals are the same also for radio and X-ray telescopes.

3.1 Basics

An astronomical telescope forms an image of a distant object by focusing parallel rays to a point, called the focus. An example of this is shown in Figure 3.1. In this case light is focused by a lens (called the ***objective lens***). In a perfect telescope, other rays that are also parallel to each other, but coming from a different direction are focused to a different point. The set of all such points forms a plane called the ***focal plane***. In this way, the telescope effectively maps direction (angle) in the sky to position in the focal plane. In a perfect telescope, this mapping is linear. In practice, there is usually some nonlinearity (called ***distortion***) that increases with distance from the center of the focal plane, but in the central region the linear assumption is a good one. Thus, we may write

$$\mathbf{x} = f\boldsymbol{\alpha} \tag{3.1}$$

Where $\mathbf{x} = (x, y)$ are Cartesian coordinates in the focal plane, $\boldsymbol{\alpha} = (\alpha, \beta)$ are angular coordinates on the celestial sphere, centered on the direction that corresponds to the point

$\mathbf{x}=0$, and f is a constant of proportionality. Since $\mathbf{x}$ has the dimension of distance and $\boldsymbol{\alpha}$ is a dimensionless angle (measured in radians), f has units of distance. For the simple telescope shown in Figure 3.1, it is evident that f is just the distance from the center of the lens to the focal plane, called the ***focal length*** of the lens.

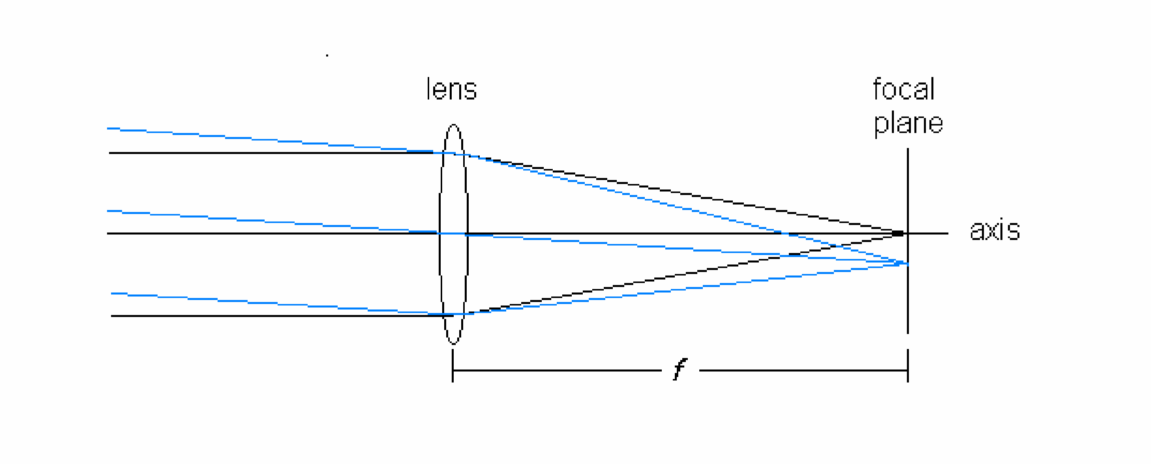


Figure 3.1. Image formation by a lens. Parallel rays are focused to a point in the focal plane. The lens maps the angle of the incident rays to position in the focal plane.

The lens focuses light by refraction. When light propagates through a medium of refractive index n, the speed of propagation is

$$v = c/n\,. \tag{3.2}$$

The wavefront is therefore retarded by an amount that is proportional to the thickness of the lens. To first order, the lens changes a plane wavefront into a spherical wavefront of radius *f* which converges to the focal point. Equivalently, once can consider ***rays***, which are families of lines that are everywhere perpendicular to the wavefront. The refraction of rays at an interface between two media is given by ***Snell's law***

$$n\sin\theta = n'\sin\theta' \tag{3.3}$$

where θ and θ' are the angles if incidence and refraction that the incident and refracted rays make with a line normal to the surface, and n and n' are the indices of refraction on the corresponding sides of the surface.

The case of ***reflection*** from a mirror is particularly simple since the angle of incidence equals the angle of reflection.

By ***Fermat's principle***, the time taken for light to propagate from any point on the incident wavefront to the focus is the same for all rays (in a perfect telescope). This is equivalent to the statement that the ***optical path length*** (OPL) (or the optical path difference OPD) from any point on the wavefront to the focus is constant, where the OPL is defined to be the physical distance multiplied by the index of refraction.

Most optical systems, such as the lens shown in Figure 3.1, have an axis of rotational symmetry which passes through the center of the focal plane. Properties of the optical system, such as the location of images, magnification, etc, can be simply found by considering only rays that are arbitrarily close to the axis. For such ***paraxial rays***, the angles of incidence and refraction are small and the sine functions that occur in (3.3) can be replaced by their linear approximation (*ie* just the angles themselves). The use of the linearized Snell's law is called the ***Gaussian-optics approximation*** or ***first-order theory***. The next terms in the series expansion of $\sin\theta$ are third order in θ, so the first correction to first-order theory is called ***third-order*** theory. As we shall see, third-order and higher terms correspond to optical ***aberrations*** which cause a blurring or distortion of the image.

The image formed at the focal plane of the lens in Figure 3.1 is called a ***real image***, because the light rays actually pass thought it. If a second convex lens is now added to the system, after the focus, one has a telescope that can be used for visual observations, as shown in Figure 3.2. Essentially, the second lens (called the ***eyepiece***) acts as a magnifier that allows the eye to view the real image produced by the objective lens. Because of the focusing effect of the eyepiece, the rays appear to originate from a magnified ***virtual image***, as shown in the figure. Real eyepieces actually contain several lenses, designed to minimize aberrations. The ratio of the angular size of the virtual image to the angular size of the source is called the ***magnification*** of the telescope. It can easily be shown that the magnification is equal to the ratio of the focal length of the objective lens to the focal length of the eyepiece (the distance at which it would focus parallel light).

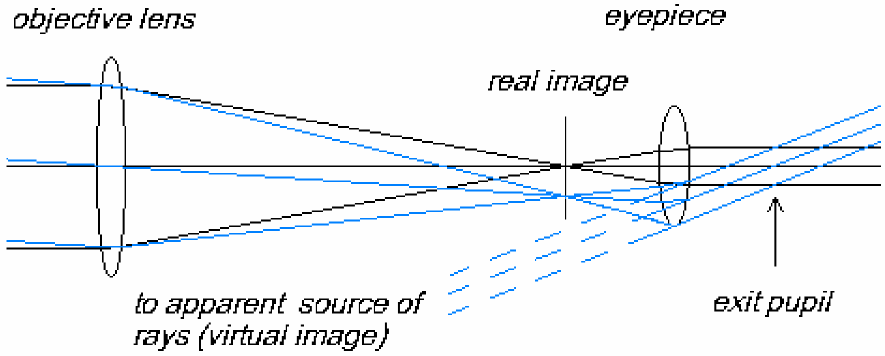


Figure 3.2. A simple telescope incorporating an eyepiece for visual observations. The eyepiece magnifies the real image, creating a virtual image of the source at infinity. The image is inverted.

The focal length of a complex system of lenses is found by the construction shown in Figure 3.3. Extend a paraxial ray, entering the lens parallel to the axis, into the lens without deflection. Similarly, extend the corresponding ray converging to the focus, backward into the lens until it intersects the extended paraxial ray. The point of intersection defines a plane perpendicular to the axis called the ***principal plane***. The distance from the principal plane to the focal plane is the focal length. Since parallel light can in principle enter the lens from behind and be focused in front, there is a second principal plane defined for such rays. Every lens thus has two principal planes. These generally do not coincide. The focal length of a complex optical system used as a telescope is also called the ***effective focal length*** (EFL), to distinguish it from the focal lengths of individual components.

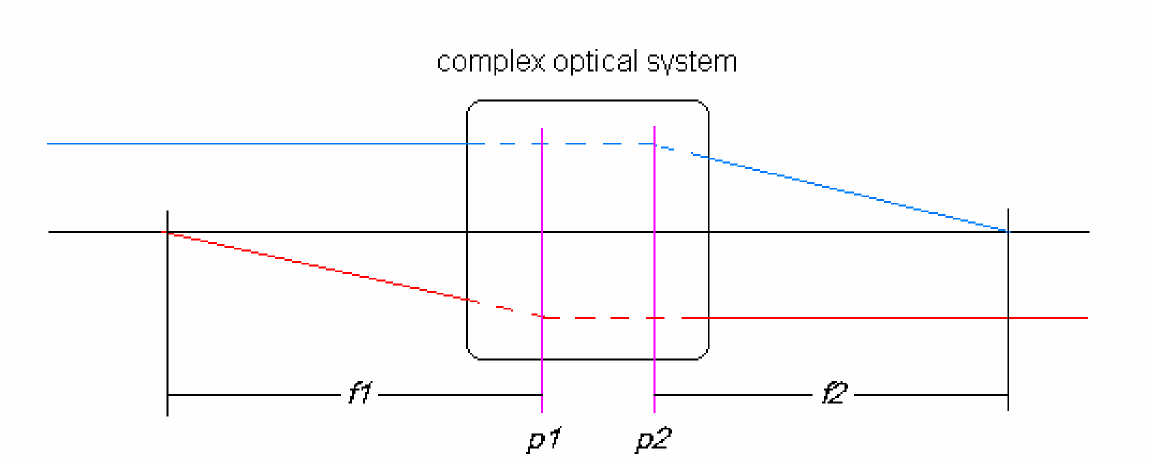


Figure 3.3. Principal planes (p1 and p2) and focal lengths of a complex optical system.

An optical system generally has some sort of aperture that defines which rays are able to pass through the system. The optical system may also produce images of this aperture. The aperture and its images are called ***pupils***. The first pupil encountered by the rays (usually the objective lens or primary mirror) is called the ***entrance pupil*** and the last pupil encountered before the rays exit the system is called the ***exit pupil***.

3.2 Optical invariants

We have already encountered one invariant quantity in geometrical optics, the ***Snell invariant*** $n\theta$, where θ is measured with respect to the normal to a surface, which is invariant for paraxial rays. We can use the conservation of intensity to obtain a second invariant.

Consider light entering an optical system. Let the area of the entrance pupil be A, and the solid angle subtended by two rays entering the system be Ω (usually the solid angle encompassed by the field of view). Now, if Ω is small (so that the cosine factor in (1.27) can be ignored) the product $I_\nu A\Omega$ is the radiant power per unit frequency interval entering the system. If there are no absorption or reflection losses, the same power must pass through every pupil and exit the system. Also, if there are no losses $I_\nu / n^2\nu^3$ is conserved (the factor of n^2 accounts for the reduced propagation speed and momentum of light in a medium of refractive index n). The frequency of the radiation is not changed by lenses, mirrors, etc, so it follows that the quantity $n^2 A\Omega$ is conserved. Now both A and Ω are purely geometric quantities, and n is an intrinsic property of the medium, so this must also be true even in the presence of absorption or reflections losses. The quantity $n^2 A\Omega$ is related to the ***Lagrange invariant*** H of geometrical optics,

$$H \propto n\left(A\Omega\right)^{1/2} \tag{3.4}$$

It is not restricted to pupils, but applies throughout the optical system to any area A that is perpendicular to the light cone described by Ω.

We can use this result to obtain a relation between the diameters of the entrance and exit pupils and the magnification in the telescope of Figure 3.2. Let primed quantities denote the exit pupil and unprimed the entrance pupil. Let D and D' be the pupil diameters. Now the angular size of the source, and its virtual image, are related by the ratio of solid angles for the rays, thus

$$\begin{aligned} M &= \left(\Omega'/\Omega\right)^{1/2} \\ &= \left(A/A'\right)^{1/2} \\ &= D/D' \end{aligned} \tag{3.5}$$

since $n^2 A\Omega$ is conserved, and n is the same for the input and output rays. This sets a lower limit on the magnification of telescopes used for visual observations. In order for the rays exiting the telescope eyepiece to enter the eye, the diameter D' of the exit pupil must be less than the diameter of the pupil of the eye. This is about 7 mm for a dark-adapted human eye. Thus the magification must be greater than $D/7\,\mathrm{mm}$, otherwise light will be lost.

We can now prove two useful and very general magnification formula. Consider a rotationally-symmetric optical system that receives rays emitted from an source at point O, and focuses them to an image at point O', as shown in Figure 3.4. Let θ be the angle, made with the axis, by a ray and θ' be the angle of the same ray as it approaches the image. If the source is displaced by an infinitesimal distance dx perpendicular to the optical axis, the image will be displaced by a distance dx'. The ratio $M_T = dx'/dx$ is called the ***transverse magnification***. Since the system has rotational symmetry, the same magnification applies to displacements in the y direction, $M_T = dy'/dy$. Therefore, the infinitesimal area $dA = dxdy$ is mapped to the area $dA' = dx'dy' = M_T^2 dA$. By rotational symmetry, all rays leaving the source within a cone of angle θ will arrive at the image within a cone of angle θ'. Then, since $n^2 A\Omega$ is conserved, we have

$$\begin{aligned} M_T^2 &= \frac{dA'}{dA} = \frac{n^2\Omega}{n'^2\Omega'} = \frac{n^2\theta^2}{n'^2\theta'^2} \\ M_T &= \frac{n\theta}{n'\theta'} \end{aligned} \tag{3.6}$$

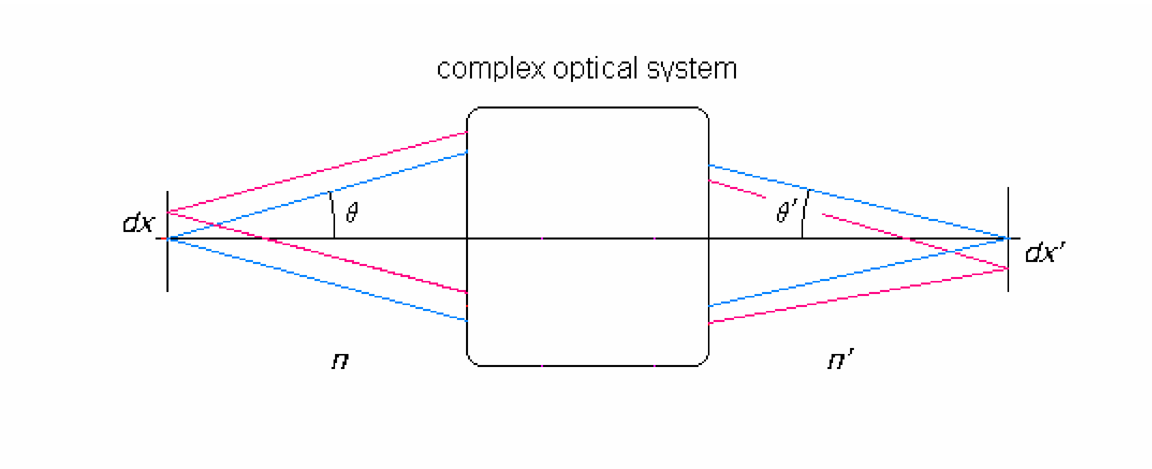


Figure 3.4. Transverse magnification. The ratio *dx'/dx* depends only on the indicated angles and indexes of refraction.

Now suppose that the source is moved an infinitesimal distance dz along the axis and the image moves a distance dz', as in Figure 3.5 The ratio $M_L = dz'/dz$ is called the ***longitudinal magnification***. The area $dA = dydz$ is mapped to $dA' = dy'dz' = M_T M_L dA$. Now, the radiant power passing through the area dA in the direction θ within a solid angle $d\Omega$ is $I_\nu \sin\theta dA d\Omega \approx I_\nu \theta dA d\Omega$, which must equal the corresponding quantity (with primes) for the area dA'. Thus,

$$\begin{aligned} n^2\theta dA d\Omega &= n'^2\theta' dA' d\Omega' \\ &= n'^2\theta' M_T M_L dA d\Omega', \\ M_T M_L &= \frac{n^2\theta^3}{n'^2\theta'^3} \\ M_L &= \frac{n\theta^2}{n'\theta'^2}. \end{aligned} \tag{3.7}$$

In summary, we have the two magnification formulae

$$\begin{aligned} M_T &= \frac{n\theta}{n'\theta'} \\ M_L &= \frac{n\theta^2}{n'\theta'^2} \end{aligned} \tag{3.8}$$

valid for small angles θ and θ' measured from the optical axis.

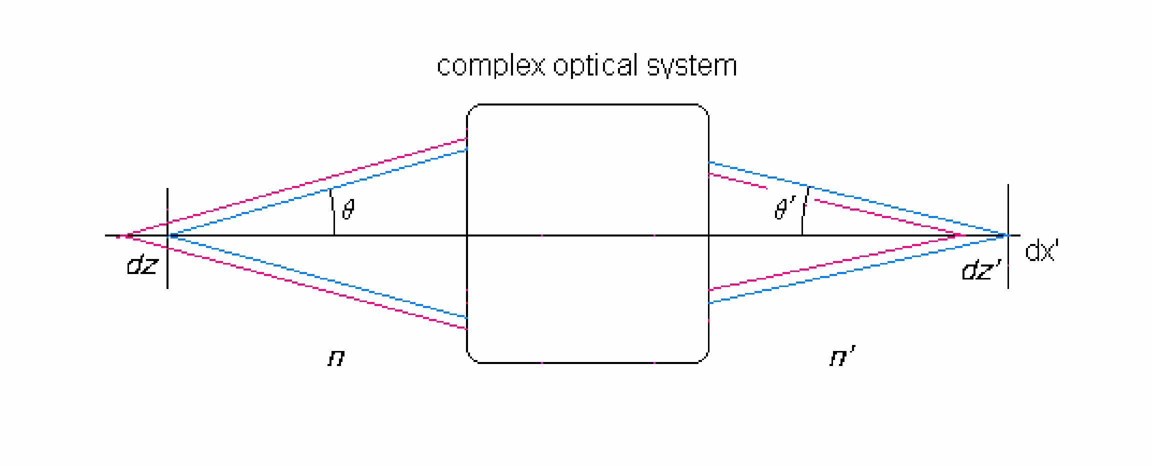


Figure 3.5. Longitudinal magnification. The ratio *dz'/dz* depends only on the indicated angles and indexes of refraction.

3.3 Optical aberrations

Gaussian optical theory tells us about the location of images and pupils, and the magnification or image scale. However, it does not describe the sharpness or quality of the images. For this we must consider higher-order theory. The main aberrations (effects which degrade the image quality) can be found from third-order theory. In practical cases it may also be necessary to include fifth or higher orders, or trace exactly the path of rays through the system using a computer program.

3.3.1 Optical shapes

The focusing effect of optical components depends on the shapes of their surfaces. These can be spherical, conic, or aspheric. Surfaces that have rotational symmetry about an axis are most conveniently described in a cylindrical coordinate system (r,ϕ,z) with the z axis coincident with the symmetry axis. A surface may then be described by giving the position *z* of a point on the surface as a function of the radial distance r from the axis. This can be written as a Taylor series as follows (because of the rotational symmetry, only even powers of *r* can appear):

$$z = z_0 + \frac{r^2}{2R} + (1+b)\frac{r^4}{8R^3} + (1+b)^2\frac{r^6}{16R^5} + A\left(r^6\right) + \cdots \qquad (3.9)$$

Here *R* is the radius of curvature of the surface, z_0 is the position of the vertex, and *b* is called the ***conic constant***. Different ranges of *b* describe conic surfaces, as indicated in Table 3.1. $A\left(r^6\right)$ denotes higher-order terms which, if present, indicate that the surface is ***aspheric*** rather than spherical or conic.

Table 3.1 Conic Surfaces

Range	Surface
$b < -1$	hyperboloid
$b = -1$	paraboloid
$-1 < b < 0$	prolate ellipsoid
$b = 0$	sphere
$b > 0$	oblate ellipsoid

3.3.2 Third-order theory

An elegant and general treatment of third-order aberrations of symmetric optical systems was published by Hamilton in 1833. Consider a plane wavefront entering a telescope. After passing through the optics, it appears at the exit pupil as a spherical wave converging to the focus. If the telescope is perfect, the wavefront will be exactly spherical and will propagate to a point (ignoring diffraction effects). However, if aberrations are present, the wavefront will deviate from a sphere. The optical path length error of any point on the wavefront, compared to the perfect sphere, is called the ***aberration function*** *W*. It can be written in terms of three fundamental parameters that can be taken to be the field angle, measured between the axis and the direction of the radiation entering the telescope, normalized to 1 at the edge of the field (σ), radial position of the point on the wavefront, normalized to 1 at the edge of the pupil (ρ), and the azimuth angle of the point measured from the plane containing the axis and the incident direction (ϕ).

Hamilton proved that because of the symmetry of the system, the aberration function must have the form

$$\begin{aligned} W(\sigma,\rho,\phi) = {}& k_{20}^{0}\rho^{2} + k_{11}^{1}\sigma\rho\cos\phi + k_{40}^{0}\rho^{4} + k_{31}^{1}\sigma\rho^{3}\cos\phi \\ & + k_{22}^{2}\sigma^{2}\rho^{2}\cos^{2}\phi + k_{20}^{2}\sigma^{2}\rho^{2} + k_{11}^{3}\sigma^{3}\rho\cos\phi + \cdots \end{aligned} \tag{3.10}$$

We see that there are two first order terms (terms quadratic in σ and ρ) and five third order terms (terms of fourth-order in σ and ρ). Their effects are summarized in Table 3.2. The first order terms represent errors in *position* of the image (ie Gaussian-optic errors). For example, a term proportional to ρ^2 means that the wavefront has a different curvature than the reference sphere. It will therefore converge to a point in front of, or behind, the nominal focal plane. This term represents a longitudinal focus error, and can be removed by properly focusing the telescope. Similarly, the term proportional to $\sigma\rho\cos\phi$ represents a tilt of the wavefront, compared to the reference sphere. This means that the images will all be displaced transversely in the focal plane. This is referred to as a guiding, or tip-tilt error, and can in principle be removed by properly guiding the telescope. In contrast, the third-order terms represent fundamental aberrations that cannot be removed by manipulating the telescope. The five third-order types are called the Seidel aberrations, after Philipp Ludwig von Seidel who first studied them.

