Wavefront sensors in Adaptive Optics

Theam Yong Chew, B.E.(Hons. I)

Department of Electrical and Computer Engineering

A thesis presented for the degree of
Doctor of Philosophy

University of Canterbury
Christchurch, New Zealand
February 2008

Dedicated to my Papa and Mama

Epigram 101: Dealing with failure is easy: Work hard to improve. Success is also easy to handle: You've solved the wrong problem. Work hard to improve.

Epigrams on Programming, Alan J Perlis

Abstract

Atmospheric turbulence limits the resolving power of astronomical telescopes by distorting the paths of light between distant objects of interest and the imaging camera at the telescope. After many light-years of travel, passing through the turbulence in that last 100km of a photon's journey results in a blurred image in the telescope, no less than 1" (arc-second) in width. To achieve higher resolutions, corresponding to smaller image widths, various methods have been proposed with varying degrees of effectiveness and practicality.

Space telescopes avoid atmospheric turbulence completely and are limited in resolution solely by the size of their mirror apertures. However, the design and maintenance cost of space telescopes, which increases prohibitively with size, has limited the number of space telescopes deployed for astronomical imaging purposes. Ground based telescopes can be built larger and more cheaply, so atmospheric compensation schemes using adaptive optical cancellation mirrors can be a cheaper substitute for space telescopes.

Adaptive optics is referred to here as the use of electronic control of optical component to modify the phase of an incident ray within an optical system like an imaging telescope. Fast adaptive optics systems operating in real-time can be used to correct the optical aberrations introduced by atmospheric turbulence. To compensate those aberrations, they must first be measured using a wavefront sensor. The wavefront estimate from the wavefront sensor can then be applied, in a closed-loop system, to a deformable mirror to compensate the incoming wavefront.

Many wavefront sensors have been proposed and are in used today in adaptive optics and atmospheric turbulence measurement systems. Experimental results comparing the performance of wavefront sensors have also been published. However, little detailed analyses of the fundamental similarities and differences between the wavefront sensors have been performed.

This study concentrates on four main types of wavefront sensors, namely the Shack-Hartmann, pyramid, geometric, and the curvature wavefront sensors, and attempts to unify their description within a common framework. The quad-cell is a wavefront slope detector and is first examined as it lays the groundwork for analysing the Shack-Hartmann and pyramid wavefront sensors.

The quad-cell slope detector is examined, and a new measure of performance based on the Strehl ratio of the focal plane image is adopted. The quad-cell performance based on the Strehl ratio is compared using simulations against the Cramer-Rao bound, an information theoretic or statistical limit, and a polynomial approximation. The effects of quad-cell modulation, its relationship to extended objects, and the effect on performance are also examined briefly.

In the Shack-Hartmann and pyramid wavefront sensor, a strong duality in the imaging and aperture planes exists, allowing for comparison of the performance of the two wavefront sensors. Both sensors subdivide the input wavefront into smaller regions, and measure the local slope. They are equivalent in every way except for the order in which the subdivision and slope measurements were carried out. We show that this crucial difference leads to a theoretically higher performance from the pyramid wavefront sensor. We also presented simulations showing the trade-off between sensor precision and resolution.

The geometric wavefront sensor can be considered to be an improved curvature wavefront sensor as it uses a more accurate algorithm based on geometric optics to estimate the wavefront. The algorithm is relatively new and has not found application in operating adaptive optics systems. Further analysis of the noise propagation in the algorithm, sensor resolution, and precision is presented. We also made some observations on the implementation of the geometric wavefront sensor based on image recovery through projections.

Acknowledgements

I would like to thank Dr Richard Lane for his guidance and assistance during his time as my supervisor, and after that, in a private capacity as Associate Supervisor. I would also like to acknowledge his help in providing funding through a Marsden fund scholarship and for conference fees. Throughout this research, his wisdom, dedication and enthusiasm have kept me motivated and driven.

I am grateful for the financial assistance from the Keith Laugesen Charitable Trust, through the Keith Laugesen Memorial Scholarship throughout most of my study. I also thank my (second) Principal Supervisor Dr Rick Millane who provided conference funding and administrative support. I am also grateful to Lincoln Ventures Limited for a part time position that helped with my living expenses. The work involved with cameras and practical image processing issues have been satisfying.