Table 3.2 The Seidel aberrations

Term	Aberration
$k_{40}^{0}\rho^{4}$	spherical aberration
$k_{31}^{1}\sigma\rho^{3}\cos\phi$	coma
$k_{22}^{2}\sigma^{2}\rho^{2}\cos^{2}\phi$	astigmatism
$k_{20}^{2}\sigma^{2}\rho^{2}$	field curvature
$k_{11}^{3}\sigma^{3}\rho\cos\phi$	distortion

It is the task of the optical designer to minimize or eliminate as many of these aberrations as possible, while using the minimum number of optical elements or surfaces. As a rule, it is theoretically possible to eliminate one aberration for each surface in the system, although in practice more surfaces may be needed, particularly if refracting elements are involved. The index of refraction of optical glasses varies with wavelength, a phenomenon known as ***dispersion***. Because of this, the imaging properties of a refractive system will vary with wavelength, which leads to ***chromatic aberration*** – a situation where different wavelengths converge to different positions. The designer must choose optical glasses and optimize the design in order to minimize this effect.

Let us now examine the Siedel aberrations individually. For an interactive display of wavefront plots and maps for these aberrations, please look at http://wyant.opt-sci.arizona.edu/seidelWavefrontMaps/seidel.htm.

3.3.3 Spherical aberration

This aberration is independent of field angle (σ), which means that the images have the same form all over the image. It corresponds to a focus error that increases quadratically with radius in the pupil. In other words, rays passing close to the center of the pupil focus at one point, but rays at different radial distances focus at different longitudinal points. There is no focus position that results in a sharp image. This aberration is most severe in fast optical systems (ie those with large numerical aperture (small f-number) due to the strong dependence on ρ. This aberration crippled the Hubble Space Telescope until it was corrected by the insertion of additional specially-designed optics. It resulted from an error in the positioning of lenses that were used to test the shape of the primary mirror when it was being made. To illustrate the effect of a small amount of spherical aberration, Figure 3.6 and Figure 3.7 show an unaberrated wavefront and ***point spread function*** (PSF, the image of a point source), including diffraction effects, and the same quantities with 2 waves of spherical aberration (a maximum OPD error of two wavelengths).

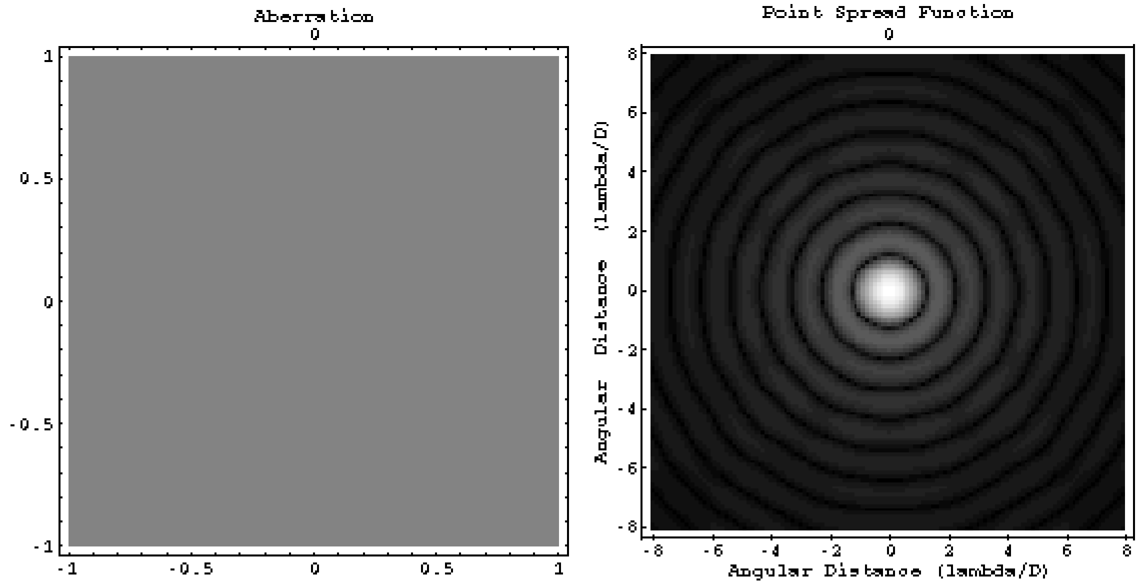


Figure 3.6 Aberration function and Point Spread Function for a perfect telescope. (Courtesy of James C. Wyant).

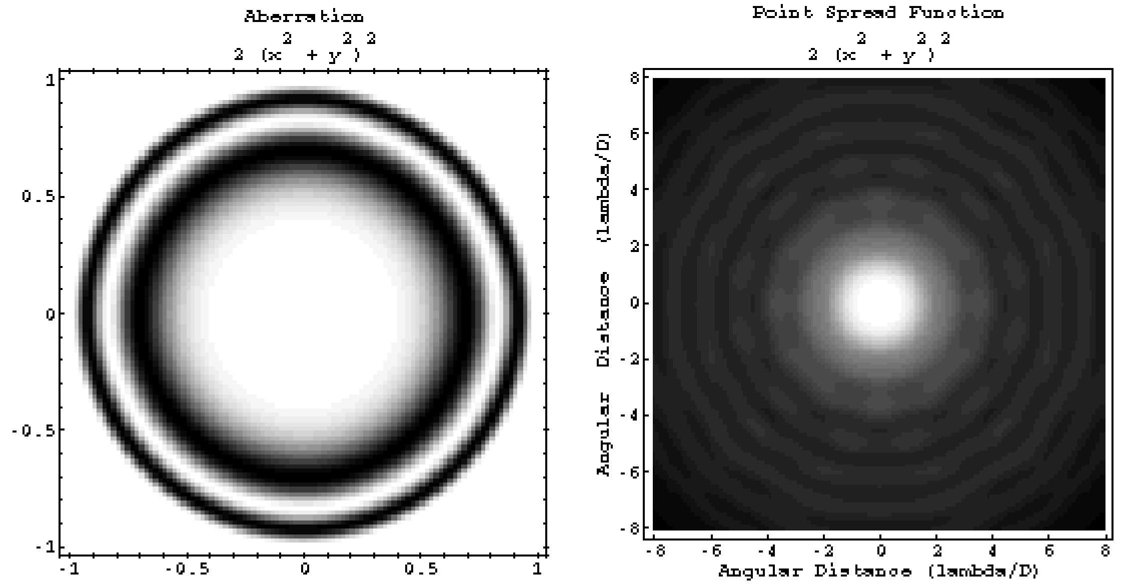


Figure 3.7 Aberration function and Point Spread Function exhibiting 2 waves of spherical aberration. The interval between successive dark bands in the aberration function corresponds to an OPD of 1 wavelength (Courtesy of James C. Wyant).

3.3.4 Coma

This is the most serious aberration after spherical aberration. Because of the linear dependence on field angle, the images at the center of the field are fine, but grow progressively worse with distance from the center. The aberration corresponds to a tilt and focus shift. Rays from at increasing radial positions in the pupil are increasingly shifted transversely away from the field center and defocused. The resulting images have the appearance of small comets pointing toward the field center, hence the name coma. Like spherical aberration, this aberration is most serious in fast optical systems. Figure 3.8 shows the aberration function and PSF of a system exhibiting 2 waves of coma. Coma and spherical aberration are so important that virtually all astronomical telescopes are designed to correct them.

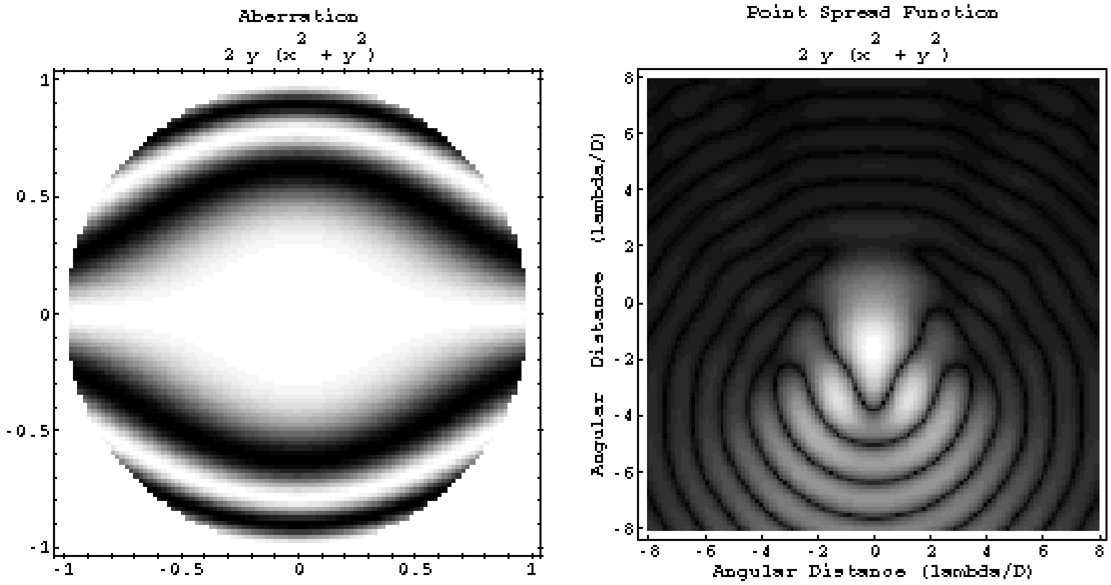


Figure 3.8. Aberration function and PSF exhibiting 2 waves of coma (Courtesy of James C. Wyant).

3.3.5 Astigmatism

This term corresponds to a longitudinal focus shift that depends on the azimuthal position of the rays in the pupil. Rays in the plane defined by the field angle focus at one distance while rays in the perpendicular plane focus at another. All other rays focus at intermediate distances. The image shape ranges from circular to linear, depending on the choice of focus. This aberration increases quadratically with field angle. Near the center, it is generally less important than coma, but grows faster with increasing angle. In most large astronomical telescopes, astigmatism (and often coma also) is removed using a set of lenses, called an optical ***corrector***, placed in the converging beam.

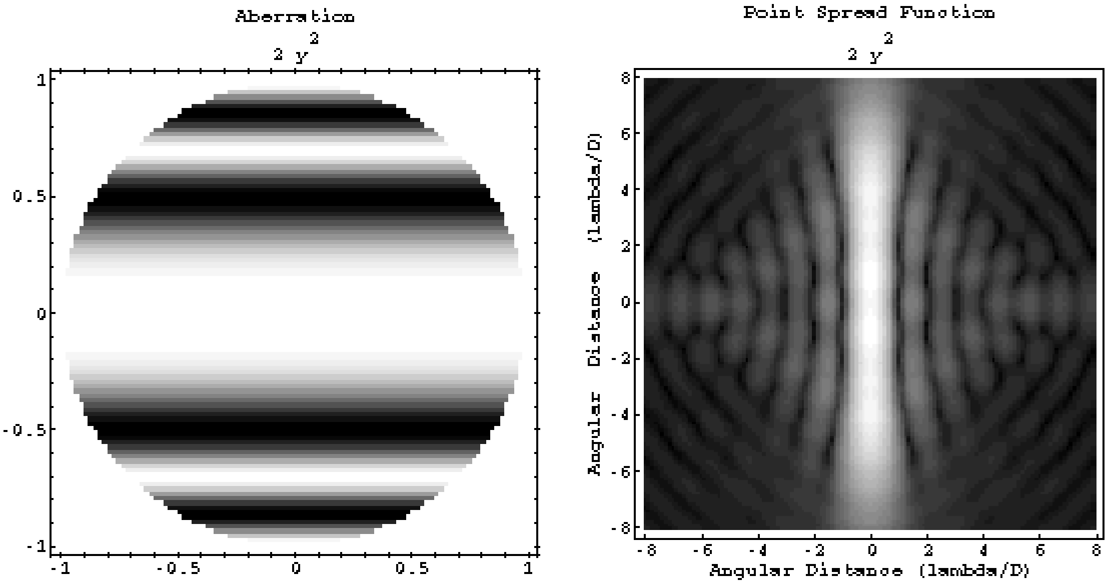


Figure 3.9. Aberration function and PSF exhibiting 2 waves of astigmatism (Courtesy of James C. Wyant).

3.3.6 Field curvature

This aberration differs from the previous three in that it is possible to find a focal surface where the image is sharp. However, the presence of this term indicates that this surface is not a plane but is curved (spherical, to first order). The radius of curvature R_P of the focal surface is related in a simple way to the radii of curvature of the optical surfaces in the system, and is independent of their position and conic shapes. It is given by the ***Petzval Sum***:

$$\frac{1}{R_P} = -n_N \sum_{i=1}^{N} \frac{\Delta n_i^{-1}}{R_i}, \quad \Delta n_i^{-1} \equiv \frac{1}{n_i} - \frac{1}{n_{i-1}} \tag{3.11}$$

where R_i is the radius of the i-th surface and n_i and n_{i-1} refer to the index of refraction that the ray is in after and before passing through the i-th surface, respectively. For reflecting surfaces, the same equation holds if we take $\Delta n_i^{-1} = -2$. This shows that it is possible to eliminate field curvature by an appropriate choice of curvatures for the individual optical surfaces in the telescope. For an all-reflecting telescope one requires that the sum of the curvatures (ie the reciprocal radii) be zero.

It is also possible to correct field curvature in many cases by inserting a lens just in front of the focal plane. This lens has a curved surface, with radius and index of refraction chosen to bring the Petzval sum to zero. Because the lens is placed very close to the focal plane it has very little effect on other aberrations. When a surface is close to the focus, the beam leading to any focal point intercepts just a small area of the surface, over which the OPD error is negligible. However, the focal point is displaced backward by a distance that is proportional to the thickness of the glass, which explains how the lens can remove field curvature. Such a lens is called a ***field flattener***.

3.3.7 Distortion

The final aberration is again one that does not affect the quality of images, but instead affects their positions in the field. This term corresponds to a transverse shift in focus position in the radial direction with a cubic dependence on field angle. This is equivalent to a magnification or image scale that varies with distance from the field center. If the magnification increases with field angle, we have ***positive*** or ***pincushion distortion***; if it decreases we have ***negative*** or ***barrel distortion***, as shown in Figure 3.10. Distortion is the least serious of the Seidel aberrations, and is often ignored. It results in the measured positions of star images being incorrect, but this is easily rectified by a suitable coordinate transformation.

There is one exception where distortion is important. Telescopes used for scanning applications, such as CCD drift scanning, require that the images move in a linear manner in order to prevent image smearing in the CCD. In this case, the telescope optics must be free of distortion.

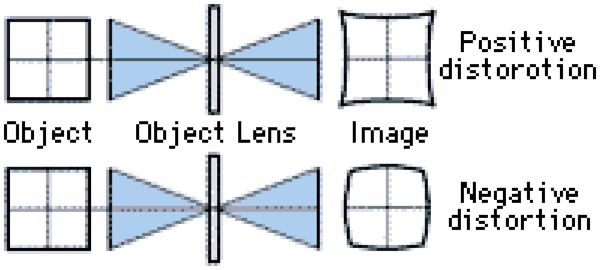


Figure 3.10. Illustration of distortion.

3.4 Telescope types

All major astronomical telescopes are of the reflecting type. That is to say, the primary optical element that collects and focuses the light is a mirror. Mirrors have a number of advantages over lenses. They can be made in large sizes, they can be supported against gravitational flexure, they require figuring of only a single surface, and they produce no chromatic aberration. Lenses are still used, in correctors and for small or auxiliary telescopes.

3.4.1 Single-mirror telescopes

The simplest telescopes employ a single ***primary mirror*** to produce an image on a detector located at what is called the ***prime focus***. By making the primary mirror parabolic ($b = -1$), spherical aberration is eliminated. Coma and astigmatism remain (and also field curvature and distortion), so large telescopes employ a ***prime focus corrector*** which consists of a set of lenses located a short distance in front of the prime focus. Most telescopes use a three-lens design, of a type first suggested by Wynne. Telescopes used for drift-scan observations, such as mercury mirror telescopes, use a four-element design to also remove distortion.

In the prime focus type, the detector and corrector block some of the incident radiation, casting a shadow on the primary mirror. For large telescopes this is not usually a serious problem as it results in only a few percent of the light being lost. In small telescopes, a flat diagonal mirror can be used to deflect the light to a focus just outside the incident light path. This technique was fist used by Newton, so such a telescope is called ***Newtonian***.

3.4.2 Two-mirror telescopes

Two-mirror telescopes are normally of the ***Cassegrain*** or ***Gregorian*** type, illustrated in Figure 3.11. In the former, a convex ***secondary mirror*** is place before the prime focus which directs the light through a hole in the center of the primary mirror to a focus (the Cassegrain focus) not far behind the primary mirror. In the latter, the secondary mirror is concave and located after the prime focus. Both systems have the advantage of providing easy accessibility to the focus – one can mount small instruments, such as a spectrograph, directly on the back of the telescope structure behind the primary mirror. Because the secondary mirror has curvature, it provides magnification factor (an increase in the EFL) typically in the range 3-12. This makes it possible to achieve focal lengths that are much longer than the physical size of the telescope – a great advantage in cost and convenience.

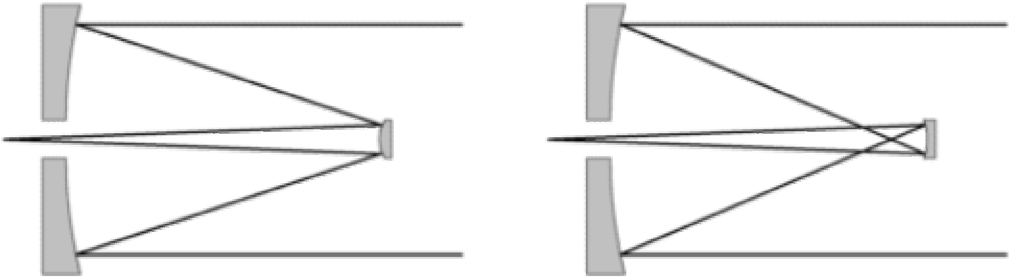


Figure 3.11. Comparison of Cassegrain (left) and Gregorian (right) telescopes.

In both the Cassegrain and Gregorian types, the primary mirror is parabolic. The secondary mirror simply re-images the light converging to, or diverging from, the prime focus and directs it to the Cassegrain or Gregorian focus. In order to do this without introducing spherical aberration, the Cassegrain secondary mirror must have a hyperbolic shape, while the Gregorian secondary is elliptical. Cassegrain telescopes are more common than Gregorian, in part because they are more compact. The relatively large focal lengths of these system (compared to the prime focus) means that they have smaller numerical apertures (larger f numbers). The off-axis aberrations (coma and astigmatism) are therefore less severe and can be removed by a simpler corrector. Often, if the desired field of view is not large, these telescopes are used with no corrector at al. They do however suffer from a larger field curvature than does the prime focus because the small, steeply-curved secondary mirror increases the Petzval sum.

Two variations of the Cassegrain system are in common use. If a flat diagonal ***tertiary mirror*** is inserted in front of the primary mirror, the beam returning from the secondary can be diverted sideways. In altazimuth telescopes (discussed in more detail below), this mirror is aligned with the altitude axis, about which the telescope optical assembly pivots, so that the focus is stationary, at least with respect to this motion. This is called a ***Nasmyth*** focus. It allows relatively large instruments to be mounted on Nasmyth platforms attached to the telescope that can be reached with only three reflections. The tertiary mirror can normally be rotated 180 degrees about the symmetry axis in order to direct the light to a second Nasmyth focus, thus providing a rapid change between two instruments.

In telescopes having mounts (axes parallel and perpendicular to the Earth's rotation axis) additional mirrors can be used to direct the light from the tertiary mirror along the polar axis and, finally, to a fixed focus that does not move no matter how the telescope is pointed. This is the ***Coudé*** focus. Here very large instruments, such as high-dispersion spectrographs, can be located.

While the Coudé provides a stationary image, and The Nasmyth focus is stationary with respect to motions along one axis, in both cases the field of view rotates. A star at the center of the field will be stationary, but other stars will rotate about it as the telescope tracks. For imaging applications, where this is important, it can be corrected by either rotating the detector itself, or by passing the light through an ***image rotator*** (a set of flat mirrors or prisms which provide opposite field rotation when they are physically rotated about the axis of the beam).

If a prime focus is not required, it is possible to use the added degree of freedom provided by two mirrors to correct two of the Seidel aberrations. These are normally taken to be spherical aberration and coma. The ***Ritchie-Chrétien*** design employs elliptical surfaces on the primary and secondary mirrors to produce a "Cassegrain" focus that is free from both these aberrations. In wide-field applications, a corrector is still required, to remove at least the astigmatism and field curvature. The 2.5-meter Carnegie telescope at Las Campanas Observatory in Chile is of the Ritchie-Chrétien type and provides a very large field of view with a diameter of 2 degrees.

3.4.3 Three-mirror telescopes

Telescopes employing three mirrors can in principle correct three Seidel aberrations. A general analysis of the theory three-mirror systems has been given by ***Korsh*** (1972). However, a clever three-mirror design was discovered much earlier by M. Paul (1935). In Paul's design, the parabolic primary and convex secondary mirrors alone form an ***afocal*** telescope (a telescope which has an infinite focal length). In other words, the beam of rays emerging from the secondary mirror does not converge or diverge. If the secondary mirror is also parabolic, the 2-mirror system is free from the first three Seidel aberrations. To produce a focus, a third mirror, concave and spherical with the same curvature as the secondary mirror, is placed between the primary and secondary mirrors. It receives the parallel beam from the secondary mirror and focuses it. If the tertiary mirror is located so that its center of curvature is located at the vertex of the secondary mirror, it will act as a Schmidt telescope. The spherical aberration of tins mirror is cancelled not with a corrector lens, but by *removing* the parabolic shape from the secondary mirror, which now also becomes spherical. The resulting ***Paul telescope***, is free from spherical aberration, coma and astigmatism. This design gives a very large field of view with well-corrected images. Because the secondary and tertiary mirrors have equal (but opposite) curvature, their contribution to the Petzval sum cancels leaving only the relatively small contribution from the primary mirror. Thus the field curvature is comparable to that of a prime-focus telescope. In a modification due to Baker (1945), the curvature of the tertiary mirror is adjusted to make the Petzval sum zero. A small aspheric correction is added to the secondary mirror in order remove the residual spherical aberration (which otherwise would no longer cancel completely). The ***Paul-Baker*** design is therefore free from all Seidel aberrations except distortion!

The advantages of a three-mirror telescope come with the penalty of increased cost and complexity. To date, no major telescope has been built using a three-mirror design. However, these systems are being seriously considered for the next generation of large telescopes.

3.4.4 Catadioptic telescopes

Catadioptic telescopes employ both mirrors and lenses as an integral part of their design. The earliest example is the ***Schmidt*** telescope. Bernhard Schmidt realized that a spherical mirror has no optical axis, only a center of curvature. It must therefore be free of all off-axis aberrations, leaving only spherical aberration. If that could be removed by another optical element without introducing an optical axis, the resulting telescope would have a very wide field of view. Schmidt achieved this by introducing a lens, of diameter comparable to the primary mirror. However, this lens had negligible curvature, so its focusing effect was minimal. The thickness of the lens was nearly constant, but included a term proportional to the fourth power of the radius. This was chosen to exactly cancel the spherical aberration of the primary mirror - recall that the wavefront error for spherical aberration is proportional to the fourth power of the radius (Table 3.2). In order to avoid introducing a preferred axis, he located this lens at the center of curvature of the primary mirror. (The lens itself has an axis, but since it has no focusing power, other than the small variation needed to correct the spherical aberration, this has only a small effect).

Schmidt telescopes make excellent wide-field survey instruments. For example, the 1.2m Ochin telescope at Palomar was used, in the early 1950s, to photograph the entire Northern sky using photographic plates, each covering 50 square degrees. The Palomar Observatory Sky Survey (POSS) has proven to be an indispensable resource for astronomy. It is presently being repeated using higher-quality plates that are now available (the POSSII), and a complimentary survey of the Southern sky has been made using a sister telescope built and operated by the European Southern Observatory (ESO) in Chile.

The Schmidt telescope has a steeply curved field of view – the center of curvature of the focal surface is the same as the center of curvature of the primary mirror, as it must be because of the spherical symmetry of the system. This can be removed by a field flattener or, in the case of the 1.2m, the glass photographic plate is bent and held at the requisite curvature by clamps.

Several variation of the Schmidt design have become popular. In the ***Baker-Schmidt*** design, a sphericial secondary mirror is added to reflect the light to an accessible focus behind the primary mirror. In other ***Schmidt-Cassegrain*** systems, commonly used in small telescopes, the corrector lens is moved closer to the primary mirror and used to support the secondary mirror, which is hyperbolic (Figure 3.12). This destroys the symmetry of the original Schmidt design, but the corrected field of view is still reasonably large. This makes for a compact high-performance telescope.

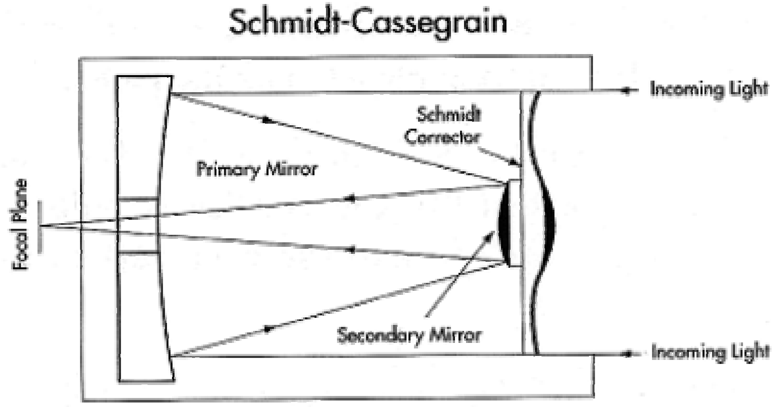


Figure 3.12. A typical Schmidt-Cassegrain telescope. The curved surface of the corrector lens is shown greatly exagurated.

The ***Maksutov*** telescope is similar conceptually to the Schmidt. However, the spherical aberration of the primary mirror is removed not by an (almost) flat lens, but by a steeply curved ***meniscus*** lens. This lens has little focusing power, because the curvatures of the two surfaces are almost equal, but it serves to introduce spherical aberration equal and opposite to that of the spherical mirror. As with the Schmidt, many variations are possible, including Cassegrain systems in which the secondary mirror is mounted on the corrector lens.

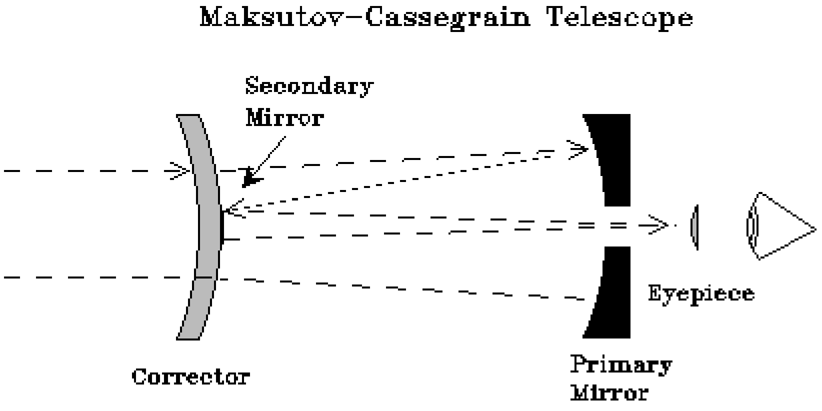


Figure 3.13. A Maksutov telescope.The spherical aberration introduced by the meniscus lens corrector cancels that of the spherical primary mirror.

3.4.5 X-ray telescope optics

The optical systems discussed so far have application at radio, microwave, infrared, optical and ultraviolet wavelengths. At shorter wavelengths the reflectivity of metals falls, so conventional optics cannot be used to image X-rays. However, if the angle of incidence is close to 90 degrees, satisfactory reflectivity can be obtained. X-ray telescopes therefore employ ***grazing-incidence*** optics. Figure 3.14 shows the optical design of the Chandra X-ray telescope. The surface shapes are like those of a Cassegrain system, with a parabolic primary and hyperbolic secondary mirror. However, as the figure shows, they are illuminated at grazing incidence. In order to increase the collecting area, four concentric sets of mirrors are employed, as shown in Figure 3.15. The entrance pupil of the telescope is a set of four concentric annuli.

3.5 Mechanical and control systems

A telescope consists of more that just optical components. The optics must all be held accurately in position with respect to each other, and a means must be provided of pointing and tracking astronomical objects. At the same time, the effects of gravitational forces, thermal variations, etc must all be compensated or minimized.

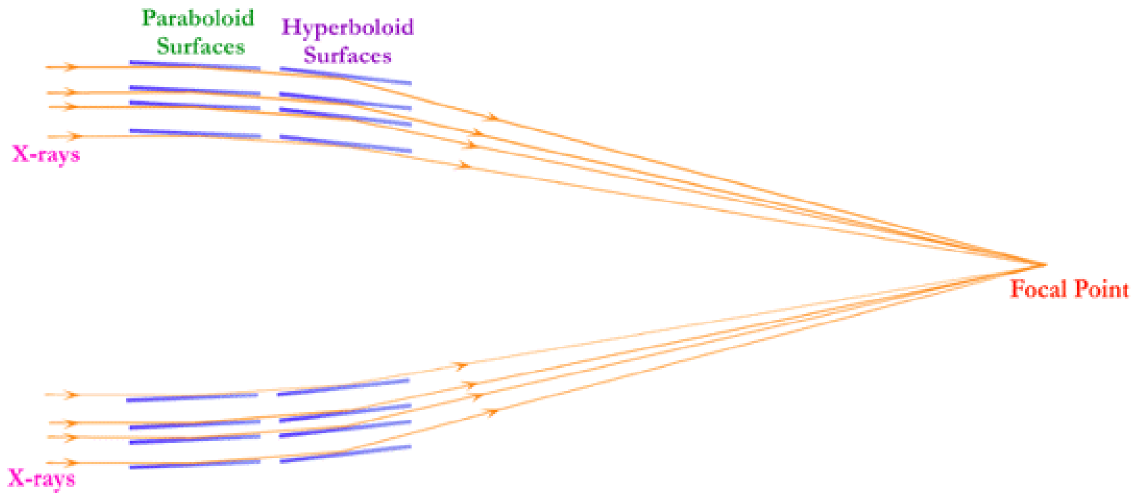


Figure 3.14. Grazing-incidence optics employed in the Chandra X-ray telescope.

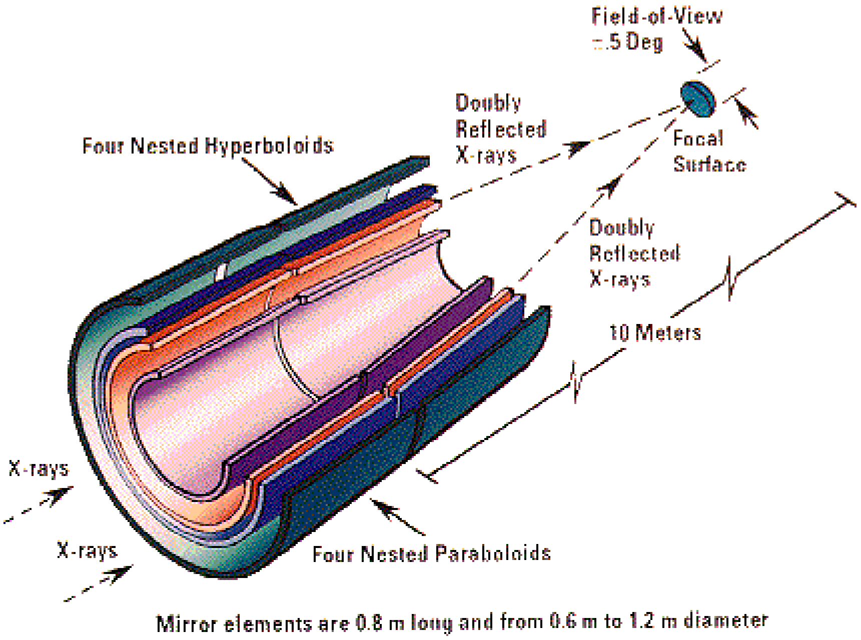


Figure 3.15. Arrangement of mirrors in the Chandra X-ray telescope.

3.5.1 Telescope mounts

The moving optical components of a telescope are generally mounted in a framework called the optical tube assembly (OTA). In small telescopes this is usually a tube with the primary mirror at one end. In large telescopes the mass and wind resistance of a tube would be a disadvantage so a lightweight open steel framework is employed.

In order for the telescope to point and track objects, the OTA and associated structure must rotate about two orthogonal axes. There are two distinctly-different ways to accomplish this. In ***equatorial*** mounts one axis, called the ***polar axis***, is aligned parallel to the rotation axis of the Earth. If the telescope is rotated about this axis, in the opposite sense as the Earth rotates, distant stars and galaxies will appear stationary regardless of their direction. The other axis (the ***declination axis***) allows the telescope to turn in a North-South direction and thereby access any point in the sky.

While equatorial mounts offer convenience in tracking, they are bulky because to the direction of the polar axis depends on latitude and is not aligned vertically or horizontally (unless the observatory is on the equator or at the North or South pole). The ***altazimuth*** mounting system offers simplicity and light weight at the expense of more complicated tracking requirements. In an alt-azimuth mount, one axis is vertical (the ***azimuth axis***) and the other (the ***altitude axis***) is horizontal. Since there axes are not alighted with that of the Earth, tracking requires motion about both axes at rates that depend on the direction that the telescope is pointed. This is accomplished by precise drive motors under computer control. The most recent large telescopes have employed this type of mount as it offers considerable cost savings. A disadvantage of altazimuth designs is that a region of sky surrounding the ***zenith*** (the direction directly overhead) is inaccessible because very high rate azimuth motions would be needed to track objects in this region. For this reason, an area of several degrees diameter is excluded. Another disadvantage, compared to the equatorial design, is that the field of view rotates as the telescope tracks. This can be compensated by rotating the instrument or by using a field rotator, but this is an added degree of complexity.

3.5.2 Telescope control systems

Telescopes are point and track with the help of encoders attached to the mechanical axes. Theses sense the position of the telescope with a resolution of a fraction of an arcsec. The actual direction in the sky to which the telescope is pointing depends on many factors in addition to the positions of the axes. These include the date and time, the type of mounting system, errors in alignment, gravitational flexure, thermal expansion, atmospheric refraction, etc. The telescope is normally designed in such a way as to minimize the effects of thermal expansion (the mirrors are always made of materials that have extremely low thermal expansion coefficients). Effects of gravitational flexure and atmospheric refraction can be calibrated by pointing the telescope to stars at known positions and mapping the pointing errors. These errors are entered into a lookup table and used to correct the telescope pointing.

The mapping between celestial coordinates and telescope position depends on the date and time. Celestial positions are described by ***right ascension*** (α) and ***declination*** (δ), which are measured with respect to the celestial equator and the vernal equinox (the position where the ecliptic intersects the celestial equator). In an equatorial mount, these are directly related to the positions of the telescope axes. As the Earth rotates, the right ascension of a fixed direction (with respect to the Earth) increases. The right ascension of the ***meridian*** (the great circle passing through the zenith and celestial poles) is called the ***sidereal time*** (*s*), and is the primary factor determining whether a particular right ascension can be observed. The rotation angle of the polar axis is called the ***hour angle*** (*h*), and is measured westward from the meridian. These quantities are therefore related by the simple formula

$$\alpha = s - h \qquad (3.12)$$

Another quantity of interest is the zenith angle *Z* of the object. This can be found using spherical trigonometry. The result is

$$\cos Z = \sin\delta \sin l + \cos\delta \cos l \cos h\,, \qquad (3.13)$$

where *l* is the latitude of the telescope.

For an altazimuth telescope, the altitude angle, measured upward from the horizon, is just $\pi / 2 - Z$, and the azimuth angle ϕ, measured from North, is given by

$$\sin\phi = -\cos\delta \sin h / \sin Z\,. \qquad (3.14)$$

While pointing is usually done entirely by reference to coordinates, accurate tracking requires a guide star and feedback loop. A suitable star in the peripheral field of the telesdcope is selected and imaged with a small video camera. The position of the star image in the video field is sensed and an error signal is generated and sent to the telescope control system (TCS). The TCS converts this into small corrections in the pointing of the telescope and the tracking rate, in order to keep the guide star stationary.

The TCS is generally coupled to other systems, such as the rotation of the dome (to keep ithe slit oriented in the direction that the telescope is pointing), the detector data acquisition system (to provide timing, pointing and other information), instruments (to provide control) and various telescope mechanical systems. A large modern astronomical telescope is a very complex piece of equipment controlled by sophisticated hardware and software systems.

Exercises

3.1 A 2048×2048 CCD array with 15-um pixels is placed at the f/8 Cassegrain focus of a 4-meter telescope. What is the image scale in arcsec/pixel and what are the angular dimensions of the region of sky covered by the CCD?

3.2 Prove that the focal length of a mirror is ½ its radius of curvature.

3.3 By direct geometrical ray tracing, determine the linear and angular diameter of the image formed at the (paraxial) focus of a spherical mirror, of an infinitely distant point source. Express the answer in terms of the focal length and *f* number of the mirror.

3.4 Show that a flat plat of glass, of thickness *t* and index *n*, placed in a converging beam, and oriented perpendicular to the axis, moves the position of the focal plane a distance $t(n-1)/n$.

Using Snell's law, evaluate the OPD to third order. What aberration is introduced by the glass?

3.5 The 3.6m Canada-France Hawaii Telescope has an f/2.8 primary mirror and an f/8 Cassegrain focus. Assume that the f/8 focus lies 1.0 metres behind the vertex of the primary mirror and determine the radius of curvature of the focal surface. In what direction does the focal surface curve?

4 Imaging

This chapter deals with a number of topics related to image formation and image quality. A quantitative description of imaging is developed and the theory of atmospheric seeing is summarized. The chapter concludes with a discussion of interferometry and aperture synthesis techniques.

4.1 Linear image theory

We begin with a general analysis of imaging systems. This applies not only to telescopes, but any kind of system that focuses or modifies images, the Earth's turbulent atmosphere for example.

4.1.1 The telescope as a linear system

A telescope can be considered as a linear system. It is a device that transforms an input image to an output image. The input image is represented by the distribution of intensity $I(x,y)$ as a function of coordinates (x,y) which we take to taken to be angular position in the sky. The output image $I'(x,y)$ is the distribution of intensity in the focal plane as a function of the same angular coordinates. Because Maxwell's equations are linear in the electric and magnetic fields, one can write the amplitude of these quantities in the output image as a linear function of the amplitude in the input image. (If polarization is important, we must consider the vector fields rather than just their scalar amplitudes, but normally the scalar treatment is sufficient.) Thus, we write

$$a' = \mathcal{O}(a) \tag{4.1}$$

where a and a' are the input and output amplitudes

$$I = aa^* \tag{4.2}$$

and $\mathcal{O}$ is a linear operator. *Linear* means that

$$\mathcal{O}(c_1 a_1 + c_2 a_2) = c_1 \mathcal{O}(a_1) + c_2 \mathcal{O}(a_2) \tag{4.3}$$

where c_1 and c_2 are constants. Given that the imaging properties of a telescope (and the atmosphere also) are fully described by a linear operator, we can now apply the standard tools of linear analysis.

4.1.2 Transfer function

We start with a Fourier decomposition. Any reasonably continuous function $a(x, y)$ can be written in terms of the Fourier integral

$$a(\mathbf{x}) = \int \tilde{a}(\mathbf{k}) e^{2\pi i \mathbf{k}\cdot\mathbf{x}} d^2k \tag{4.4}$$

where $\mathbf{x} = (x, y)$, the varibles $\mathbf{k} = (k_x, k_y)$ describe spatial frequencies, and the integral is taken over the infinite two-dimensional plane. The complex function $\tilde{a}(\mathbf{k})$ describes the amplitudes and phases of the spatial frequencies that comprise the image. It can be obtained from the inverse transform of the amplitude

$$\tilde{a}(\mathbf{k}) = \int a(\mathbf{x}) e^{-2\pi i \mathbf{k}\cdot\mathbf{x}} d^2x \tag{4.5}$$

Similarly, we may write

$$a'(\mathbf{x}) = \int \tilde{a}'(\mathbf{k}) e^{2\pi i \mathbf{k}\cdot\mathbf{x}} d^2k \tag{4.6}$$

for the output image.

According to (4.1) and (4.3),

$$a'(\mathbf{x}) = \int \tilde{a}(\mathbf{k}) \mathcal{O}\left(e^{2\pi i \mathbf{k}\cdot\mathbf{x}}\right) d^2k \,. \tag{4.7}$$

Comparing (4.7) and (4.6), we see that $\mathcal{O}$ must have the form

$$\mathcal{O}\left(e^{2\pi i \mathbf{k}\cdot\mathbf{x}}\right) = O(\mathbf{k}) e^{2\pi i \mathbf{k}\cdot\mathbf{x}} \tag{4.8}$$

where $O(\mathbf{k})$ is some function of $\mathbf{k}$ alone, and

$$\tilde{a}'(\mathbf{k}) = O(\mathbf{k}) \tilde{a}(\mathbf{k}) \,. \tag{4.9}$$

In other words, the Fourier components of the output image are just the Fourier components of the input image multiplied by the function $O(\mathbf{k})$. This function is called the ***optical transfer function*** (OTF) of the system.

4.1.3 Impulse response

Consider the response to an input image that has the form of a delta function (a distant star, for example),

$$a(\mathbf{x}) = \delta(\mathbf{x}) \,. \tag{4.10}$$

The delta function has the integral representation

$$\delta(\mathbf{x}) = \int e^{2\pi i\mathbf{k}\cdot\mathbf{x}} d^2k \tag{4.11}$$

so its Fourier transform is the unit function

$$\tilde{\delta}(\mathbf{k}) = 1 \tag{4.12}$$

Using (4.9), (4.11) and (4.12) we find, for the amplitude of the output image,

$$\begin{aligned} a'(\mathbf{x}) &= \int O(\mathbf{k}) e^{2\pi i\mathbf{k}\cdot\mathbf{x}} d^2k \\ &= \tilde{O}(\mathbf{x}), \end{aligned} \tag{4.13}$$

which is just the inverse Fourier transform of the OTF. This amplitude is called the ***impulse response*** (IR). Equivalently, the OTF is the Fourier transform of the impulse response.

4.1.4 Point spread function

We are usually more interested in the intensity distribution in the image, as this is what most detectors respond to. The intensity distribution $I'(\mathbf{x})$ for the image of a point source is called the ***point spread function*** (PSF). From (4.2),

$$I'(\mathbf{x}) = \left|a'(x)\right|^2 \tag{4.14}$$

so the PSF is the square of the IR. It is usual to normalize the PSF so that its integral, over the image plane, is unity

$$\int I'(\mathbf{x})\, d^2x = 1 \tag{4.15}$$

4.1.5 Modulation transfer function

If the light is incoherent, so that interference effects can be ignored, the linear analysis can be applied directly to intensities. As before, we obtain the OTF, but now it refers to intensities rather than amplitudes, and is the Fourier transform (F) of the PSF. In general, the OTF is a complex quantity and can be written in the form

$$O(\mathbf{k}) = M(\mathbf{k}) \exp\left(iP(\mathbf{k})\right) \tag{4.16}$$

Where $M(\mathbf{k})$ and $P(\mathbf{k})$ are real-valued function functions, called the ***modulation transfer function***, (MTF) and ***phase transfer function*** (PTF) respectively. The MTF is therefore the modulus of the OTF.

4.2 Fourier imaging theory

In this section we consider diffraction effects and show that imaging devices are in fact Fourier transform machines.

Consider a telescope imaging a very distant source of radiation. In the entrance pupil P of the telescope, we set up coordinates $\mathbf{q} = (q_x, q_y)$ which describe the position of any point in the pupil. Let the amplitude of the radiation of wavelength λ (that is, radiation with wavelength in a small but finite interval containing λ) at that point be $a(\mathbf{q})$. In the focal plane we set up angular coordinates $\mathbf{x}$. For any given point in the focal plane, the components of x are the angles, measured from the optical axis, of the incident radiation that is focused that point.

According to the Huygens-Fresnel principle, to obtain the amplitude in the focal plane, we sum over all possible contributions from the pupil including the phase factors that result from propagation

$$a(\mathbf{x}) = \int_P a(\mathbf{q}) \exp\left[\frac{2\pi i}{\lambda} L(\mathbf{q}, \mathbf{x})\right] d^2q \tag{4.17}$$

Here $L(\mathbf{q}, \mathbf{x})$ is the optical path length through the telescope from the point $\mathbf{q}$ to the point $\mathbf{x}$.