In 2002, I had the opportunity to work with Judy Mohr on the experimental wavefront sensing and tip/tilt correction workbench for the McLellan Telescope at Mt John Observatory, University of Canterbury. The assistance rendered by Associate Professor Peter Cottrell from the Physics Department, Graeme Kershaw from the Mechanical Workshop, Alan Gilmore (Mt John Superintendent), and Rachel Johnston was instrumental in the success of that project, and is greatly appreciated.

I acknowledge the work done by Rachel Johnston, Marcos van Dam, and Richard Clare, especially for their simulation routines, which I have inherited and learnt from. I thank Jeffrey Hsiao for the assistance rendered in formatting and rendering this thesis, and Nic Blakely for initially creating the Latex style files.

I am also grateful for the friendship of Jeffrey Hsiao, Wei-Lun Chiu and Steve Weddell. My conversations with them along random directions have been useful for bouncing ideas and for entertainment.

Finally and most importantly, I appreciate the support and encouragement from my Mom and Dad through these interesting times.

Contents

Preface

I began my postgraduate studies as a Masters student in 2002, under the supervision of Dr Richard Lane. As part of the requirements for the Masters degree, I took courses in Optical Engineering, Computational Image Recovery, Advanced Systems and Control, Techniques in Observational Astronomy and Applied Electromagnetism. The thesis component of the course involved a simulation of atmospheric turbulence for wide field imaging and correction.

The Electrical and Computer Engineering department collaborated with the Physics and Astronomy Department on an atmospheric sensing and tip/tilt correction system for the university's Mt John Observatory. I had the opportunity to work on the camera software and optical layout calibration for the rig, and was invited to the observatory on several trips to test the system and gather data.

In order to investigate further the performance limits of wavefront sensors, I also began to study the operation of wavefront sensors. To study alternatives to adaptive optics, I adapted the simulations for wide-field imaging through turbulence to examine the phase retrieval problem and employed phase diversity to resolve the ambiguity inherent in phase retrieval. Most of that work was exploratory in nature, and is not documented here.

After upgrading my Masters degree into a PhD degree, initial work with wavefront sensors involved a comparison between the curvature sensor and the geometric wavefront sensor. The exact geometric optics model in the geometric sensor provided it with the obvious advantage when solving for the inverse solution. However, since real images also contain photon noise, how does this advantage translate to practical applications? This motivates the work (Chapter 7) into the comparison between the geometric and curvature wavefront sensors. During the work with photon noise in the geometric wavefront sensor, some interesting properties of the Zernike polynomials under projections were observed, and are described in the Appendix.

Around the time I was analysing noise propagation through the geometric wavefront sensor, I inherited some Matlab code from Richard Clare for simulating the pyramid wavefront sensor. The wavefront estimation routines use direct inversion of a linear model. However, an empirically determined and turbulence dependent scale factor is required to account for the changing sensitivity in the pyramid sensor during operation. Difficulties in determining this scale factor are compounded in closed-loop compensated systems.

This motivated a return to the analysis of the quad-cell (Chapter 5) even though it is well-covered in the literature. The result of that analysis, along with the Fourier duality property, is useful for comparing the performance of the Shack-Hartmann to the pyramid wavefront sensor in Chapter 6. Simulations with different lenslet sizes were carried out to demonstrate the precision-resolution trade-off in the Shack-Hartmann sensor.

0.1 Thesis organisation

The contents of each chapter in this thesis are summarised here. Chapter 1 to Chapter 4 introduce all the preliminaries required to understand the subsequent chapters, and contain no new materials. My new contributions are presented in Chapters 5 to 7.

Chapter 1 provides an introduction to the field of astronomical imaging and the role of adaptive optics in combating atmospheric turbulence. Alternatives to adaptive optics like computer post-processing are also discussed.

Chapter 2 introduces the mathematical techniques and notations used in the subsequent chapters of this thesis. Linear systems theory and probabilistic or statistical techniques are fundamental to the description of atmospheric turbulence, optics, and wavefront sensors.