Now suppose that the telescope is free from aberrations. That is to say, it focuses all incident plane waves to points in the focal plane. Then by Fermat's theorem, the optical path length from a point $\mathbf{x}$ in the focal plane to all points in a plane, centered on the entrance pupil, whose normal is oriented at angle $\mathbf{x}$ with respect to the optical axis must be constant (we have assumed here that the angle is small). To within an unimportant constant phase factor, we may then replace $L(\mathbf{q}, \mathbf{x})$ in (4.17) with the perpendicular distance from the point $\mathbf{q}$ to the plane that corresponds to point $\mathbf{x}$. Thus,

$$L(\mathbf{q}, \mathbf{x}) = \mathbf{x} \cdot \mathbf{q} + const \tag{4.18}$$

Ignoring the constant factor, (4.17) now becomes

$$a(\mathbf{x}) = \int_P a(\mathbf{q}) \exp\left[\frac{2\pi i}{\lambda} \mathbf{x} \cdot \mathbf{q}\right] d^2q\,. \tag{4.19}$$

If we define the dimensionless coordinates $\mathbf{u} = \mathbf{q}/\lambda$, and define the amplitude $a(\mathbf{u})$ to be zero outside the pupil so that the integral can be extended to infinity, (4.19) becomes

$$a(\mathbf{x}) = \lambda^2 \int_{-\infty}^{\infty} a(\mathbf{u}) e^{2\pi i \mathbf{x}\cdot\mathbf{u}} d^2u \qquad (4.20)$$

From this it can be seen that the amplitude distribution in the focal plane is proportional to the inverse Fourier transform of the amplitude distribution in the entrance pupil.

For functions that have circular symmetry, the two-dimensional Fourier transform becomes the Hankel transform

$$\begin{aligned} a(\theta) &= \lambda^2 \int_0^{\infty}\int_0^{2\pi} a(\rho) e^{2\pi i \rho\theta\cos\phi} d\phi \rho d\rho \\ &= 2\pi\lambda^2 \int_0^{\infty} a(\rho) J_0(2\pi\rho\theta)\rho d\rho \end{aligned} \qquad (4.21)$$

where $\theta = |\mathbf{x}|, \quad \rho = |\mathbf{u}|$ and we have used the integral form of the Bessel function

$$J_v(x) = \frac{1}{2\pi}\int_0^{2\pi} e^{i(x\cos\phi - v\phi)} d\phi\,. \qquad (4.22)$$

4.2.1 The Airy profile

Let us now apply this result to determine the PSF of an aberration-free telescope having a circular aperture. Consider a plane wave traveling parallel to the optical axis. Inside the entrance pupil the amplitude is constant and outside it is zero, so we write

$$a(\mathbf{u}) = a_0 P(\mathbf{u}) \qquad (4.23)$$

where $P(\mathbf{u})$ is called the ***pupil function***. For a circular aperture of diameter *D*, it has the form

$$P(\mathbf{u}) = \Pi(\lambda\rho / D) \qquad (4.24)$$

where

$$\Pi(\rho) \equiv \begin{cases} 1 & \rho <= 1/2 \\ 0 & \rho > 1/2 \end{cases} \qquad (4.25)$$

From (4.21), the amplitude distribution in the focal plane is now given by

$$\begin{aligned} a(\theta) &= 2\pi a_0 \lambda^2 \int_0^{D/2\lambda} \mathrm{J}_0(2\pi\rho\theta)\rho d\rho \\ &= \frac{a_0 \pi D^2}{4}\left[\frac{2\mathrm{J}_1(\pi\theta D/\lambda)}{\pi\theta D/\lambda}\right] \end{aligned} \tag{4.26}$$

where we have dropped the primes used earlier. Thus, the intensity in the focal plane is

$$I(\theta) = I_0 A(\theta D/\lambda), \tag{4.27}$$

where I_0 is the central intensity and $A(x)$ is the ***Airy profile***, defined by

$$\mathsf{A}(x) = \left[\frac{2\mathrm{J}_1(\pi x)}{\pi x}\right]^2 \tag{4.28}$$

and illustrated in Figure 4.1. This function fall from unit value at the center to zero at the zeros of J_1. The angular radius of the first zero is approximately $1.22\lambda/D$.

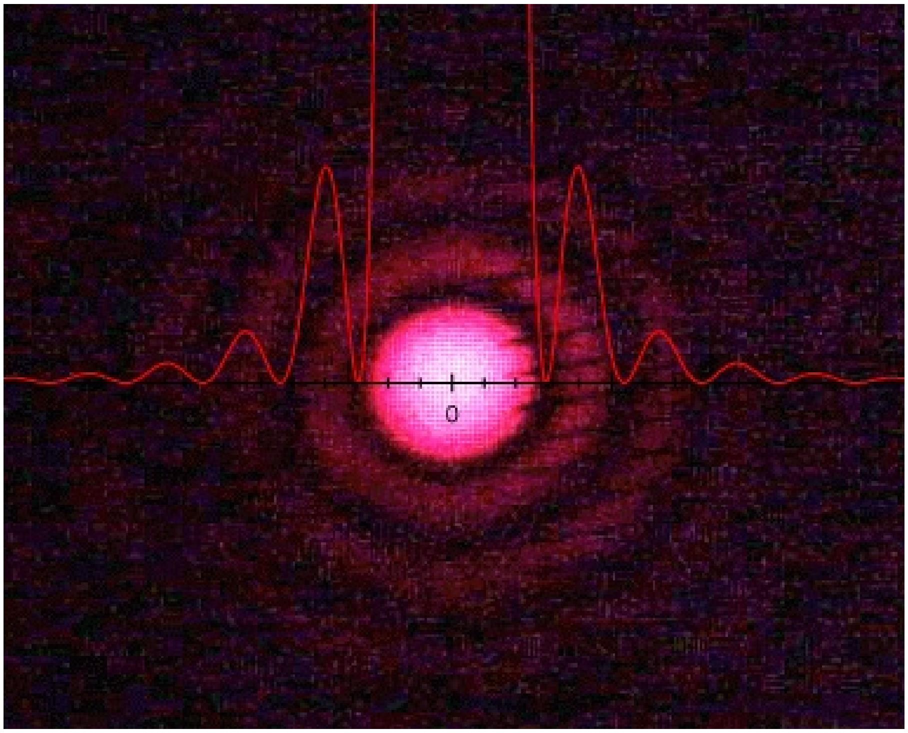


Figure 4.1 PSF of an ideal telescope. The angular diameter of the central Airy disk is $2.44\lambda/D$, where λ is the wavelength and D is the aperture diameter.

The OTF for a perfect telescope can be obtained by taking the Fourier transform of the PSF $I'(\mathbf{x})$.

$$\begin{aligned} O(\mathbf{k}) &= \mathcal{F}\left(I\right)(\mathbf{k}) \\ &= \mathcal{F}\left(a' \cdot a'^{*}\right)(\mathbf{k}) \\ &= \mathcal{F}\left(a'\right) * \mathcal{F}\left(a'^{*}\right)(\mathbf{k}) \\ &= \tilde{a}'(\mathbf{k}) * \tilde{a}'^{*}(\mathbf{k}) \end{aligned} \tag{4.29}$$

Where we have used the Fourier convolution theorem. (Note that by virtue of (4.4) and (4.20) the variables $\mathbf{k}$ and $\mathbf{u}$, both dimensionless, are identical.) Thus, the OTF is the convolution of the complex amplitude distribution in the exit pupil with itself (more accurately, with its complex conjugate). The OTF is therefore proportional to the convolution of the pupil function with itself

$$O(\mathbf{u}) \propto P(\mathbf{u}) * P(\mathbf{u}) = \Pi(\lambda\rho / D) * \Pi(\lambda\rho / D) \tag{4.30}$$

This function is real and depends only on the separation ρ. Therefore, the MTF equals the OTF. Evaluating the right hand side of (4.30) gives

$$M_0(\rho) = 1 - \frac{2}{\pi}\left[\sin^{-1}(\lambda\rho / D) + \frac{\lambda\rho}{D}\sqrt{1 - \lambda^2\rho^2 / D^2} \right]. \tag{4.31}$$

which is plotted in Figure 4.1. There is a steady decline with increasing spatial frequency because large frequencies are less well sampled by the aperture. The maximum spatial frequency that can be sampled by the telescope is D/λ. The diffraction-limited MTF (4.31) is the best possible response that an optical system with this diameter entrance pupil can provide. In practice, the response is generally less, due to optical aberrations, atmospheric seeing, and other such effects.

4.3 Atmospheric seeing

The Earth's atmosphere is an inhomogeneous medium that affects the propagation of light. The index of refraction of air depends on wavelength, pressure and temperature

$$\begin{aligned} n &\approx n_0 + 0.79 \times 10^{-6} P / T \\ n_0 &= 1.000272643 + 1.2288 \times 10^{-12} \lambda^{-2} + 0.03555 \times 10^{-6} \lambda^{-4} \quad (\lambda \text{ in nm}) \end{aligned} \tag{4.32}$$

The troposphere contains layers of turbulent air with variations in density and temperature. Mechanical energy, injected on large scales by convection and wind, is transferred to smaller and smaller scales by viscous forces until it is dissipated as heat. This results in a distribution of turbulent eddy sizes that is well described by Kolmogorov's theory of turbulence (see Landau & Lifshitz, Fluid Mechanics).

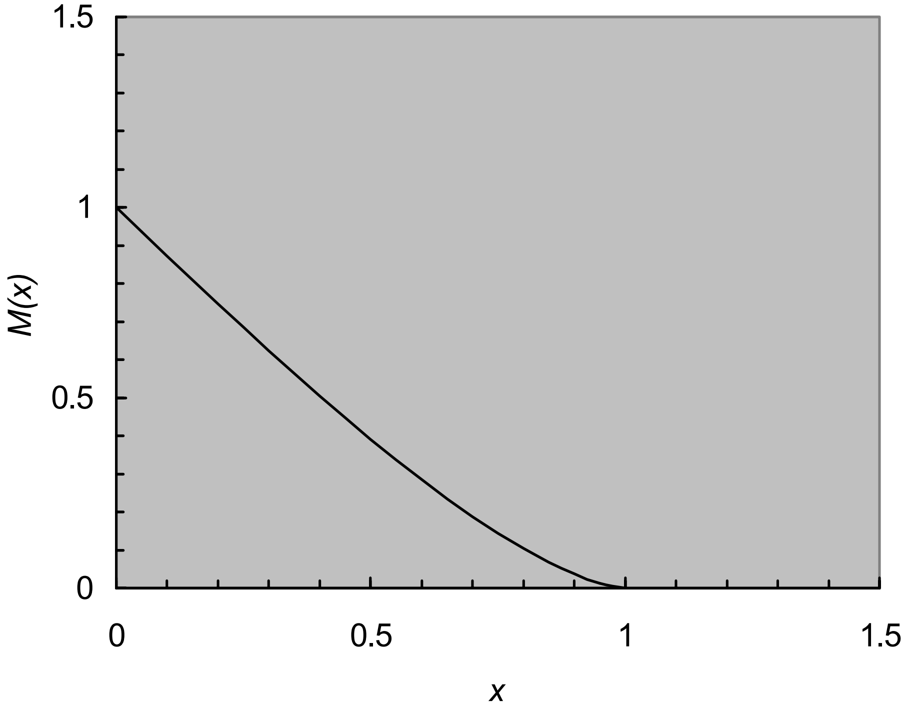


Figure 4.2. Diffraction-limited MTF for a circular aperture. The horizontal axis refers to spatial frequency in units of *D*/*λ*.

The propagation of electromagnetic radiation through a turbulent atmosphere has been studied in detail (eg. Tatarski 1967, see also the review by Roddier 1981), at least for the Kolmogorov spectrum. Essentially, the atmosphere acts as a phase screen, introducing random phase errors into the wavefront. Consider light propagating to Earth from a distant star and imagine that the turbulence is confined to a single thin layer, of thickness δh, at a height h above the telescope. The distance from the telescope to this layer, along the line of sight, is $z = h\sec\varsigma$, where ζ is the zenith angle of the star. The complex amplitude of the radiation field at the bottom of the layer can be written as

$$\Phi(\mathbf{x}, z) = \exp\left(i\phi(\mathbf{x}, z)\right) \tag{4.33}$$

where $\mathbf{x}$ represents two-dimensional coordinates in the plane perpendicular to the line of sight and the amplitude has been normalized to unity at the top of the layer. For astronomical observations at moderate zenith angles, $\phi(\mathbf{x}, z) << 1$ and so we may write

$$\Phi(\mathbf{x}, z) \approx 1 + i\phi(\mathbf{x}, z) \tag{4.34}$$

The propagation of light from the bottom of the layer to the telescope is well-described by Fresnel diffraction. Accordingly, the complex amplitude at the detector is given by

$$\begin{aligned}\Phi(\mathbf{x}) \equiv \Phi(\mathbf{x},0) &= \Phi(\mathbf{x},z) * \frac{1}{i\lambda z}\exp\left(i\pi\frac{x^2}{\lambda z}\right) \\ &= 1 + \phi(\mathbf{x},z) * \frac{1}{\lambda z}\exp\left(i\pi\frac{x^2}{\lambda z}\right)\end{aligned} \qquad (4.35)$$

where ‘ $*$ ’ denotes convolution.

The real part of the second term on the right-hand side of (4.35)

$$\chi(\mathbf{x}) = \phi(\mathbf{x},z) * \frac{1}{\lambda z}\cos\left(\pi\frac{x^2}{\lambda z}\right) \qquad (4.36)$$

describes fluctuations in the logarithm of the amplitude $|\Phi(\mathbf{x},z)|$. The imaginary part

$$\varphi(\mathbf{x}) = \phi(\mathbf{x},z) * \frac{1}{\lambda z}\sin\left(\pi\frac{x^2}{\lambda z}\right) \qquad (4.37)$$

describes fluctuations in the phase.

These fluctuations vary on a timescale that is typically on the order of a few milliseconds and are essentially random in nature. Therefore, we can describe only statistical properties of these fluctuations. The covariance of two random variables f and g is defined by

$$C_{fg} = \langle fg \rangle - \langle f \rangle \langle g \rangle \qquad (4.38)$$

Extending this idea to fluctuations of the same function but at different positions, one defines ***spatial covariance functions***

$$\begin{aligned}C_\chi(\mathbf{r}) &= \langle \chi(\mathbf{x})\chi(\mathbf{x}+\mathbf{r}) \rangle \\ C_\varphi(\mathbf{r}) &= \langle \varphi(\mathbf{x})\varphi(\mathbf{x}+\mathbf{r}) \rangle\end{aligned} \qquad (4.39)$$

where the brackets indicate a spatial average over the variable $\mathbf{x}$. In these the second term on the right-hand side of (4.38) is missing because the mean values of the fluctuations are zero. Even though the phase and amplitude fluctuations vary randomly, these statistical averages are well defined and suffice to characterize the effects of the turbulence.

According to the ***Weiner-Kinchine theorem***, the Fourier transforms of the covariance functions are the power spectra $W_\chi(\mathbf{k})$ and $W_\varphi(\mathbf{k})$ defined by

$$\begin{aligned} W_\chi(\mathbf{k}) &= \left|\tilde{\chi}(\mathbf{k})\right|^2 \\ W_\varphi(\mathbf{k}) &= \left|\tilde{\varphi}(\mathbf{k})\right|^2 \end{aligned} \tag{4.40}$$

To compute these, take the Fourier transform of (4.36) and use the convolution theorem

$$\begin{aligned} \tilde{\chi}(\mathbf{k}) &= F\left[\phi(\mathbf{x},z) * \frac{1}{\lambda z}\cos\left(\pi\frac{x^2}{\lambda z}\right)\right] \\ &= \tilde{\phi}(\mathbf{k})\cdot F\left[\frac{1}{\lambda z}\cos\left(\pi\frac{x^2}{\lambda z}\right)\right] \\ &= \tilde{\phi}(\mathbf{k})\sin\left(\pi\lambda z\kappa^2\right) \end{aligned} \tag{4.41}$$

Putting this result in (4.40) gives

$$\begin{aligned} W_\chi(\mathbf{k}) &= W_\phi(\mathbf{k})\sin^2\left(\pi\lambda z\kappa^2\right) \\ W_\varphi(\mathbf{k}) &= W_\phi(\mathbf{k})\cos^2\left(\pi\lambda z\kappa^2\right) \end{aligned} \tag{4.42}$$

where $W_\phi(\mathbf{k}) = \left|\tilde{\phi}(\mathbf{k})\right|^2$ is the power spectrum of the complex phase fluctuations at the bottom of the turbulence layer. Our next task is to determine the form of $W_\phi(\mathbf{k})$, the power spectrum of phase fluctuations produced by atmospheric turbulence. One begins by assuming that the statistical properties of the atmospheric turbulence do not depend on the orientation perpendicular to the line of sight. Therefore the power spectrum can depend only on the magnitude κ of the spatial frequency $\mathbf{k}$. To find the form of the power spectrum, we need a theory of turbulence.

4.3.1 Atmospheric turbulence

Over a wide range of scales, atmospheric turbulence is well-described by the ***Kolmogorov turbulence*** spectrum. In Kolmogorov's theory, large-scale turbulent cells are produced by the action of wind. The largest scale of such cells is called the ***outer scale*** of turbulence, L_o. The air flow within these cells is itself unstable, producing smaller cells. These in turn produce cells that are smaller still and the process continues until viscous damping becomes important. At this point energy is dissipated as heat. The size of the smallest turbulent cells is call the ***inner scale*** L_i.

Kolmogorov turbulence therefore involves a cascade of energy from large scales to small scales. At any scale, the rate at which energy enters this scale, per unit mass, is the same as the rate at which energy leaves this scale. This rate, denoted by ϵ, is therefore independent of scale. Now, associated with every cell, we have a typical length l, and a typical air velocity υ. A characteristic time for the cell is $\tau = \upsilon / l$.

The energy per unit mass is

$$\varepsilon \sim \upsilon^2 \tag{4.43}$$

and the energy dissipation rate is

$$\epsilon \sim \varepsilon / \tau \sim \upsilon^3 / l \tag{4.44}$$

Therefore, the typical velocity in a cell of size l is

$$\upsilon \sim \left(\epsilon l\right)^{2/3} \sim l^{2/3} \tag{4.45}$$

since ϵ is constant.

Now, consider the energy per unit mass contained within turbulent cells having scales in some finite range, say l to $2l$. From (4.43) and (4.45) this must be proportional to $l^{2/3}$. This energy can also be found by integrating the power spectrum of the turbulent energy $W_E(\kappa)$, over the corresponding range of spatial frequencies, from $\kappa \sim 1/l$ to $\kappa/2$. Therefore we must have

$$\int_{\kappa/2}^{\kappa} W_E(\kappa)\kappa^2 d\kappa \sim l^{2/3} \sim \kappa^{-2/3} \tag{4.46}$$

which can only be true if

$$W_E(\kappa) \sim \kappa^{-11/3} \tag{4.47}$$

One can now argue that the power spectrum of other physical quantities such as the temperature and ultimately the index of refraction must have the same form. Therefore, we can write the power spectrum of index of refraction fluctuations as

$$W_N(\kappa) = 0.033\, C_N^2\, \kappa^{-11/3} \tag{4.48}$$

where the proportionality constant $0.033\, C_N^2$, a measure of the strength of the turbulence, is a function of height. The parameter C_N^2 is called the ***structure constant*** of index of refraction fluctuations. The factor of 0.033 appears because C_N^2 is actually defined as the coefficient of the structure function (defined in 4.3.4) of index of refraction fluctuations.

The two-dimensional phase power spectrum can now be obtained from the three-dimensional index of refraction power spectrum. One multiplies by $(2\pi/\lambda)^2$, to obtain phase from index of refraction, and integrates over the z variable (see Roddier 1981, Section 3). The result is

$$W_\phi(\mathbf{k}, z) = 0.38 \sec\zeta\, \lambda^{-2} \kappa^{-11/3} C_N^2(h)\delta h \tag{4.49}$$

Generally, turbulence occurs over the entire lower atmosphere, extending from the surface to a height of tens of kilometers. We can treat this general case by dividing the atmosphere into many (an infinite number) of thin layers. Each layer contributes a small amount to the phase fluctuations, and these contributions add linearly. The total effect is found by integrating equations (4.42) over h. The result is

$$\begin{aligned} W_{\phi}(\mathbf{k}) &= 0.38\sec\zeta\,\lambda^{-2}\kappa^{-11/3}\int_0^{\infty} C_N^2(h)\,dh \\ W_{\chi}(\mathbf{k}) &= 0.38\sec\zeta\,\lambda^{-2}\kappa^{-11/3}\int_0^{\infty} C_N^2(h)\sin^2\left(\pi\lambda h\sec\zeta\,\kappa^2\right)dh \\ W_{\varphi}(\mathbf{k}) &= 0.38\sec\zeta\,\lambda^{-2}\kappa^{-11/3}\int_0^{\infty} C_N^2(h)\cos^2\left(\pi\lambda h\sec\zeta\,\kappa^2\right)dh \end{aligned} \tag{4.50}$$

In practice, the argument of the trigonononometric functions in the integrals is small and we may make the near-field approximation. This gives

$$\begin{aligned} W_{\chi}(\mathbf{k}) &= 0.38\pi^2\sec^3\zeta\,\kappa^{1/3}\int_0^{\infty} C_N^2(h)\,h^2dh \\ W_{\varphi}(\mathbf{k}) &= 0.38\sec\zeta\,\lambda^{-2}\kappa^{-11/3}\int_0^{\infty} C_N^2(h)\,dh \end{aligned} \tag{4.51}$$

4.3.2 Scintillation

The first equation in (4.51) describes the power spectrum of log-amplitude fluctuations that give rise to scintillation. The variance of these fluctuations is given by

$$\sigma_{\chi}^2 = \left\langle \chi(\mathbf{x})^2 \right\rangle = C_{\chi}(0) \tag{4.52}$$

Since intensity is the squared modulus of the amplitude, the variance of intensity fluctuations seen by a telescope is

$$\begin{aligned} \sigma_I^2 &= 4\sigma_{\chi}^2 = 4C_{\chi}(0) \\ &= 8\pi\int_0^{\infty} W_{\chi}(\kappa)G(\kappa)\kappa\,d\kappa \end{aligned} \tag{4.53}$$

To obtain the last line we have used the fact that the covariance function is the inverse Fourier transform of the power spectrum, so its value at the origin is equal to the integral of the power spectrum. The weighting function $G(\kappa)$ accounts for the fact that the telescope aperture averages over spatial frequencies. Fluctuations of spatial frequency κ have a spatial oscillation scale of order κ^{-1}. If this is much smaller than the diameter of the telescope aperture, the fluctuation will be effectively averaged out. Therefore, we expect that $G(\kappa)$ will drop to zero at a frequency $\kappa \approx 1/D$ effectively cutting off the integral. In fact, this weighting function is the square of the Fourier transform of the telescope pupil function, defined to have unit value inside the aperture and be zero outside the aperture.

$$G(\mathbf{k}) = \left| \int P(\mathbf{x}) e^{-2\pi i \mathbf{k}\cdot\mathbf{x}} d^2 x \right|^2 \tag{4.54}$$

For distances $z << D^2 / \lambda$, the argument of the trigonometric functions in the integrals of (4.50) is small, so a small angle approximation can be used. This is called the ***near field*** limit. From(4.50) and (4.53) we see that, in the near field, the variance of intensity fluctuations increases with distance from the turbulent layer in proportion to z^2. Therefore the RMS intensity fluctuations increase linearly with distance. *Scintillation increases in proportion to the distance to the turbulent layer*. It is also evident that *scintillation decreases as the aperture size of the telescope increases*, because of the cutoff, at frequency $\kappa \approx 1/D$ due to averaging.

4.3.3 Seeing

The third equation in (4.50) describes fluctuations in phase. Phase errors correspond to a random tilt and distortion of the wavefront. A tilt of the wavefront results in a transverse displacement of the image at the focal plane. A distortion of the wavefront results in a blurring of the image of a star. There is a characteristic scale r_0 (to be defined more precisely in the next section) that describes the wavefront distortion With small-aperture telescopes ($D << r_0$), wavefront tilt is the dominant effect and the image is sharp, having size $\sim \lambda / D$, but dances erratically with a timescale of a fraction of a second. In long exposure images, the random motions of the image produce a smooth Gaussian-like PSF with characteristic size $\sim \lambda / r_0$.

With large telescopes ($D >> r_0$), wavefront distortion is the dominant effect and the image is stable but is composed of a large number of "speckles", each with size $\sim \lambda / D$, The overall size of the envelope of speckles is $\sim \lambda / r_0$. In long exposure images, the random motions of the speckles produce a smooth Gaussian-like PSF with characteristic size $\sim \lambda / r_0$.

At visible wavelengths, $r_0 \sim 10\,\text{cm}$, so atmospheric seeing is the principle factor that determines the resolution of ground-based telescopes. Recently, it has become possible to correct the phase errors, at least in part and at near-infrared wavelengths, by means of Adaptive Optics systems, discussed in 4.4.

4.3.4 Atmospheric structure function

Although the wavefront phase errors produced by the atmosphere are random, and time-varying, their statistical properties are well-defined and related to the physical properties of the turbulence. We have already encountered the ***phase covariance function***

$$C_\varphi(\mathbf{r}) = \langle \varphi(\mathbf{x})\varphi(\mathbf{x}+\mathbf{r}) \rangle . \tag{4.55}$$

Another useful quantity is the ***phase structure function*** to be the ensemble average of the square of the difference between the phase errors at two points

$$D_\varphi(\mathbf{r}) = \left\langle \left(\varphi(\mathbf{x}) - \varphi(\mathbf{x}+\mathbf{r})\right)^2 \right\rangle \tag{4.56}$$

It is easy to see that these two functions are related. By expanding the square in (4.56), and using the linearity of the ensemble average, we find

$$\begin{aligned} D_\varphi(\mathbf{r}) &= \left\langle \varphi(\mathbf{x})^2 \right\rangle + \left\langle \varphi(\mathbf{x}+\mathbf{r})^2 \right\rangle - 2\left\langle \varphi(\mathbf{x})\varphi(\mathbf{x}+\mathbf{r}) \right\rangle \\ &= 2C_\varphi(0) - 2C_\varphi(\mathbf{r}) \end{aligned} \tag{4.57}$$

where

$$C_\varphi(0) = \left\langle \varphi(\mathbf{x})^2 \right\rangle \tag{4.58}$$

is the variance of phase fluctuations at any point on the wavefront. It is usually reasonable to assume that the atmosphere is sufficiently homogeneous that the phase variance is independent of position, at least over distances comparable to the diameters of telescopes. It is also reasonable to assume that the phase fluctuations are isotropic, that is to say the covariance and structure functions depend only on the distance between the two points, and not on the orientation. We then have

$$\begin{aligned} C_\varphi(\mathbf{r}) &= C_\varphi(r) \\ D_\varphi(\mathbf{r}) &= D_\varphi(r) . \\ r &= |\mathbf{r}| \end{aligned} \tag{4.59}$$

Let us now compute the structure function for atmospheric turbulence. Using (4.57) we write the structure function in terms of the covariance function and then use the fact that the covariance function is the (inverse) Fourier transform of the power spectrum given in (4.50). Recalling that for circular symmetry the Fourier transform becomes a Hankel transform (see (4.21)), and using the near-field approximation, we find.

$$\begin{aligned} D_\varphi(r) &= 2C_\varphi(0) - 2C_\varphi(r) \\ &= 4\pi \int_0^\infty W_\varphi \left[1 - J_0(2\pi\kappa r)\right] \kappa d\kappa \\ &= 1.52\pi \sec\zeta\, \lambda^{-2} \int_0^\infty C_N^2(h)\, dh \int_0^\infty \kappa^{-11/3} \left[1 - J_0(2\pi\kappa r)\right] \kappa d\kappa \end{aligned} \tag{4.60}$$

The integral over spatial frequency can be done analytically. To three digit accuracy the result is

$$D_\varphi(r) = 6.88\left(r / r_0\right)^{5/3} \tag{4.61}$$

where

$$r_0 = 0.185\left[\sec\zeta\, \lambda^{-2} \int_0^\infty C_N^2(h)\, dh\right]^{-3/5} \tag{4.62}$$

This result was first obtained in 1966 by David Fried. The parameter r_0 has therefore come to be known as the ***Fried parameter***. In deriving it we used the near-field approximation, which is generally quite good for astronomical observations. One can easily show that the result holds for the combined structure function

$$\begin{aligned} D(r) &\equiv D_\varphi(r) + D_\chi(r) \\ &= 6.88\left(r / r_0\right)^{5/3} \end{aligned} \tag{4.63}$$

without any approximation. The near field approximation is therefore equivalent to the statement $D_\varphi(r) >> D_\chi(r)$.

From (4.56) and (4.61) it can be seen that the RMS phase difference between two points on the wavefront grows with separation in proportion to $r^{5/6}$ and has the value 2.62 radians when $r = r_0$. Thus r_0 represents a characteristic scale above which the phase fluctuations are large enough to cause significant blurring of the image.

From equation (4.62) it is evident that the Fried parameter depends on wavelength

$$r_0 \propto \lambda^{6/5}. \tag{4.64}$$

At typical telescope sites, $r_0 \approx 0.1\,\mathrm{m}$ at a wavelength $\lambda = 0.5\,\mathrm{nm}$ and increases to about 0.5 m at a wavelength of 2 um.

4.3.5 Long-exposure MTF and PSF

Lets now apply the tools that we have learned to estimate the MTF and PSF for long-exposure images taken through the atmosphere. As in section 4.2.1 we write the amplitude of the electromagnetic wave in the entrance pupil as

$$a(\mathbf{r}) = a_0 e^{i\varphi(\mathbf{r})}\Pi(r/D) \tag{4.65}$$

Now, from (4.29), the instantaneous MTF is just the convolution

$$\begin{aligned} M(\mathbf{r},t) &= (a * a^*)(\mathbf{r}) \\ &= a_0^2 \int e^{i[\varphi(\mathbf{x})-\varphi(\mathbf{r}-\mathbf{x})]}\Pi\left(|\mathbf{x}|/D\right)\Pi\left(|\mathbf{r}-\mathbf{x}|/D\right)d^2\mathbf{x} \end{aligned} \tag{4.66}$$

The long-exposure MTF is obtained by averaging,

$$M(\mathbf{r}) = a_0^2 \int\left\langle e^{i[\varphi(\mathbf{x})-\varphi(\mathbf{r}-\mathbf{x})]}\right\rangle\Pi\left(|\mathbf{x}|/D\right)\Pi\left(|\mathbf{r}-\mathbf{x}|/D\right)d^2\mathbf{x} \tag{4.67}$$

By virtue of the central limit theorem, the individual fluctuations $\phi(\mathbf{x})$ have a distribution function that is essentially Gaussian with zero mean. We may therefore make use of the following relation, valid for any Gaussian random variable x,

$$\begin{aligned} \langle \exp(x)\rangle &= \exp\left(\langle x\rangle + \frac{1}{2}\mathrm{Var}(x)\right) \\ &= \exp\left(\langle x\rangle + \frac{1}{2}\left[\langle x^2\rangle - \langle x\rangle^2\right]\right) \end{aligned} \tag{4.68}$$

Thus,

$$\begin{aligned} \left\langle e^{i[\varphi(\mathbf{x})-\varphi(\mathbf{r}-\mathbf{x})]}\right\rangle &= e^{-\frac{1}{2}\left\langle[\varphi(\mathbf{x})-\varphi(\mathbf{r}-\mathbf{x})]^2\right\rangle} \\ &= e^{-\frac{1}{2}D_\varphi(r)} \end{aligned} \tag{4.69}$$

and (4.67) becomes

$$M(r) = e^{-\frac{1}{2}D_\phi(r)}M_0(r) \tag{4.70}$$

where $M_0(r)$ is the diffraction-limited MTF (4.31) shown in Figure 4.2. Substituting Fried's relation gives

$$M(r) = e^{-3.44(r/r_0)^{5/3}}M_0(r)\,. \tag{4.71}$$

From this we see that, for $r_0 \ll D$, atmospheric seeing introduces an attenuation factor into the MTF that effectively reduces the resolution to that of a diffraction-limited telescope of approximate diameter r_0.

The PSF can now be determined as it is proportional to the Fourier transform of the MTF. There is no exact analytic form for this, but we can easily get an approximate solution. The MTF (4.71) has a form very similar to a Gaussian function of *r*, but with a 5/3 power instead of the normal square. Since the Fourier transform of a Gaussian is also a Gaussian, we conclude that the PSF will have an approximately Gaussian form, with a characteristic angular width of order $\theta \simeq \lambda / r_0$. Since $r_0 \propto \lambda^{6/5}$, $\theta \propto \lambda^{-1/5}$ and the image quality improves as the wavelength increases (contrary to the result for a diffraction limited telescope). At a sufficiently large wavelength $r_0 \simeq D$ and the telescope will become diffraction limited.

4.3.6 Image quality

The sharpest images that a telescope can produce occur when the resolution is limited only by diffraction. In practice, the resolution of most telescopes is limited by aberrations, atmospheric seeing, etc. There are a number of ways of expressing the resolution of a telescope, both in an absolute sense and relative to the diffraction limit.

One could of course give a complete description of the PSF or MTF of the telescope, but more often a single number is desired which describes the image quality. A commonly used parameter is the full-width at half maximum (***FWHM***) of the PSF. This is the effective diameter of the isophote whose intensity is half that of the central intensity of the PSF, and corresponds to our intuitive notion of the “width” of the PSF. The FWHM is often used to describe the seeing at a telescope site, since the seeing is usually the primary factor determining the image quality for ground-based optical telescopes.

Another parameter is the 50% encircled-energy diameter (***EED***). This is the diameter of the smallest circle that contains 50% of the light. Similarly, one can define the 75% EED, the 90% EED, etc.

A very useful parameter describing images that are nearly diffraction limited is the ***Strehl ratio*** (*S*). It is defined as the ratio of the central intensity of the PSF to the central intensity of a perfect diffraction-limited PSF (at the same wavelength). Clearly $S \leq 1$.

4.3.7 Zernike functions

A useful way of describing wavefront aberrations is by an expansion in terms of orthogonal functions. For circular apertures, the Zernike polynomials are commonly used as they are defined within a circle and are closely related to the polynomial expansion of Hamilton and the Seidel aberrations. The Zernike functions are defined as

$$Z_n^m(r,\phi) = R_n^{|m|}(r)e^{im\phi}\,, \tag{4.72}$$

where $n = 0,1,2,\cdots$, $m = 0,\pm1,\pm2,\cdots,\pm n$ and $n-m$ is an odd number. $R_n^m(r)$ are the Zernike polynomials,

$$R_n^m = \sum_{k=0}^{(n-m)/2} \frac{(-1)^k (n-k)!}{k!((n+m)/2-k)!((n-m)/2-k)!} r^{n-2k}\,. \tag{4.73}$$

These functions are defined on the unit disk ($r \le 1$) and are orthogonal in the sense that

$$\int_0^{2\pi}\int_0^1 Z_n^m Z_{n'}^{m'} r dr d\phi \propto \delta_{mm'}\delta_{nn'} \tag{4.74}$$

The parameter *n* is called the radial order and *m* the azimuthal order.

A listing and interactive display of the 36 lowest-order Zernike functions can be found on-line at http://wyant.opt-sci.arizona.edu/zernikes/zernikes.htm. The first 12 Zernike polynomials are listed in Table 4.1. The image effects of the first 11 Zernike functions are described in Table 4.2.

Table 4.1 Low-order Zernike polynomials

n\m	0	1	2	3	4	5
0	1					
1		r				
2	$2r^2-1$		r^2			
3		$3r^3-2r$		r^3		
4	$6r^4-6r^2+1$		$4r^4-3r^2$		r^4	
5		$10r^5-12r^3+3r$		$5r^5-4r^3$		r^5

Table 4.2 Low-order Zernike functions

(n,m)	Name	Effect
$(0,0)$	Piston	None, phase change only
$(1,\pm1)$	Tip-tilt	Shift of image position

(2,0)	Defocus	Focus shift
(2,±2)	Astigmatism	3rd-order astigmatism
(3,±1)	Lateral coma	3rd-order coma
(3,±3)	Triangular coma	Aberration with 3-fold symmetry
(4,0)	Spherical aberration	3rd-order spherical aberration

4.4 Adaptive optics

It has recently become possible to largely correct the phase errors introduced by the atmosphere, at least at near-infrared wavelengths. This is accomplished by inserting a deformable mirror into the beam, at or near a pupil, prior to the final focus. The surface shape of this mirror can be altered rapidly by means of electrical signals. A wavefront sensor determines the residual phase error by analyzing light from a reference star, and a feedback system drives the deformable mirror in such a way as to cancel the phase error. This technique is called ***adaptive optics*** (AO) and is now used at many major optical observatories (Figure 4.3).

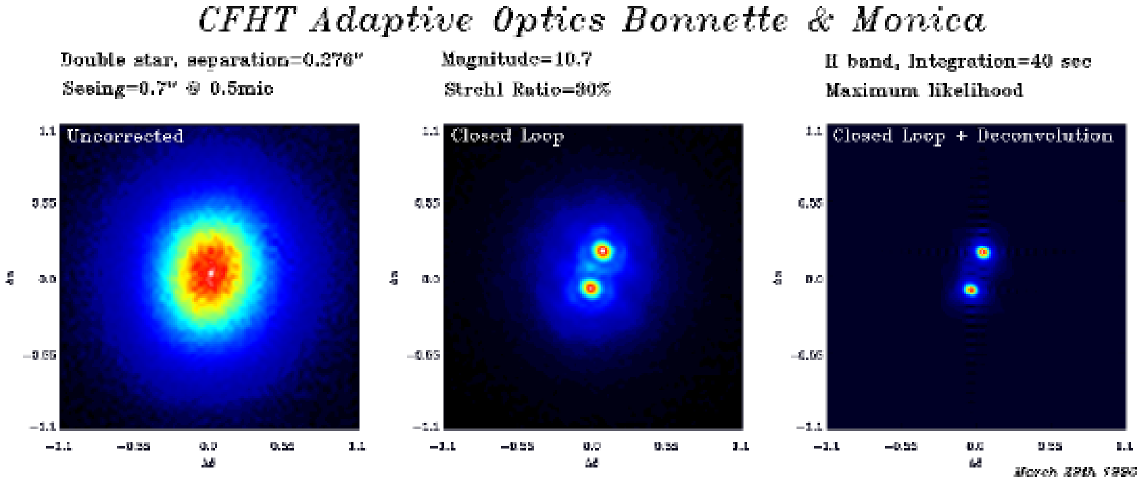


Figure 4.3 A binary star imaged under normal atmospheric conditions (left), is unresolved. With adaptive optics, the double star is clearly resolved (center). Image processing further improves the image (right). Figure courtesy of the Canada-France-Hawaii Telescope Corp.

4.4.1 Adaptive optics systems

A typical AO system has at least the following components: a collimator, tip-tilt mirror, deformable mirror, camera, detector, wavefront sensor and control system. The collimator receives diverging light from a focus, and forms a parallel beam. Located in this beam is a plane mirror that can pivot rapidly about two orthogonal axes. This ***tip-tilt mirror*** removes wavefront tilt errors (corresponding to the $Z_1^{\pm 1}$ Zernike terms) that would be difficult for the deformable mirror to correct. Following the tip-tilt mirror, the light strikes a nearly-plane ***deformable mirror*** (Figure 4.4) This mirror is usually located at a pupil position (image of the primary mirror), or at a position near it that is optically conjugate to (ie. an image of) the main atmospheric turbulence layer. The surface of this mirror is divided into zones called ***subapertures***, each of which can be deflect a distance of up to several wavelengths by application of an electrical voltage. After reflection from the deformable mirror, the light passes through a beam splitter and then to a ***camera*** mirror or lens which produces an image on the detector. A small portion of the light is reflected by the beam splitter to a ***wavefront sensor***. The function of this device is to analyze the residual wavefront errors (after correction by the deformable mirror) and to generate an error signal which is amplified and applied to the deformable mirror in a null-seeking feedback loop. Practical systems include a number of other elements such a atmospheric dispersion correctors, filters, folding mirrors, etc and are quite complex (Figure 4.5).

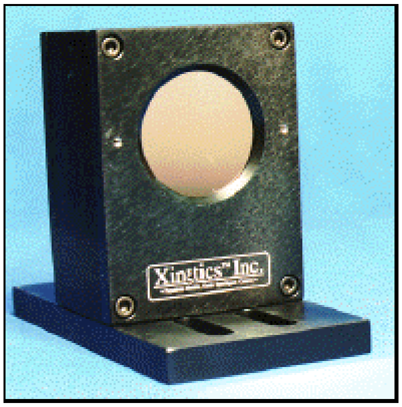


Figure 4.4 A commercially-available deformable mirror

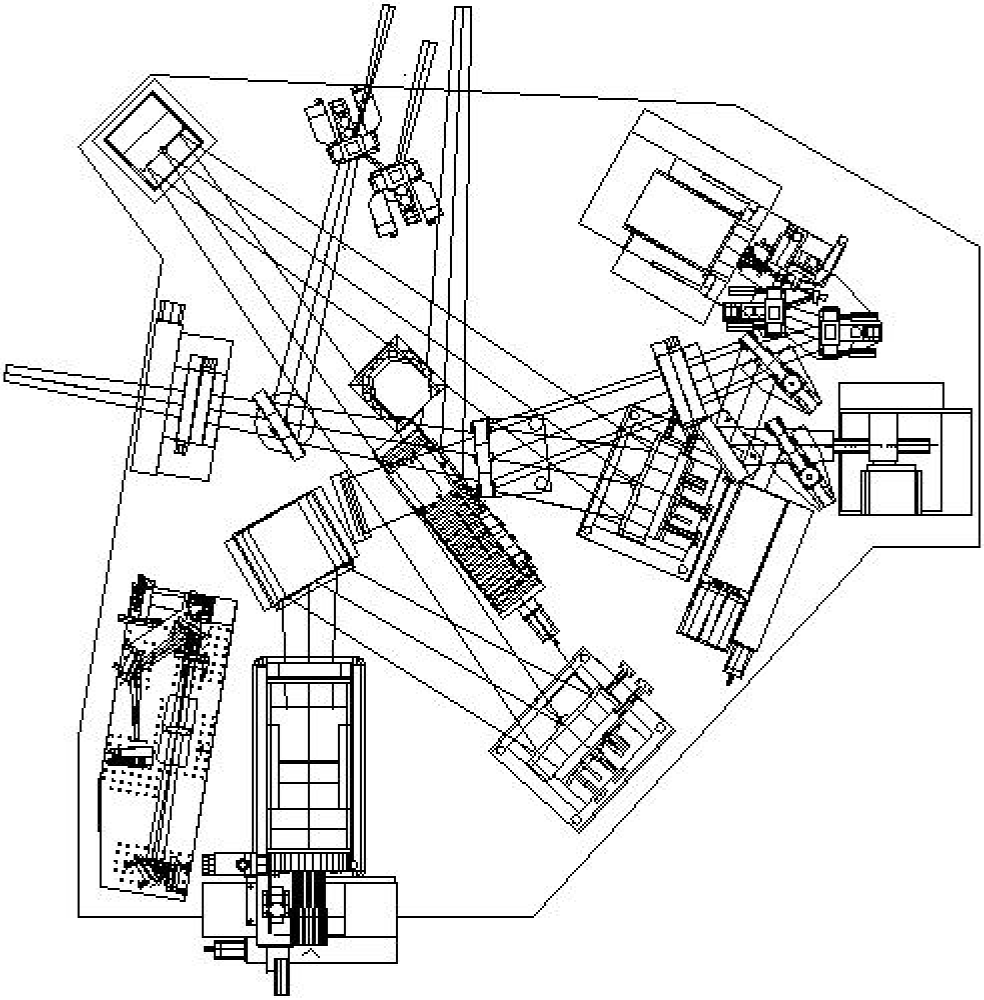


Figure 4.5 Optical layout of an adaptive optics system used by the Keck telescopes.