Chapter 3 reviews the field of optics. Beginning with the geometric ray tracing model, the laws of refraction and reflection and their use in optical systems are described. Diffraction is approximated with the Fresnel and Fraunhofer diffraction models, which are the main mathematical tools used in this thesis. Lastly, the newer optical techniques of Fourier optics and the field transport equations are introduced.

Chapter 4 provides an overview of the statistical properties of atmospheric turbulence, and the wavefront sensors used to detect them. The four main sensors studied in this thesis —the Shack-Hartmann, the pyramid, the curvature, and the geometric wavefront sensors are introduced and a unifying theme is suggested.

Chapter 5 lays the groundwork for the analyses of the Shack-Hartmann and pyramid wavefront sensors by characterising the quad-cell wavefront slope sensor. The performance of the quad-cell is examined and the result is extended to apply to the analysis of wavefront sensors. Conventional quad-cell analysis cannot be applied to closed-loop adaptive optics systems, so the novel contribution from this analysis is a simplified closed-loop analysis of wavefront sensors.

Chapter 6 compares the performance of the Shack-Hartmann and the pyramid wavefront sensors. After developing the Fourier duality of the two wavefront sensors, the quad-cell analysis is applied to compare the slope estimation performance of both wavefront sensors. A unique aspect of this work lies in the use of Fourier duality, which provides a neat classification of the various sensor functions for direct comparison.

Chapter 7 compares the performance of the geometric and the curvature wavefront sensors. The two sensors are physically identical, and their only difference lies in their wavefront estimation algorithm. The geometric wavefront sensor is shown to be an exact model of geometric optics through ray-tracing, while the curvature sensor is shown to be a simplified approximation of the geometric sensor. This chapter proposes a new noise propagation analysis for the geometric wavefront sensor and explores the resolution limits posed by diffraction.

Chapter 8 concludes with a summary and some discussions on future work.

The Appendix also includes some interesting observations and a conjecture on the properties of the projections of Zernike polynomials, which provides a potentially useful tool for the recovery of images through projections.

0.2 Supporting publications

A number of journal and conference publications resulted from work on this thesis. These are listed below.

T.Y. Chew and R.G. Lane, “Estimating phase aberrations from intensity data”, in *Proceedings of Image and Vision Computing New Zealand 2003 (IVCNZ’03)* , D. G. Bailey ed., 181-186.

T.Y. Chew, R.M. Clare and R. G. Lane, “A Cramer-Rao bound analysis of the Shack-

Hartmann and pyramid wavefront sensors", in *Proceedings of Image and Vision Computing New Zealand 2004 (IVCNZ'04)* , D. Pairman, H. North and S. McNeill, eds., 227-232.

T.Y. Chew and R.G. Lane, "Benefits of a single photon wavefront sensor", in *Proceedings of Image and Vision Computing New Zealand (IVCNZ'05)* , B. McCane, ed. 85-89.

T.Y. Chew, R.M. Clare and R.G. Lane, "A comparison of the Shack-Hartmann and pyramid wavefront sensors", in *Optics Communications* , **268** (2), 189-195 (2006).

Chapter 1

Introduction to Astronomical Imaging

The increasing size of ground-based astronomical telescopes has led to the ability to see fainter objects. In the absence of the atmosphere, larger telescope sizes not only increase the light gathering power of telescopes, but also increase the resolution of telescopes, allowing for finer details in astronomical images to be measured. In practice, the increase in size has not been matched by increased resolution, since the atmospheric refractive index fluctuations caused by turbulence distort the light rays from distant stars unevenly across the telescope aperture. This degrades the resolution of all ground-based telescopes to about 1" (arc-second), regardless of telescope size. When astronomical objects are viewed through large astronomical telescopes, they appear blurred and distorted, with the distortion changing over time. Figure 1.1(a) shows a simulated image of a pair of binary stars blurred by atmospheric turbulence when viewed through a large telescope, at an instance in time. The same image is shown in Figure 1.1(b) with adaptive optics to partially cancel the effects of the atmosphere.