4.4.2 Wavefront-sensors

Presently, two types of wavefront sensors are in common use, Shack-Hartman sensors and curvature sensors. The ***Shack-Hartman sensor*** employs an array of microlenses, located at a pupil, to form multiple images of a star on a CCD. Each microlens produces a single image of the reference star on the CCD (Figure 4.6). The position of the centroid of this image is directly related to the slope of the wavefront at the location of the microlens. By measuring the centroids of the images produced by all the microlenses, one obtains a map of the wavefront slopes (derivatives of the wavefront phase error) at all locations in the pupil. The wavefront error at any point may then be determined by integration.

The curvature sensor makes use of the fact that the intensity distribution in a defocused star image is proportional to the local mean curvature of the wavefront phase error $I(\mathbf{x}) \propto \nabla^2\phi(\mathbf{x})$. In a typical curvature sensor, two defocuses images of the reference star are formed on a CCD detector, one a certain distance before the focus and the other the same distance behind the focus. The difference between this images is calculated from which a map of the wavefront curvature over the pupil is determined. Boundary conditions are obtained from the positions of the edge of the image, and the map is integrated to obtain a map of the wavefront error.

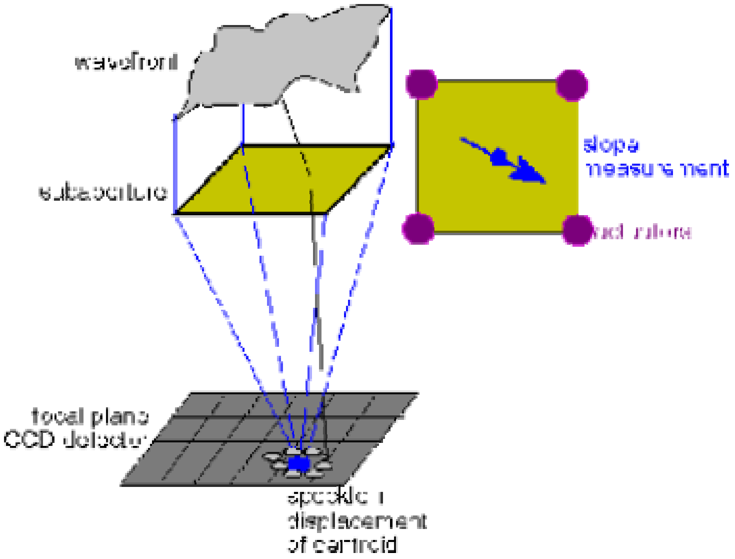


Figure 4.6 Principle of the Shack-Hartman wavefront sensor. The illustration shows how a single microlens produces an image of the reference star on a CCD detector. The position of the image gives the local mean wavefront slope.

4.4.3 Adaptive optics performance

In order for the AO system to provide good correction of the wavefront, the size of the subapertures should not be much larger than the Fried parameter r_0, otherwise the atmospheric phase errors will have significant power on scales smaller than the subaperture size and will not be fully corrected. Of course, the resolution of the wavefront sensor must be matched to the subaperture size of the deformable mirror. It is also necessary to collect sufficient photons from the reference star, over an area in the pupil of roughly r_0^2, in order to accurately determine the phase error.

Since $r_0 \propto \lambda^{6/5}$, there will be a wavelength below which r_0 becomes smaller than the subaperture size and the AO performance becomes severely degraded. As r_0 decreases, fewer photons are detected and the integration time must also be shorter (because for a given wind speed *v*, the timescale for significant phase change is r_0 / v). For these reasons, adaptive optics works best in the infrared, typically J, H and K bands, and gives little if any performance gain in the optical with present technology.

Another problem occurs if the object of interest is not sufficiently close to a suitably bright reference star. Then, the light arriving from the object traces a different path through the atmosphere than does the light from the reference star and the phase errors calculated for the star are no longer suitable for the object. If the turbulence arises at a height h above the telescope, and the deformable mirror is located at a pupil (conjugate to the primary mirror), the ***isoplanatic angle*** (the angle within which the phase differences are small) is

$$\theta_{iso} = r_0 / h. \tag{4.75}$$

Thus, good adaptive optics performance can only be expected within an angle θ_{iso} of the reference star. This situation can be improved by locating the deformable mirror conjugate to the turbulence layer, but if the turbulence is distributed over a range of altitudes this will be less effective. A better solution is to have several deformable mirrors, each conjugate to a different altitude. A corresponding number of reference stars and wavefront sensors are also required, in order to assign the appropriate wavefront correction to each mirror. This technique is called ***multi-conjugate adaptive optics*** (MCAO) or ***atmospheric tomography***. It is a subject of current research and has not yet been demonstrated.

A lack of suitable reference stars has motivated the development of ***laser beacons***. A powerful laser is used to illuminate a region of the atmosphere creating an artificial star. Light scattered back from this region is used as a reference to determine the atmospheric phase error. Either Rayleigh scattering or flourescence of the atmospheric sodium layer can be used. The latter is most effective as the light returned from the 90km high sodium layer is more nearly parallel to the light from distant astronomical sources. Laser beacons offer greater flexibility, but they have not achieved as high a degree of performance as AO systems employing natural guide stars.

Fig shows an image from the Gemini telescope with and without adaptive optics. The gain in resolution and sensitivity provided by AO is very impressive. An interesting animation showing a stellar image with no correction, tip-tilt correction only, and then full high-order adaptive optics correction, can be found online at http://www.aoainc.com/technologies/adaptiveandmicrooptics/aos.html.

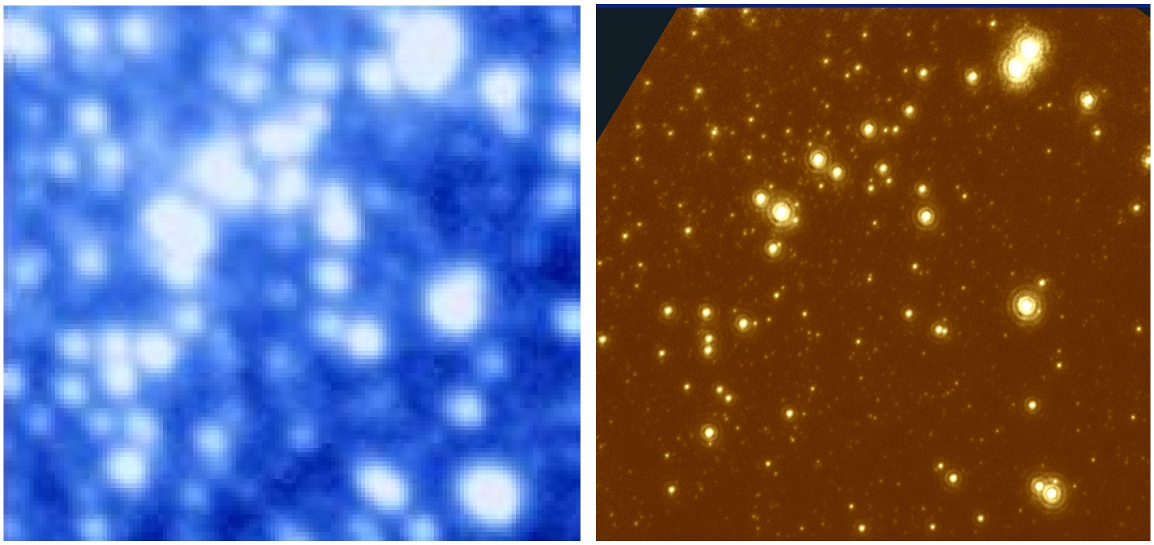


Figure 4.7 Comparison of an image in natural atmospheric seeing (left, 0.8" FWHM) and the same field imaged with an adaptive optics system (right). Both images were obtained with the Gemini North 8-meter telescope.

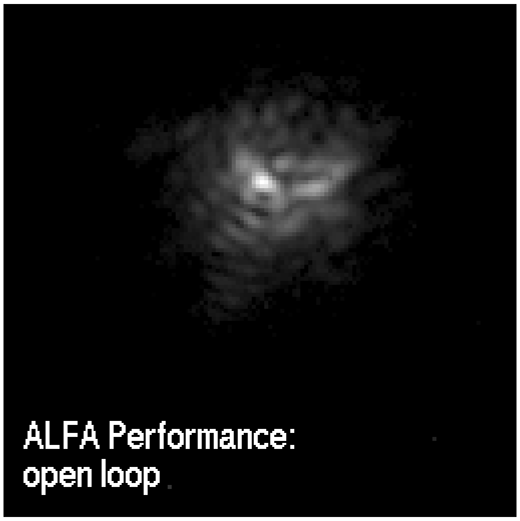


Figure 4.8 Animation showing the effect of an adaptive optics system. The sequence shows 100 ms images of the 8th magnitude star Omega-Cass, taken with the 3.5m Max-Planck telescope at Calar Alto. The first 100 images show the the star image in natural seeing, the next 100 with tip-tilt correction, and the last 100 with full adaptive-optics correction.

4.5 Interferometry

An ***interferometer*** is a telescope that employs separate apertures, but combines the radiation received be each in a way that preserves the phase information. An early example is the optical interferometer used by Michelson to measure stellar diameters. Michelson attached two small auxiliary mirrors to the 1.5-metre telescope at Mt. Wilson and used them to direct light from a star to the primary mirror that brought the light to a common focus. For a small separation between the mirrors, interference fringes were observed. These disappeared at sufficiently large separation for a number of nearby stars of large angular diameter. From a graph of fringe visibility vs mirror separation, Michelson was able to infer the stellar diameter.

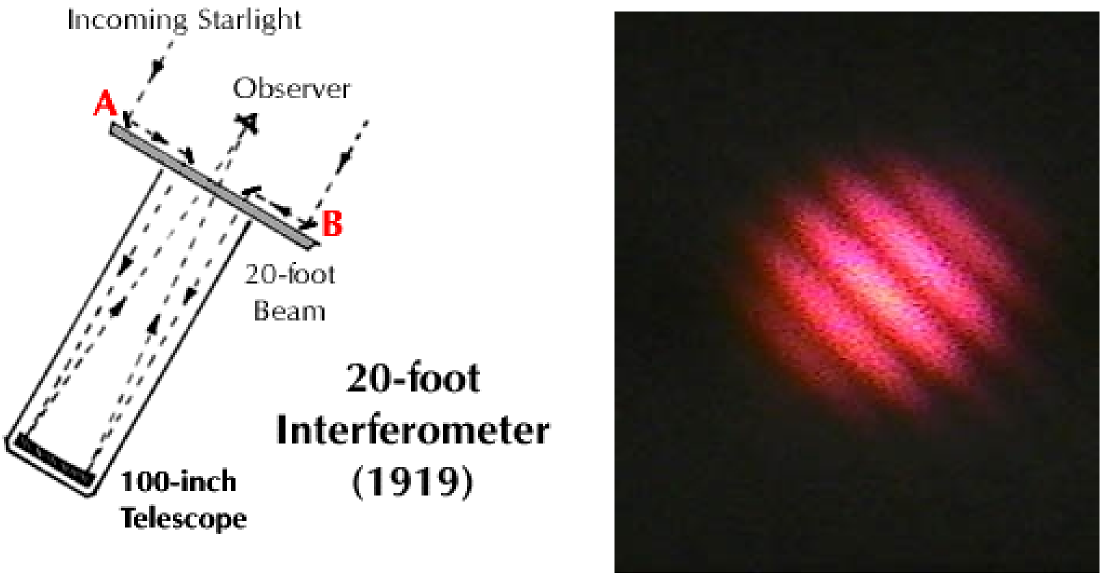


Figure 4.9. Principle of the Michelson stellar interferometer.

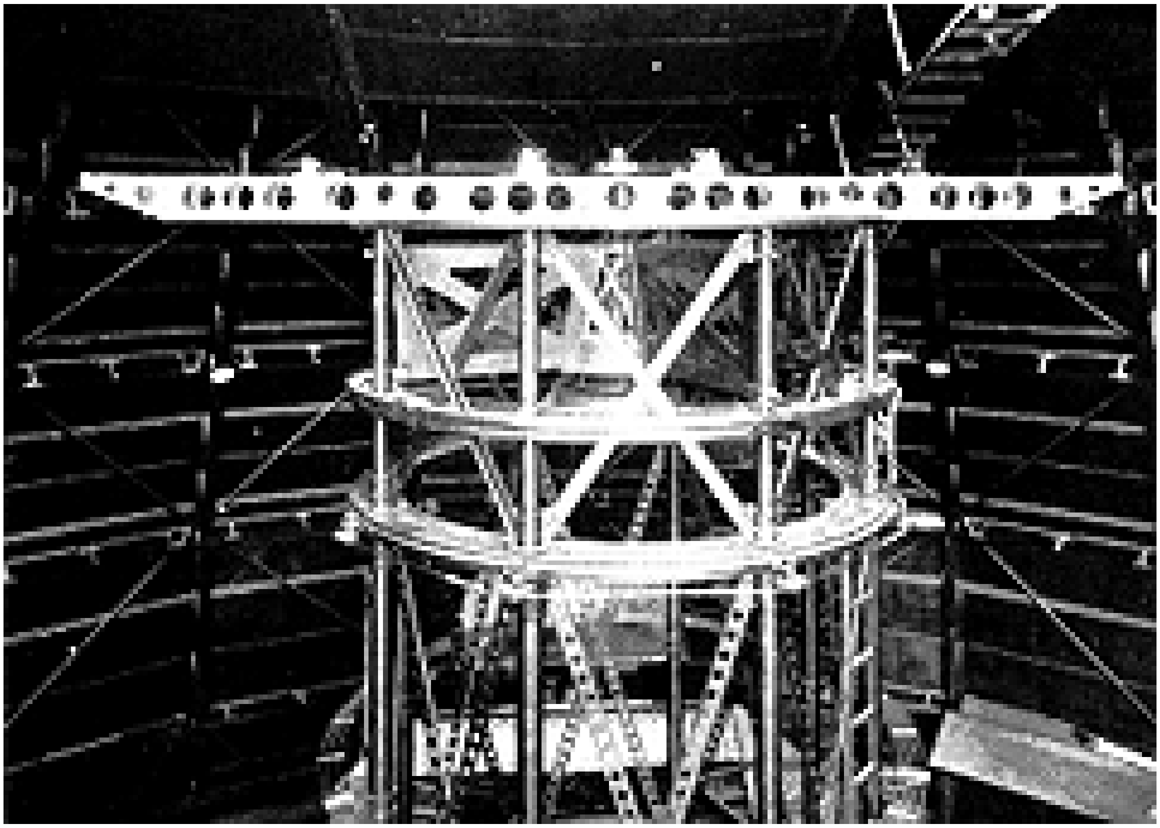
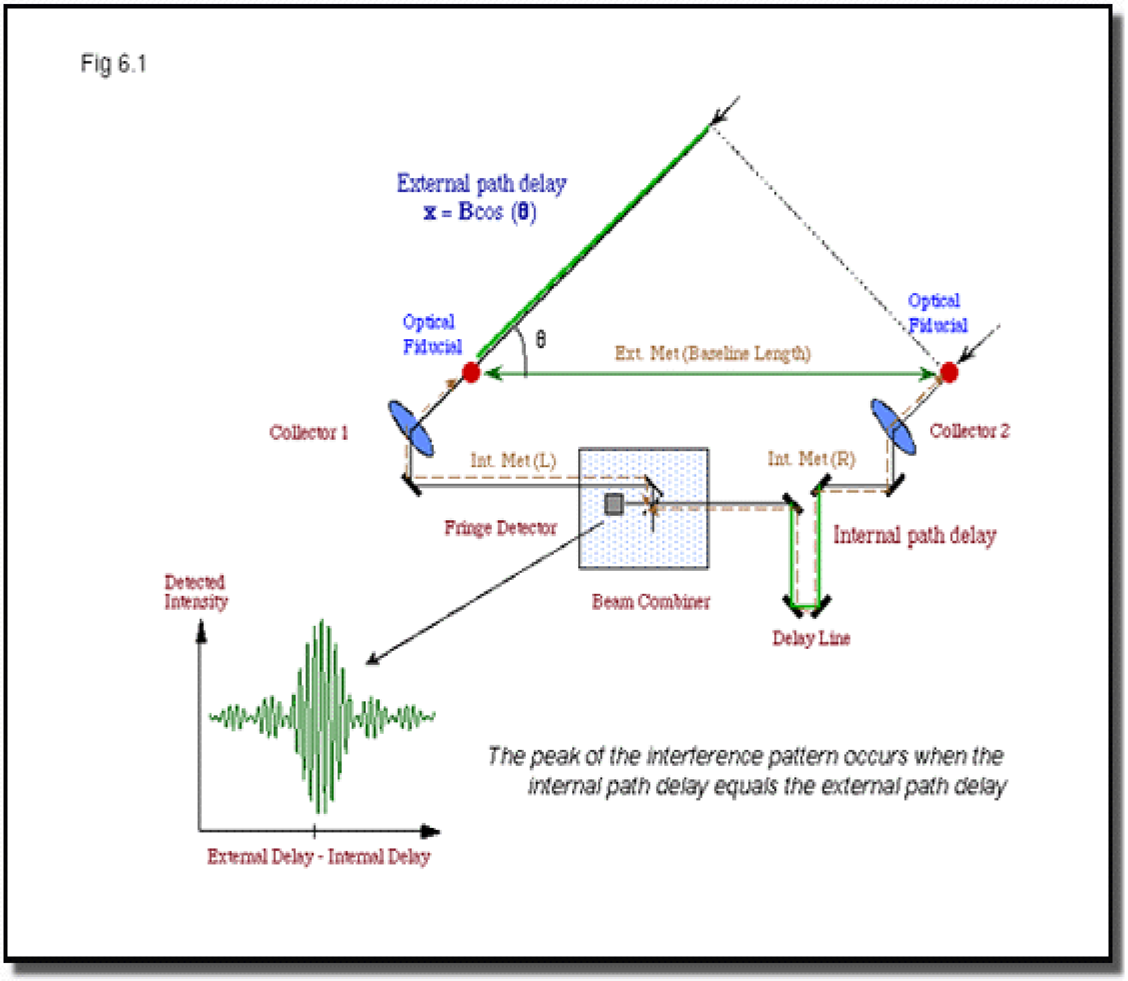


Figure 4.10. Original Michelson interferometer on the Mount Wilson 100-inch Hooker telescope.

Interferometry was popularized by radio astronomers, who were at a disadvantage compared to optical astronomers when in came to resolution. Recall that the angular size of the PSF of a telescope of diameter D is approximately λ / D. Because radio waves have wavelength typically six orders of magnitude larger than optical radiation, the resolution of single-dish radio telescopes is relatively poor. To overcome this limitation, radio astronomers combine the signals from separate antennae whose separation can be much larger than their individual diameters.



The peak of the interference pattern occurs when the internal path delay equals the external path delay

Figure 4.11. Schematic diagram of a two-element interferometer.

In radio interferometry, the signals (specifically, the intermediate-frequency signals produced by mixing a local oscillator signal with the radio-frequency signal received by the antenna) from two antenna are multiplied together. The result is an oscillating signal whose phase is proportional to the difference in optical path length from an incoming plane wave to the two antennae. As the Earth rotates, this difference varies and the signal oscillates.

Let the amplitude of the radiation field, measured in a plane perpendicular to the direction of propagation, be $a(\mathbf{q}_1,t)$ and $a(\mathbf{q}_2,t+\tau)$. Here τ represents the time difference between the two signals reaching the correlator, due to the different propagation distances to the two antenna and internal delays in the transmission of the electrical signal to the correlator. The output of the correlator is a signal proportional to the ***autocorrelation function***

$$\Gamma(\mathbf{q}_1,\mathbf{q}_2,\tau)=\left\langle a(\mathbf{q}_1,t)a^*(\mathbf{q}_2,t+\tau)\right\rangle, \qquad (4.76)$$

also called the ***mutual coherence function*** in optics. Here the bracket denote a time average that is long compared to the reciprocal of the frequency of the radiation, but short compared to variations in τ caused by the Earth's rotation. For astrophysical sources (which are incoherent and very distant) the mutual coherence function depends only on the coordinate difference $\mathbf{q}_1 - \mathbf{q}_2 = \lambda \mathbf{u}$, where $\mathbf{u} = (u, \upsilon)$ is the vector corresponding to the ***projected baseline*** (the line joining the two antenna, projected onto a plane perpendicular to the direction of the source) divided by the wavelength. Therefore,

$$\Gamma(\mathbf{q}_1, \mathbf{q}_2, \tau) = \Gamma(\mathbf{u}, \tau) = \left\langle a(\mathbf{q}_1, t) a^*(\mathbf{q}_1 + \lambda \mathbf{u}, t + \tau) \right\rangle \tag{4.77}$$

where the brackets now include a spatial average. This function has a sinusoidal dependence on τ, which corresponds to fringes, just as for the Michaelson interferometer. The fringe amplitude is $\Gamma(\mathbf{u}) \equiv \Gamma(\mathbf{u}, 0)$. According to the Weiner-Khinchin theorem, the Fourier transform of an autocorrelation function is the Power spectrum of the signal. In other words, the spatial Fourier transform of $\Gamma(\mathbf{u})$ is proportional to the squared modulus of $\tilde{a}(\mathbf{x})$, the Fourier transform of the amplitude in the pupil. From Fraunhoffer diffraction theory, the amplitude in the pupil is the Fourier transform of the angular distribution of the amplitude distribution in the source. Therefore, the Fourier transform of $\Gamma(\mathbf{u})$ is proportional to the intensity distribution in the source

$$I(\mathbf{x}) \propto \int \Gamma(\mathbf{u}) e^{-2\pi i \mathbf{u} \cdot \mathbf{x}} d^2 u \,. \tag{4.78}$$

Each pair of antennae gives a measure of $\Gamma(\mathbf{u})$ at a different point in the u-v plane. Now, as the Earth rotates, the projected baseline changes because the antennae are fixed to the Earth. Thus, each antenna pair sweeps out a curve in the u-v plane, along which measurements of $\Gamma(\mathbf{u})$ are recorded. After a sufficiently long time interval, we know $\Gamma(\mathbf{u})$ at enough points in the u-v plane that the image of the source can be reconstructed. This technique is called ***aperture synthesis*** and is the principle of operation of all modern radio interferometers, such as the Very Large Array (VLA) in New Mexico (Figure 4.12).