Several methods are available to combat the distortions introduced by the atmosphere. The ideal method is to avoid the atmosphere, by using space-based telescopes. In 1990, NASA deployed the 2.4m Hubble space telescope into low earth orbit. It was initially plagued by spherical aberrations, but was successfully repaired in-orbit, and has provided astronomers with deep space images of the universe for close to 15 years. From the original estimated cost of about US$400 million, the telescope eventually cost over US$2 billion, and has been estimated to cumulatively cost up to US$14 billion (inflation adjusted). The Hubble space telescope has not been operating since 2004 following the failure of the imaging spectrograph, and now has an uncertain future. Without further repairs and maintenance,

(a) A pair of double stars seen through a frozen instance of atmospheric turbulence.

(b) The double stars with partial correction of the same turbulence using adaptive optics.

Figure 1.1 Simulations of the effects of adaptive optics on atmospheric turbulence induced blurring.

the telescope will eventually re-enter the atmosphere [4].

The successor to the Hubble Space Telescope, the James Webb Space Telescope [5], is in its preliminary design stages and is planned for launch in 2013. The telescope is designed with a 6.5m folding mirror and operates at infra-red wavelengths of 0.6 to 28 μm. The telescope will orbit the sun at the L2 Lagrange point between the Sun and the Earth, 1.5 million km away. The budget for the James Webb Space Telescope project is currently about US$ 3.5 billion, which is US$ 1 billion over-budget.

Due to the prohibitive cost of space telescopes, more practical solutions are needed to overcome the effects of atmospheric turbulence. Adaptive optics systems for ground-based telescopes provide an alternative solution. Using optical elements which modify the propagation of light to cancel the effects of atmospheric turbulence, image quality can be restored to near the ideal performance. More importantly, this correction can be achieved over a range of light frequencies and consequently is more useful than computer post-processing methods.

Adaptive optics systems are used today in most large optical telescopes for compensating the effects of atmospheric turbulence. The design of an adaptive optics system must consider the costs and appropriateness of ever changing technology, the characteristics of atmospheric turbulence at a specific observation site, and the type of observation to be performed at the site. These observations consist mainly of spectroscopy, photometry and direct imaging.

Spectroscopy [99] is the analysis of the composition of stars or materials by decomposing light into its component spectra. The spectrograph performance (spectral resolution) is determined by a narrow slit in the spectrograph. The size of the slit is traditionally matched to, and limited by, the blurred focal plane image of a point-source object, so adaptive optics can be used to reduce the slit size and increase the spectrograph resolution.

Photometry is the measurement of stellar magnitudes (intensity) [108]. The main objective is consistent measurement of intensities, so image resolution is normally not important for performance. Adaptive optics has limited applicability in this area, and in fact, by reducing the light throughput, actually degrades system performance.

Direct imaging, whether by a recording medium like photographic plates or CCD cameras, is similar to sight in the human eye. The intensity distribution of a distant object is re-imaged with a lens or mirror, and then recorded. The imaged objects can be point-source stars, double stars, distant extended objects like star systems, galaxies and nebulae, or nearby extended objects like planets and comets. Image sharpness, resolution and contrast are important, and adaptive optics can play a crucial role in such applications. Unlike spectroscopic applications, images recorded directly may be further enhanced with computer post-processing.

A promising new technique for high resolution imaging, interferometric imaging, provides very high but selective resolution by using multiple telescopes arranged on long baselines of nearly hundreds of meters. The long distances involved require precise calibration of the phase delay arising from the different imaging path lengths and atmospheric turbulence. The major contribution to image degradation comes from the phase piston term between widely separated apertures. Techniques in adaptive optics have also been adapted to this specialised application.

1.1 Adaptive optics

In 1953, Babcock [9] suggested the first adaptive optics system [62, 85, 101] for real-time aberration compensation. An adaptive optics system is shown in Figure 1.2, consisting of a wavefront sensor to measure aberrations caused by the atmosphere, a wavefront corrector or deformable mirror driven in closed loop by a command computer, and an imaging channel that carries out the scientific observations.