For an array of *N* antennae, the number of baselines is $N(N-1)/2$, so the quality of the reconstructed image increases rapidly with the number of antennae. The resolution of the image is determined by the length of the longest baselines D_{max} and is of order $\theta \simeq \lambda / D_{max}$. Long-baseline interferometers, like the VLA, achieve a resolution comparable to that of the best optical telescopes (Figure 4.13). In very-long-baseline interferometry (***VLBI***), the separation between antennae is comparable to the radius of the Earth, so a resolution of a few milli-arcseconds or better can be achieved.

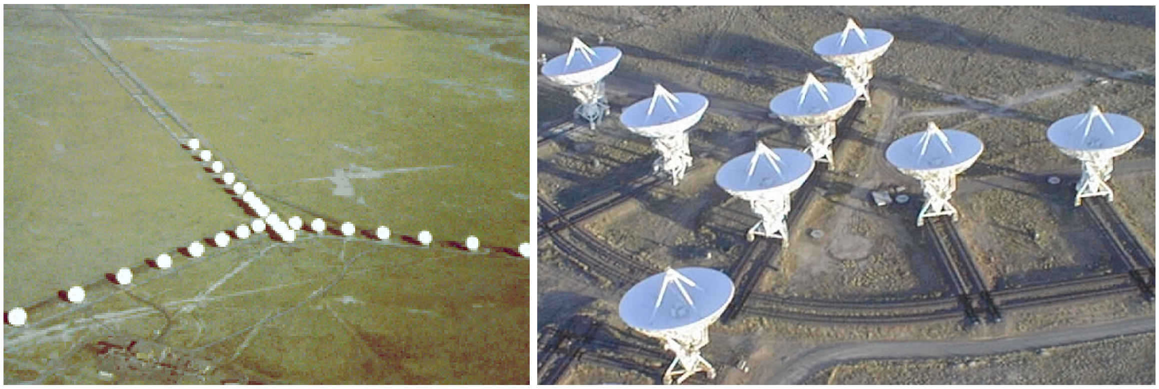


Figure 4.12 The Very Large Array (left) and a closeup of the central region (right). The array, located near Socorro, New Mexico, has 27 antenna, each 30 metres in diameter, deployed along a Y-shaped track. The array has four configurations (A, B, C and D) corresponding to increasing antenna separations and resolution.

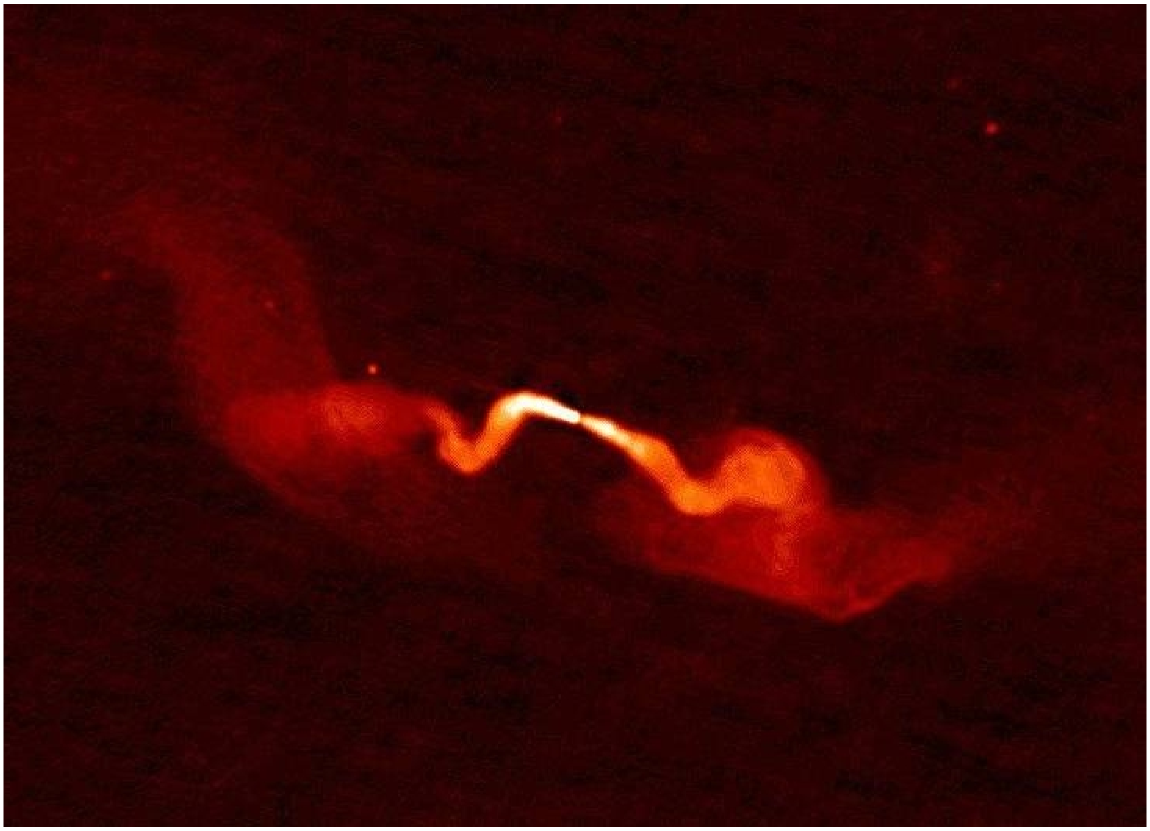


Figure 4.13 The radio galaxy 3C31, imaged by the VLA. The motion of this galaxy within a cluster causes its jets to be swept back by the pressure of the intergalactic medium.

Exercises

4.1 The seeing at the Canada-France-Hawaii Telescope is typically 0.8" FWHM in the V band. At approximately what wavelengths would you expect the 3.6-metre telescope to give diffraction-limited images?

4.2 Why do you think that the wavefront sensor (and beamsplitter) in an adaptive optics system is placed after the deformable mirror, rather than before it?

4.3 The Terrestrial Planet Finder (TPF) is a proposed space telescope that would be able to image an Earth-like planets orbiting solar type. One concept for TPF is an optical interferometer consisting of four diffraction-limited telescopes operating in the 3-30 um wavelength range with separations between telescopes in the range 25-1000 m. Estimate the maximum resolution (the PSF FWHM) that could be achieved by TPF. How does this compare with the angular size of the Earth seen at a distance of 10 pc? How does it compare with the angular separation between the Earth and Sun as seen from the same distance?

4.4 Prove (4.68).

4.5 By numerically evaluating the Hankel Transform of the MTF (4.71), and assuming that $r_0 \ll D$, compute the atmospheric PSF. Plot the logarithm of intensity vs radius. How does it compare with a Gaussian function (which is a parabola in this type of plot)?

5 Spectroscopy

The remarkable progress in astrophysics over the last century owes much to the development of spectroscopic techniques. Spectra can provide a wealth of information about the physical conditions, chemical composition, and state of motion of a source. Because the flux or intensity must be measured as a function of wavelength, more photons are required than for simple imaging. Spectroscopy is therefore inherently more difficult than imaging, particularly for faint sources. Increasingly-sophisticated and sensitive instruments have been developed to maximize the efficiency of spectroscopic observations. This chapter is primarily concerned with optical spectroscopy, although much of it is applicable to other wavelengths, particularly the infrared and ultraviolet. Spectroscopic techniques used for coherent detection (radio astronomy) are generally rather different and will be discussed elsewhere.

5.1 Spectrographs

In spectroscopy, we seek to determine the functions $F_\nu, F_\lambda, I_\nu,$ or I_λ over some range of frequency or wavelength. Of particular interest is the ***spectral resolution***, ($\Delta\nu$ or $\Delta\lambda$) which describes the smallest frequency or wavelength interval over which differences in the flux or intensity can be discerned. For example, two spectral lines separated by a wavelength interval less than $\Delta\lambda$ are not individually distinguished but instead appear blended. Closely related is the spectral ***resolving power***

$$R = \frac{\lambda}{\Delta\lambda} = \frac{\nu}{\Delta\nu}. \qquad (5.1)$$

Optical spectroscopy can be divided into low $(R < 100)$, medium $(100 < R < 5000)$ and high $(R > 5000)$ resolution categories, each requiring rather different instrumentation and techniques.

The basic instrument for optical spectroscopy is the ***grating spectrograph***, illustrated in Figure 5.1. A spectrograph has five essential components: a slit, collimator, diffraction grating, camera and detector. The typical layout is illustrated in Figure 5.2. The ***slit***, located at the telescope focus, allows light from a star or galaxy to enter the spectrograph while blocking light from other objects and the night sky. Light passing through the slit strikes the ***collimator*** that focuses it into a parallel beam. This light then strikes the diffraction ***grating*** that disperses it. Different wavelengths are diffracted by different angles, so the diffracted beam has a wavelength-dependent angular spread in the direction perpendicular to the slit. The light is then re-imaged onto a ***detector*** by means of a ***camera*** that may consist of mirrors, lenses or both. Because of the angular dispersion created by the grating, different wavelengths focus to different points on the detector, so the spectrum of an object is spread along a line on the detector, perpendicular to the direction of the slit. In other words, a spectrograph is a kind of imaging system that produces an image of the slit on the detector. However, the position of this image is wavelength dependent.

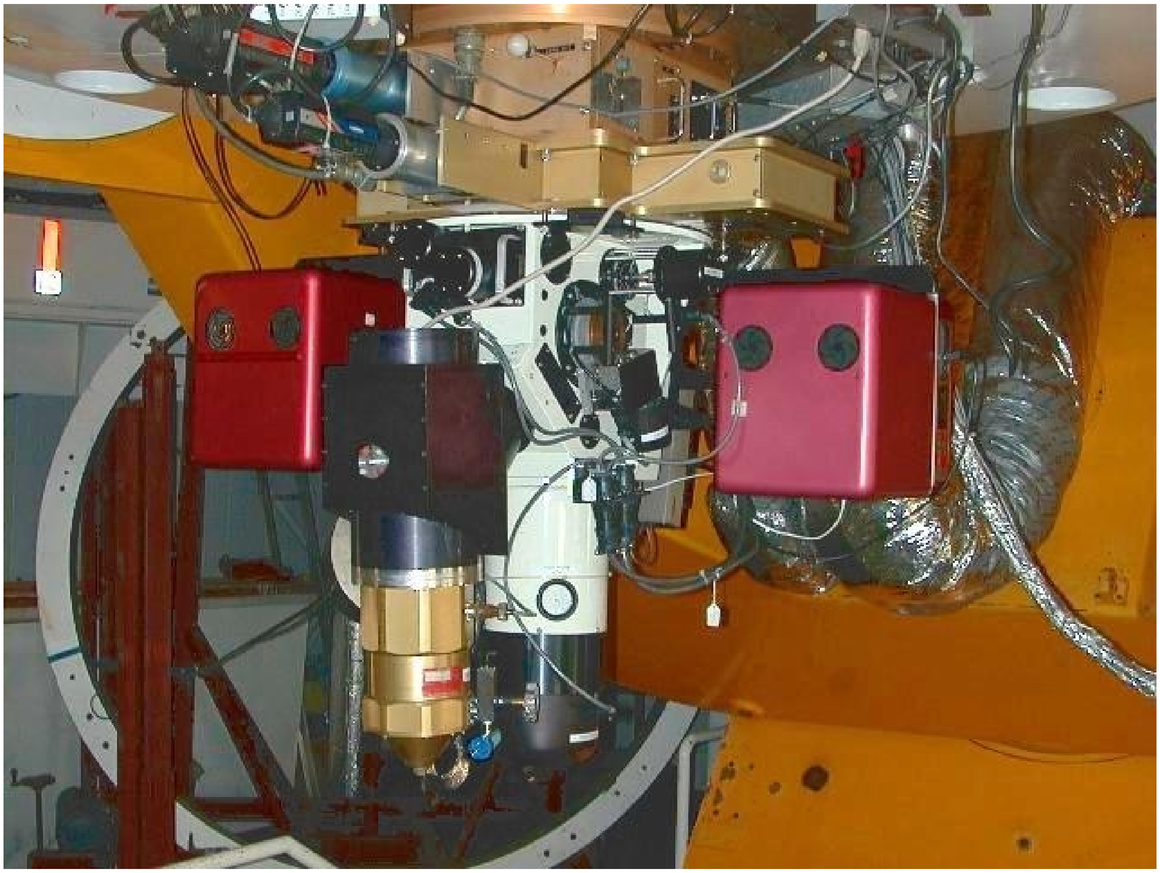
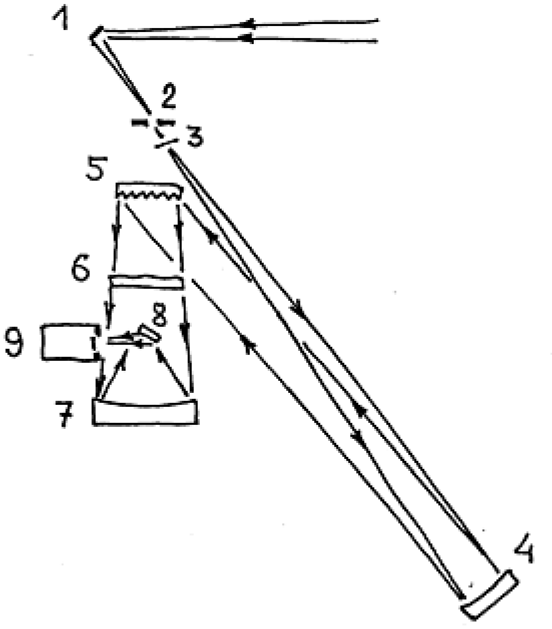


Figure 5.1 The Boller and Chivens CCD Spectrograph (the white structure) used on the MDM Observatory 2.4-m and 1.3-m telescopes. It is a conventional optical grating spectrograph operating in the 3200 - 9500 Å region. Seven gratings are available providing spectral resolutions between 550 and 9100. The slit width is continuously adjustable between 0.5 and 13 arcsec. The detector is a Loral 1200 × 800 pixel CCD (the gold coloured cylinder). Lamps mounted above the spectrograph are used for wavelength calibration and flat fielding. At the lowest dispersion, the peak total system quantum efficiency is about 19% at 5200 Å. An intensified camera records the image provided by the telescope on the polished entrance slit jaws and for an integration time of a few seconds, point sources with magnitudes just below the POSS limit can be seen over most of the 5 arcmin field of view at the 2.4-m telescope. Low resolution spectra of objects as faint as 20.5 mag can be obtained at the 2.4-m telescope.



Components:

1. Folding mirror
2. Slit
3. Shutter
4. Collimator
5. Grating
6. Schmidt corrector
7. Camera mirror
8. Folding mirror
9. Detector

Figure 5.2 Optical layout of a typical grating spectrograph. Light from the telescope enters from the upper right on its way to a focus, located at the position of the slit.

5.1.1 Diffraction gratings

A reflective diffraction grating is a mirror on which is ruled a series of parallel grooves with a constant spacing d comparable to the wavelength of light λ. Transmission gratings are also available, which transmit light instead of reflecting it. The groove spacing is usually specified in terms of the number of lines per mm, which can range from a hundred to several thousand. The fundamental relation which describes the behaviour of the grating is the ***grating equation***

$$\sin\theta + \sin\theta' = n\lambda / d \tag{5.2}$$

where θ and θ' are the angles of incidence and reflection and *n* is an integer called the ***order***, $n = 0, \pm 1, \pm 2, \cdots$ This equation is easily derived by requiring that the optical path length for rays diffracted from adjacent grooves be a multiple (*n*) of λ for constructive interference.

The efficiency of a diffraction grating (the fraction of light of a given wavelength diffracted in a given order) varies with wavelength. The wavelength, and order, of peak efficiency depends on the angle of the surfaces of the grooves, called the ***blaze angle***. Most spectrographs have a selection of gratings from which to choose, with various line spacing and blaze wavelengths (wavelengths of peak efficiency).

The wavelength range of the spectrum on the detector can be adjusted by rotating the grating. This changes the angle of incidence, and therefore the angle of diffraction of a particular wavelength. Most spectrographs have a mechanical control, manual or motorized, to allow the central wavelength (the wavelength imaged at the center of the detector) to be adjusted.

Because *n* can take any integer value, diffraction gratings produce an infinite number of spectra, centered at different angles. For typical gratings, most of the light is diffracted into the first order, but the amount of light in second or higher orders can be significant. These higher-order spectra overlap the first-order spectrum which would cause problems if the wavelength range is unrestricted. For example, a wavelength of 400 nm in the second order appears at exactly the same angle (and position on the detector) as a wavelength of 800 nm in the first order spectrum. In order to eliminate contamination of the first-order spectrum by higher orders, blocking filters are used. For example, a filter which passes only wavelengths of 500 nm or more would allow an uncontaminated first-order spectrum with a range of 500-1000 nm, a filter passing 300 nm and greater would allow a first-order spectrum with a range 300-600 nm, etc.

Because of the limited wavelength range of high grating efficiency, and the need for blocking filters, two or more spectrograph configurations (gratings, filters, etc) are often required to cover the desired spectral range.

5.1.2 Spectrograph parameters

From the grating equation we can obtain the ***angular dispersion*** of the grating

$$\frac{d\lambda}{d\theta'} = \frac{d\cos\theta'}{n} \tag{5.3}$$

Dividing by the focal length of the camera f_{cam} gives the ***linear dispersion*** of the spectrograph (the wavelength change per unit length on the detector).

$$\frac{d\lambda}{dx} = \frac{d\cos\theta'}{nf_{cam}} \tag{5.4}$$

The cosine factor introduces a small nonlinearity into the mapping of wavelength onto position in on the detector. The actual position of wavelength λ on the detector is usually determined empirically by reflecting light from spectral calibration lamps into the slit. The light produced by these lamps has numerous emission lines of know wavelengths. From the resulting calibration spectrum, the mapping between wavelength and position on the detector can be determined.

Viewed as an imaging system, the spectrograph has a transverse ***magnification*** M which is the ratio of the width of the image of the slit, on the detector, to the actual width of the slit. It is given by the ratio of camera and collimator focal lengths

$$M = f_{cam} / f_{col} . \tag{5.5}$$

If the slit is uniformly illuminated, each wavelength will give rise to an image of the slit, at a position on the detector determined by the grating equation and camera focal length. The light distribution along the image on the detector is therefore the one-dimensional convolution of the spectrum of the object with the image of the slit. If aberrations and diffraction effects in the spectrograph optics can be ignored, the spectrum is convolved by the function $\Pi(x / wM)$, where w is the physical slit width. This limits the resolution of the spectrograph. From (5.4), we find

$$\begin{aligned} \Delta\lambda \simeq \frac{d\lambda}{dx} wM &= \frac{wMd \cos\theta'}{nf_{cam}} \\ &= \frac{wd \cos\theta'}{nf_{col}}. \end{aligned} \tag{5.6}$$

Normally, the slit is adjusted to have a width comparable to size of the image of a star or galaxy in the telescope focal plane. However, in some applications the slit width is made smaller, in order to improve the resolution. In that case much of the light from the object will be blocked by the slit and sensitivity will be reduced. To overcome this problem, one can use an ***image slicer***. This is a device, placed before the slit, which rearranges the light from an object so that more of it can pass through the slit. Intensity must be conserved in this operation, so the resulting image is wider, occupying a greater length along the slit.

5.1.3 Multi-object spectrographs

With a conventional spectrograph, it is normally only possible to observe one object at a time (unless more than one object lies on the slit). This, coupled with the long exposure times required for faint objects, makes the acquisition of spectra for large samples of galaxies very difficult. ***Multi-object spectrographs*** are designed to take spectra of many objects simultaneously. There are several techniques for doing this. ***Fiber spectrographs***, employ optical fibers to bring the light from the telescope focal plane to the spectrograph slit. The fibers are positioned in the focal plane at the positions of the objects, one fiber per object. In the most recent systems, this is done robotically, by reference to an image of the field. The other ends of the fibers are placed adjacent to each other along the slit. In this way, the spectra of several hundred galaxies may be obtained in a single exposure.

A second technique uses multiple short slits, located at the positions of galaxies in the telescope focal plane (see Figure 5.4). One system, such as the MOS spectrograph at the Canada-France-Hawaii Telescope, uses a laser to cut slits, at the appropriate positions, in thin sheets of aluminum. These are then placed in position before each exposure. A new technique being investigated for the Next-Generation Space Telescope (NGST), uses a programmable micro-mirror array to deflect light from galaxies into the spectrograph.

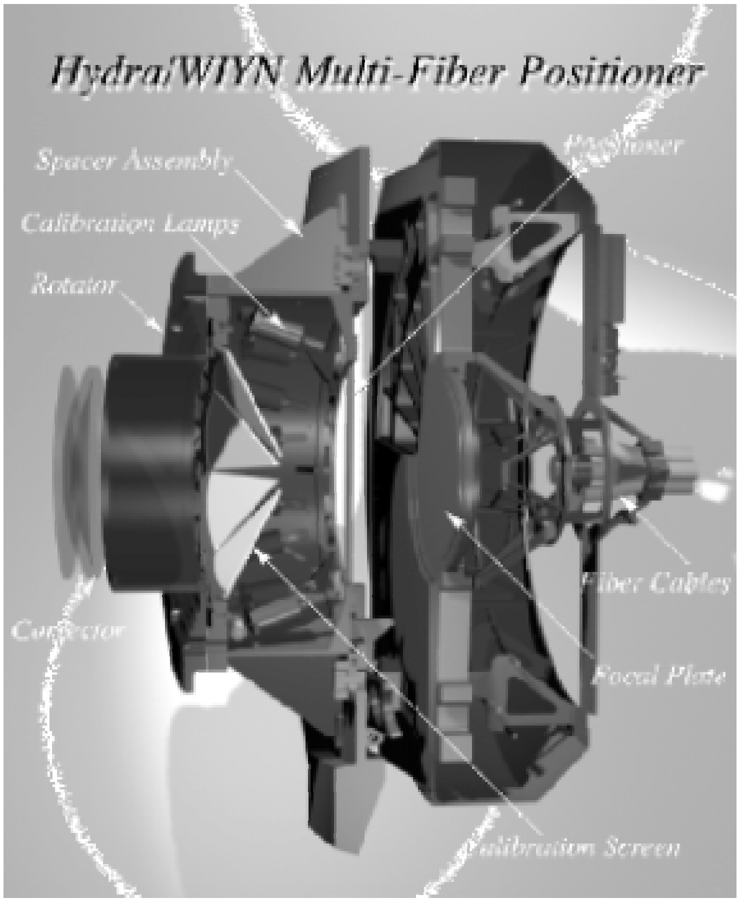


Figure 5.3. The Hydra fiber positioner on the WIYN 3.5 meter telescope allows multi-object spectroscopy of up to 100 objects over a 1 degree field. An optimized, bench mounted spectrograph provides low, moderate, or high dispersion spectra.