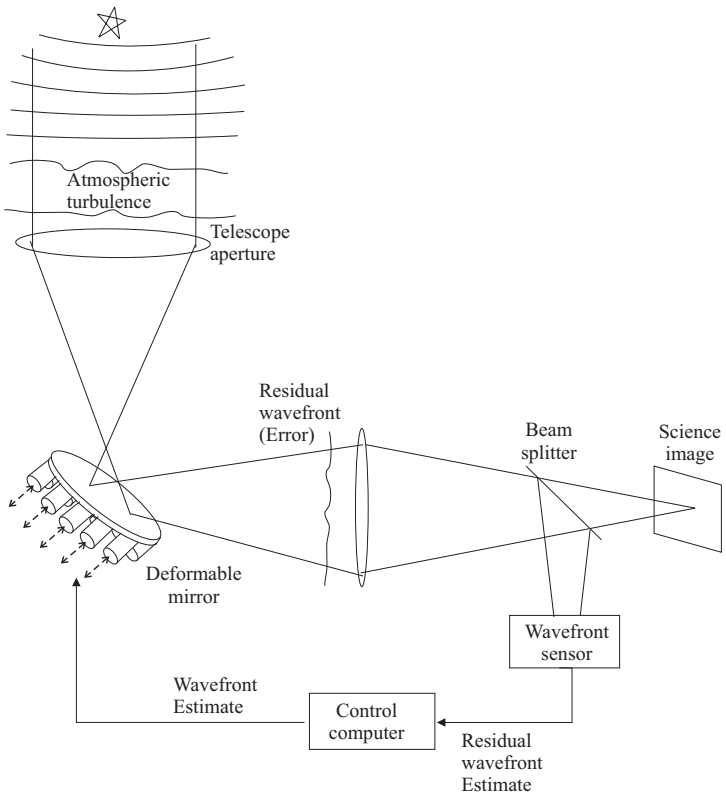


Figure 1.2 Real-time correction of atmospheric turbulence with a closed-loop adaptive optics system.

1.1.1 Atmospheric turbulence

Atmospheric turbulence is caused by the mixing of air of different temperature and pressure, and water vapour. The turbulent motion of air starts from large scale motions, but through the viscosity and friction of moving air, the motion ends at smaller and smaller scales, and eventually dissipates as heat. This process results in fluctuations in the refractive index of air, so these “Tremors of the atmosphere”, as described by Sir Isaac Newton [64], are perceived from the ground by the naked eye as the “twinkling of fixd Stars”.

Although turbulence is considered to be present in the whole of the atmospheric troposphere and stratospheric layers, its strongest measurable effects are usually localised to several strong layers, typically located about 10km high in the sky. Often, strong ground layers are also present. The increasing awareness of the presence of strong ground layers in recent years has led to greater care being taken to select suitable sites, and the adoption of construction practices that reduce the observatory thermal signature to reduce ground layer turbulence [85, 101].

Several statistical measures of the severity of turbulence are used as rule-of-thumb indicators of the image quality or "seeing" achievable at an observatory site [89]. They are typically expressed as angles of seeing (size of a blurred point-source), the isoplanatic angle θ_0, usually around a few arc-seconds, turbulence cell size (also known as Fried's parameter r_0, usually from 5 to 20 cm), or a rate of change (Greenwood's frequency f_G, typically in the 20 to 100 Hz range). All these quantities are derived from the refractive index fluctuations of turbulence (as measured by the structure constant C_N^2) and wind speed. The statistical properties of atmospheric turbulence are examined further in Section 4.1.

1.1.2 Wavefront sensors

Wavefront sensors are used to estimate the image aberrations caused by the atmosphere. The most practical way today to measure turbulence is to measure its effect on light. The most commonly used wavefront sensor, the Shack-Hartmann wavefront sensor [72] (to be examined in Section 6.3) is shown in Figure 1.3, and illustrates most of the basic principles of wavefront sensors.

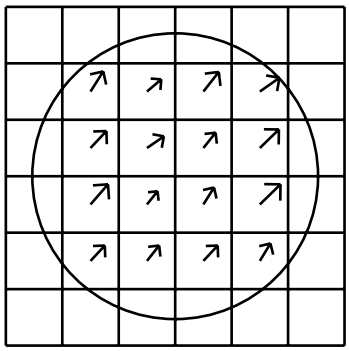


Figure 1.3 The Shack-Hartmann wavefront sensor divides the circular telescope aperture into smaller regions, and combines the local slope signals (shown as arrows in each subregion) to form the full wavefront estimate over the whole aperture.