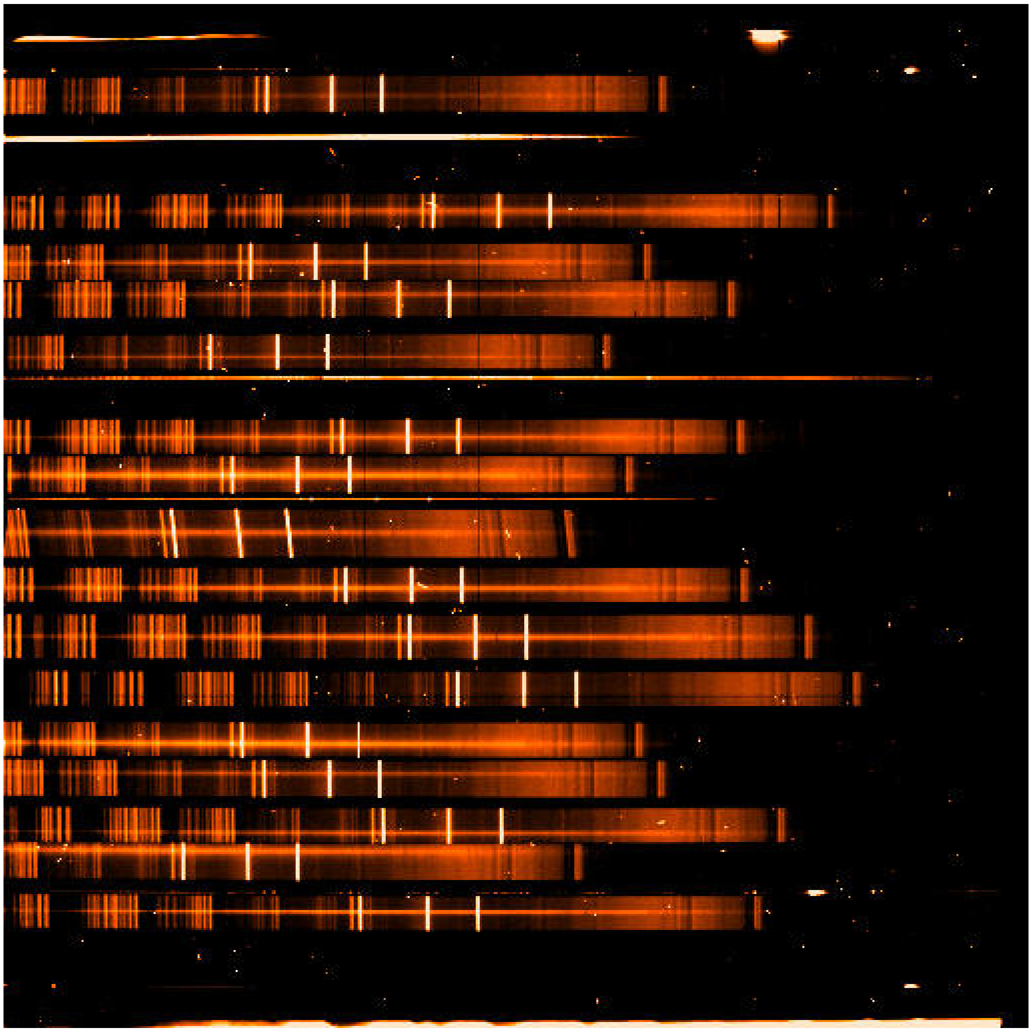


Figure 5.4. Mult-slit spectrum of a field of galaxies taken with the MOS spectrograph of the Italian Telescopio Nationale Galilieo.

5.1.4 Slit-less spectroscopy

Another way of obtaining spectra of many objects simultaneously is to use no slit at all. In a spectrograph designed for imaging, this can be done by simply removing the slit. Another way is to place a thin prism in front of the telescope aperture. The image at the telescope focal plane then becomes a collection of small spectra and no spectrograph is needed at all. Such ***objective prisms*** have been commonly used on small to moderate-sized telescopes to obtain low-resolution spectra of stars for classification purposes. For practical reasons, objective prisms cannot be manufactured for large telescopes, but there is an alternative. A transmission grating placed in the converging beam before the focus of a telescope will serve the same purpose, but it also introduces optical aberrations and causes a large deflection of the images. However, a combination of a prism and a diffraction grating (called a ***grism***, see Figure 5.6) cancels much of this deflection and reduces chromatic coma, producing images much like an objective prism.

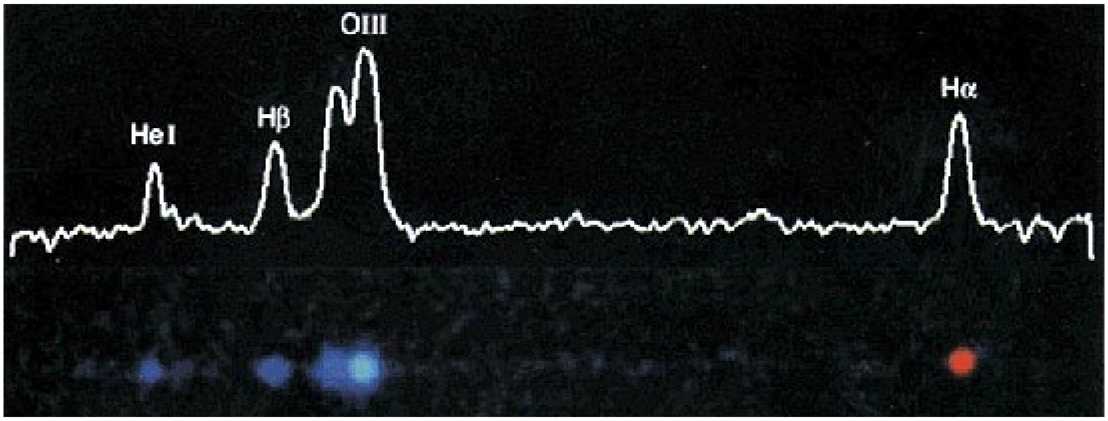


Figure 5.5. Spectrum of the planetary nebula NGC 7662 obtained with an objective prism. The image is shown below, and a profile of the intensity, along the direction of dispersion, is shown above.

There are several disadvantages to this approach. Because there is no slit, all of the light from the night sky passes through to the focal plane. Because of the photon noise from the sky slit-less spectroscopy is limited to relatively bright objects, at least from ground-based telescopes. It is however competitive for space telescopes and has been used very effectively with the STIS spectrograph on the Hubble Space Telescope.

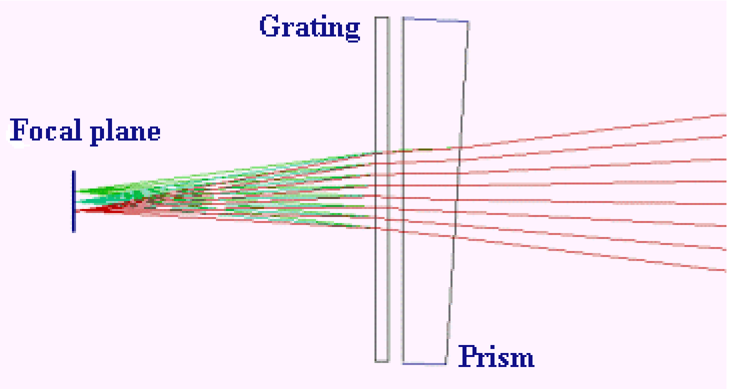


Figure 5.6. Combination of a prism and a diffraction grating to remove chromatic coma. The grating can be ruled directly on the prism, which is then called a grism.

5.1.5 Echelle spectrographs

The problem of overlapping orders, discussed in 5.1.1, can be eliminated in an elegant manner. If a small prism is inserted in the beam, between the slit and the collimator, with angle of dispersion perpendicular to that of the diffraction grating, a ***cross dispersion*** of the spectrum will result. On the detector, the spectrum is not only dispersed in the usual direction, but there is also a small wavelength-dependent displacement in the perpendicular direction as a result of the prism. This cross dispersion breaks the degeneracy of different orders. For example, the wavelengths 800 nm and 400 nm in the first and second orders, respectively, that would have appeared at the same position on the detector, are now separated in the perpendicular direction by a small but sufficient amount. Echelle gratings are designed to be used at high order (*n* ~ 30 for example). The spectra are highly dispersed so that the range of wavelengths in a single order corresponds to the difference in wavelengths between adjacent orders. In this way, an entire high-resolution spectrum can be fit into a rectangular format on the detector (Figure 5.7).

Figures 5.8 -5.10 shows a CAD drawing of UVES, an echelle spectrograph at the Nasmyth focus of one of the 8-meter telescopes of the VLT.

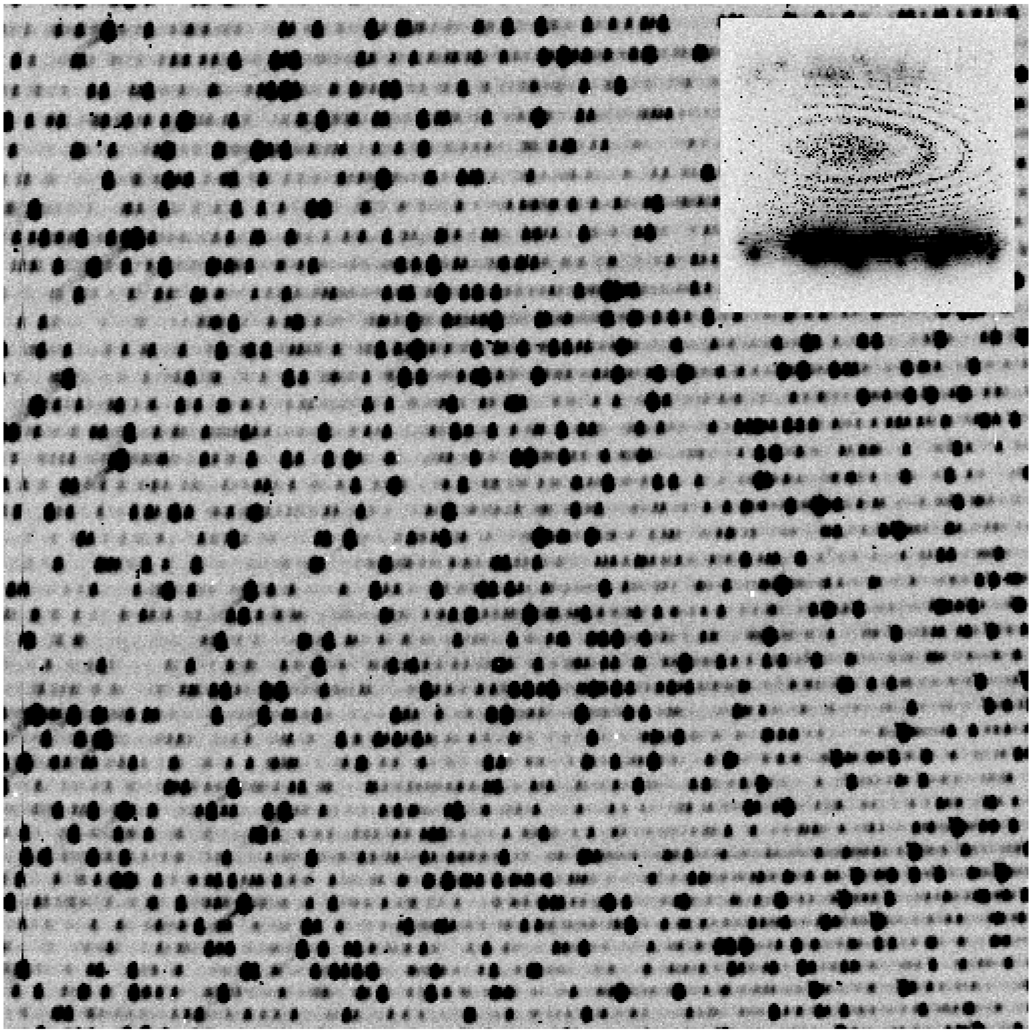


Figure 5.7 Echelle spectrum of a Th-Ar calibration lamp taken with the ARC spectrograph of the Apache Point 3.5-metre telescope. Each horizontal line corresponds to an order.

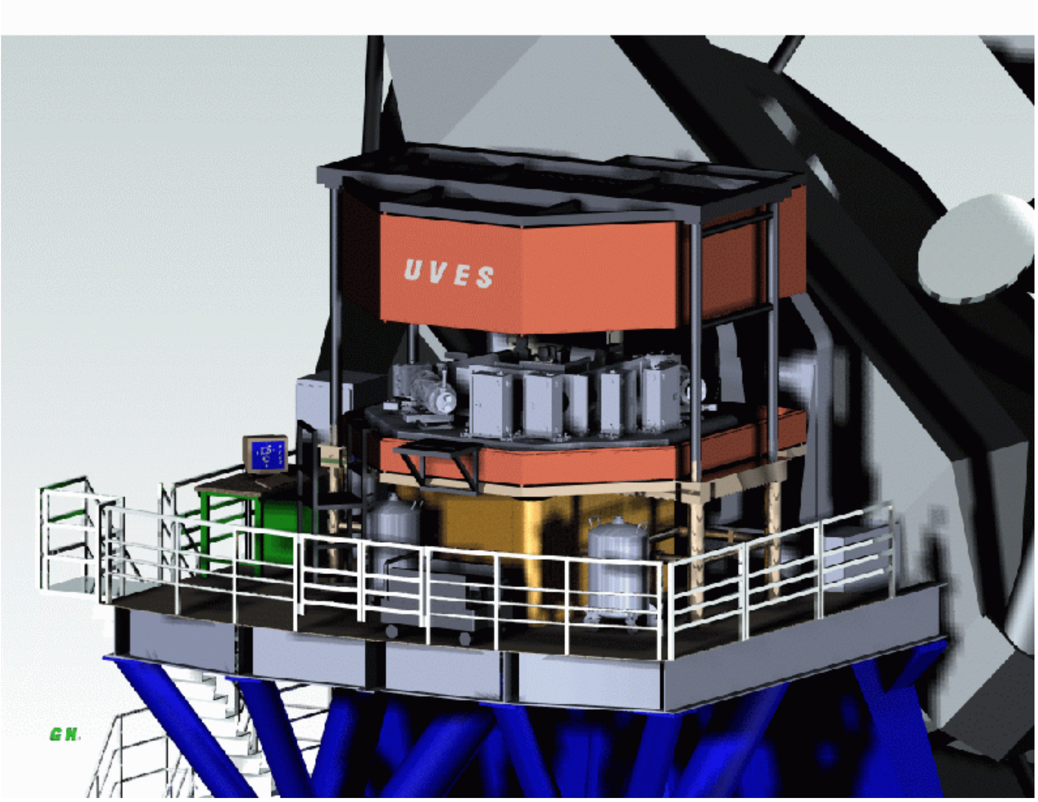


Figure 5.8 the UVES echelle spectrograph on a Nasmyth platform of the VLT.

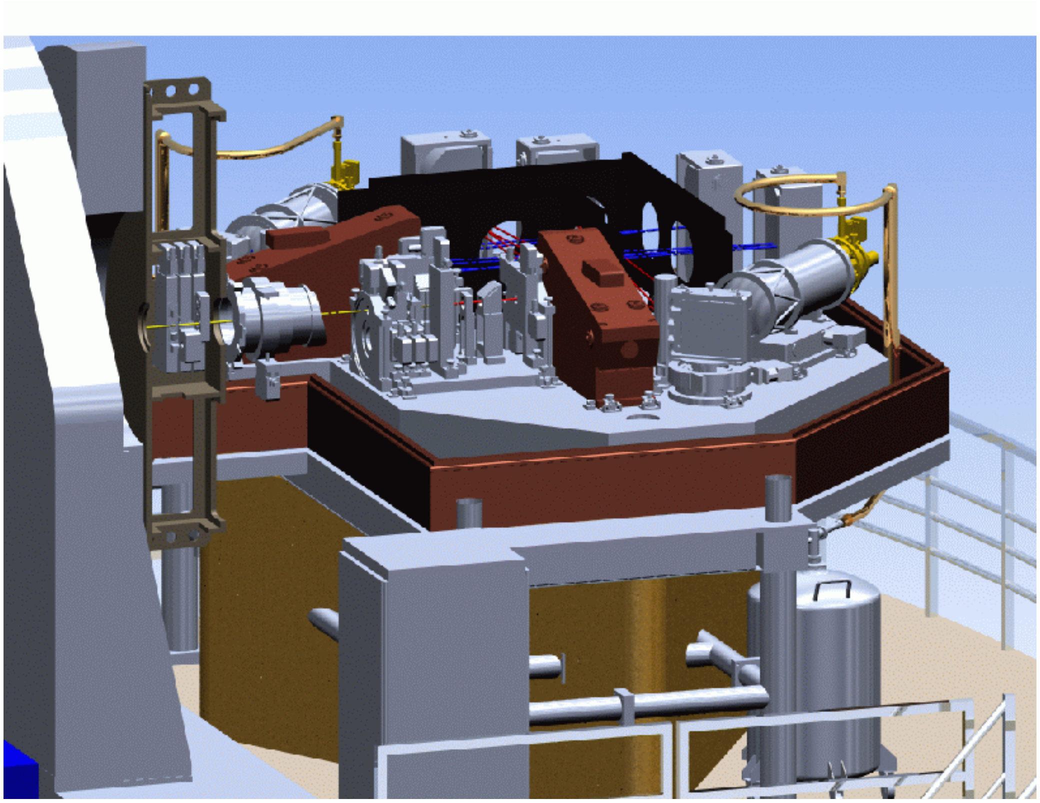


Figure 5.9 View of UVES with the cover removed.

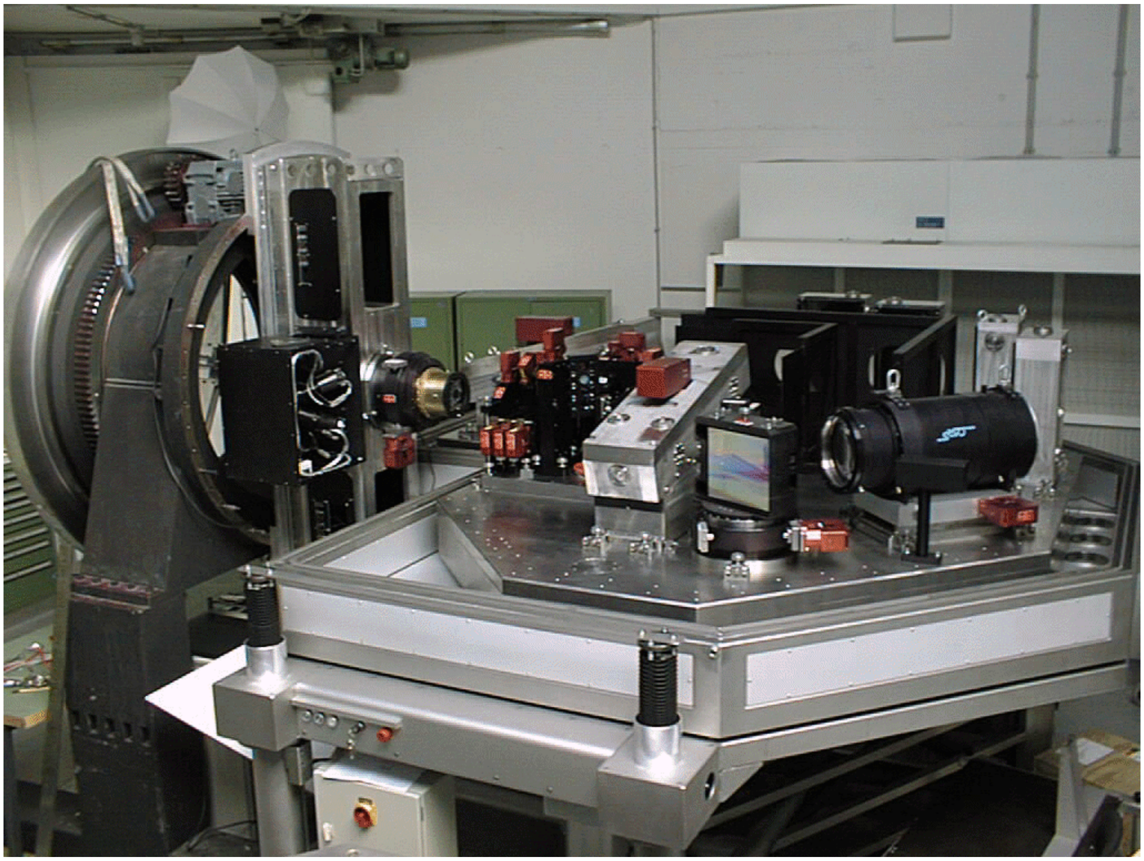


Figure 5.10 UVES under construction in Garching. The echelle grating is visible in the foreground.

Exercises

4.1 The seeing at the Canada-France-Hawaii Telescope is typically 0.8” FWHM in the V band. At approximately what wavelengths would you expect the 3.6-metre telescope to give diffraction-limited images?

4.2 Why do you think that the wavefront sensor (and beamsplitter) in an adaptive optics system is placed after the deformable mirror, rather than before it?

4.3 The Terrestrial Planet Finder (TPF) is a proposed space telescope that would be able to image an Earth-like planets orbiting solar type. One concept for TPF is an optical interferometer consisting of four diffraction-limited telescopes operating in the 3-30 um wavelength range with separations between telescopes in the range 25-1000 km. Estimate the maximum resolution (the PSF FWHM) that could be achieved by TPF. How does this compare with the angular size of the Earth seen at a distance of 10 pc? How does it compare with the angular separation between the Earth and Sun as seen from the same distance?

4.4 Prove (4.68).

4.5 By numerically evaluating the Hankel Transform of the MTF (4.71), and assuming that $r_0 \ll D$, compute the atmospheric PSF. Plot the logarithm of intensity vs radius. How does it compare with a Gaussian function (which is a parabola in this type of plot)?