The Shack-Hartmann sensor consists of an array of lenslets spread across the telescope aperture, subdividing it into smaller regions. The effect of atmospheric turbulence is localised within each lenslet and proportionately reduced. The lenslet image is displaced randomly over time, but since the lenslet size is usually chosen to be approximately equal to r_0, maintains an image size close to the un-aberrated case. The image displacement is

linearly proportional to the slope of the atmospheric wavefront [93]. In general, all wavefront sensors produce a vector signal derived from the wavefront slope over sub-regions in the aperture and invert the linear relationship to recover the full wavefront across the whole telescope aperture.

Like all wavefront sensor, the Shack-Hartmann sensor does not work well with dim objects and requires at least 20 photons in each sub-aperture to provide useful wavefront estimates. Therefore, the wavefront estimate is frequently obtained from measurements on a nearby guide star instead of the target star itself. This avoids the loss of light from the target star measurements and may even allow a brighter star to be used for wavefront sensing. However, a nearby natural guide star is often unavailable. Observatories today [2, 3, 6, 7] are equipped with artificial laser guide star systems that can form a bright spot high in the atmosphere at selected positions.

1.1.3 Imaging camera

The first imaging devices are based on photosensitive materials coated on photographic film. Today, most imaging devices have been replaced by semiconductor technology, such as linear arrays of charge-coupled devices (CCD), or more recently, complementary metal-oxide-semiconductor (CMOS) sensors. Most wavefront sensors use CCD sensors as their light detector, so the characteristics of CCD detectors play an important role in the performance of wavefront sensors.

The most important characteristics of CCD imaging devices are their efficiency, spectral sensitivity and noise levels. CCD devices can detect as much as 90% of available photons if substrate thinning and back illumination are employed with close packing of the individual photosites (fill factor). Additionally, anti-reflection and fluorescence (expanding the spectral sensitivity range) coatings are often used [1].

CCD cameras are also affected by noise during their operation. Thermal or dark noise arises from accumulated random fluctuations of electrons in thermal motion. Dark noise obeys Poisson statistics and accumulates over time at a rate proportional to the temperature of the imaging site. It is reasonably consistent for any individual pixel. Read-out noise or read noise is a bias introduced mainly by the amplifiers in the on-board measurement electronics. It is roughly proportional to the amplifier gain. Knowledge of the noise statistics allows their effects to be reduced by proper calibration of images with averaged noise frames.

Aside from instrumentation noise, images are sometimes affected by cosmic rays, which typically saturate individual pixels, creating a salt and pepper noise effect. Typically, single saturated pixels are removed in a separate preprocessing step to remove this noise. At low light levels, randomness in the photon arrival process gives rise to noise obeying Poisson statistics. Poisson noise cannot be reduced except by increasing the received light level. Witthoft [113] investigated a way to reduce photon noise relative to detector read noise by image intensification.

1.1.4 Control computer

A control computer transforms signals from a wavefront sensor into the appropriate actuation voltage signals. The signals drive the deformable mirror in a closed loop control system. The corrections applied by the system have to take place faster than the atmospheric time constant, which is typically a few milliseconds [101].

The control computer is modelled as a simple closed loop control system. The most significant effect arising from turbulence is image displacement, caused by the tip/tilt term, so providing a separate flat mirror significantly reduces the demand on the deformable mirror. For this reason, it is sufficient to consider only a single channel here. Figure 1.4 shows the layout of a tip and tilt only adaptive optics system or an image displacement stabiliser. The incoming light, the light from a distant star, provides the input signal. The output is the stabilised image used for scientific observations. Instrumentation noise (at the tilt sensor) and photon noise (inherent in the input) are also present.

The performance of the system is determined by the classical factors [23] in a control system: the noise level, the delay introduced by each component in the adaptive optics system, and the rate of change of atmospheric turbulence.

1.1.5 Deformable mirrors

Optical phase compensation devices work by introducing a phase shift along the light path of an optical system. Devices based on birefringent materials or LCD phase shifters can be used for manipulating optical phase directly. However, the aberrations created by atmospheric turbulence, resulting from irregular refractive index fluctuations in the air, is dependent on the light wavelength. Deformable mirror membranes are used in practical systems, since they cancel aberrations by physical path differences instead of phase and have no wavelength dependence. Furthermore, deformable mirrors show a uniform response, and

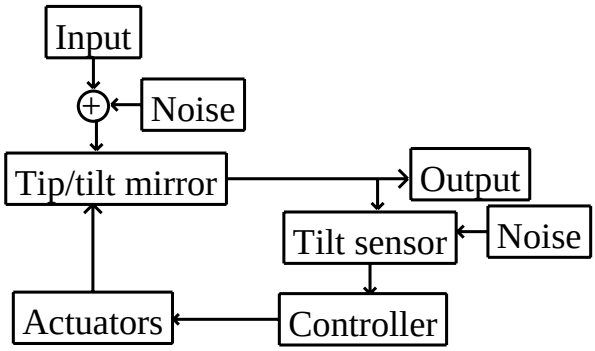


Figure 1.4 The feedback control loop for an image stabiliser adaptive optics system.

have short response times.

Segmented mirrors were used in early mirror prototypes, but have fallen from favour because of their high wavefront fitting errors at the edges. The most commonly used deformable mirror today has continuous facesheets. The facesheet is a flexible reflecting membrane supported by many micro-actuators that can be adjusted at high speeds to shape the mirror surface. The micro-actuators are usually built from ferroelectric ceramic materials that have a piezoelectric response to strong electric fields. The bimorph mirror, another deformable mirror with continuous facesheets, uses two piezoelectric ceramic wafers that locally contract in opposing directions when a voltage is applied through an electrode, causing a local deformation around the electrode [85].

The mechanical properties of the mirror actuators determine the characteristics of deformable mirrors. The stroke (amount of movement) determines the maximum phase correction that can be compensated by the mirror. The number and positions of the actuators limit the complexity of the phase function that can be compensated. The geometry of the actuators also affects the coupling between actuators, with the elasticity of the coupling determining the response time (typically in the millisecond range) of deformable mirrors.

1.2 Post-processing of images

Computer post-processing of images is an attractive alternative to adaptive optics. This may also complement an adaptive optics system at a later stage, to enhance the output images

from the adaptive optics system.

The model for the blurring introduced by the atmosphere is given by

$$d(x,y) = f(x,y) \odot h(x,y) + n(x,y) \tag{1.1}$$

where $f(x,y)$ is the original image, blurred by $h(x,y)$, the instantaneous atmospheric point-spread-function at a certain time, and $n(x,y)$ is the additive noise. The contaminated final image $d(x,y)$, along with constraints made using assumptions of the properties of atmospheric blurring, is used to recover the original image. Image restoration using deconvolution belongs to the class of inverse problems, where a model of the forward problem is inverted to recover the original image, often solved using iterative techniques [74].

Post-processing techniques require light to be detected before processing offline, unlike fully online adaptive optics systems. This has the disadvantage that it cannot be used in cascade with non-imaging observations like spectroscopy or interferometry, which require real-time compensation.

1.2.1 Image deconvolution

In conventional image deconvolution, the contaminated image $d(x,y)$ is known along with an approximate model of the blurring, $h(x,y)$. Given these two datasets, the original image $f(x,y)$ can be recovered by reversing the equivalent filtering operation. A knowledge of the energy statistics of the original image compared to the noise can be used to design optimal filters known as Wiener filters. In astronomical imaging, the shot noise from randomness in photon arrivals dominates the $n(x,y)$ term. For this class of problems, alternatives like the CLEAN and Richardson-Lucy iterative algorithms are more commonly applied. They are maximum likelihood solutions for Poisson noise statistics [55, 79].

A more challenging class of problems is encountered when the only measurement available is from $d(x,y)$, so the original image $f(x,y)$ has to be recovered along with the blurring function $h(x,y)$ too. In astronomical imaging, this is mitigated by storing and processing a large number of frames of the image (refer Equation 1.2). Over all frames, the original image remains the same, while the atmospheric blurring function and noise vary, giving

$$d_i(x,y) = f(x,y) \odot h_i(x,y) + n_i(x,y) \tag{1.2}$$

with i representing an image frame index.

The blind deconvolution problem is frequently under-constrained due to the small number of measurements compared to the image that is to be recovered. The possible solutions to the problem are frequently restricted by additional constraints arising from the physical limits of the imaging problem, such as positivity, smoothness and finite support of images. A related and more restricted class of problems can be used for recovery of the atmospheric phase aberrations that lead to image blurring.

1.2.2 Phase retrieval

Phase retrieval refers to the class of techniques used to recover the phase information using the information from intensity and prior information [24, 59, 92]. It is applied in fields as diverse as astronomical imaging, microscopy, crystallography, sonar, and radar, among others [60]. Most spectacularly, it has been used to estimate the aberrations in the Hubble space telescope [82] using only the aberrated stellar images captured from the telescope while in Earth orbit.

When a distant star ($f(x,y)$ being a point-source object) is imaged through the atmosphere, the measured image is given by

$$\begin{aligned} d(x,y) &= \delta(x,y) \odot h(x,y) + n(x,y) \\ &= h(x,y) + n(x,y) \end{aligned} \tag{1.3}$$

Usually, the atmospheric turbulence is approximated by a single layer of phase-screen that adds random phase perturbations to the passing light. As shown later in Section 3.3.2, $h(x,y)$ is derived from the Fourier Transform of the telescope aperture function and phase aberrations propagated from the phase-screen.

$$h(x,y) = \left| \mathscr{F} \left\{ A(u,v) e^{i\phi(u,v)} \right\} \right|^2 \tag{1.4}$$

Here, the image magnitude at the imaging plane, $h(x,y)$, and the aperture magnitude, $A(u,v)$ (usually taken to be $\mathrm{circ}(\sqrt{u^2+v^2})$), are known. Using two images, the phase $\phi(u,v)$ needs

to be recovered. From this point of view, all imaging results in a lost of phase information, since only the magnitude of a complex field is measurable.

From Equation 1.4, three classes of solutions to the phase estimate $\hat{\phi}(u,v)$ that produce the same output image $h(x,y)$ exist

$$\{\hat{\phi}(u,v)+c\},\{\hat{\phi}(u,v)+2\pi k(u,v)\},\{\hat{\phi}(-u,-v)\} \tag{1.5}$$

for integer values of $k(u,v)$ and for $A(u,v)=A(-u,-v)$ (circular symmetry is common in telescope apertures).

The absolute phase value has no effect on the imaging problem, so the first class of ambiguity is usually resolved by setting the DC term to zero, $\sum_u \sum_v \hat{\phi}(u,v)=0$. The second class of ambiguity results from the 2π wrap-around in the phase representation. This may be resolved by phase unwrapping techniques commonly used elsewhere in signal processing, or by applying a smoothness constraint to the solution. The third ambiguity is not resolvable, and in practice, additional information is required from other sources (an estimate of the original solution provides a good starting point).

Setting aside these ambiguities, most solutions to the phase retrieval problem are iterative techniques aimed at reducing some error measure. For example, using an initial guess of the phase function $\hat{\phi}(u,v)$, an estimate for the image $\hat{h}(x,y)$ is produced, and compared to the actual image $h(x,y)$. The initial estimate of the phase is modified iteratively to reduce the difference between the corresponding image estimate and the measured image. Alternatively, in the Gerchberg-Saxton method [32], the estimate of the complex field is transformed back and forth through the Fourier domain. In each domain, a projection operation based on constraints imposed by the measured intensity of the image is performed.

Resolving the ambiguity to the phase retrieval problem requires additional measurements. Additional measurements not only collect more light (increasing the signal to noise ratio), but can also measure slightly different aspects of the data. Phase diversity is a concept similar to diversity in wireless radio communications. The same phase aberration is measured through different “channels” to provide multiple viewpoints on the same data.

1.2.3 Phase diversity

In phase diversity [35, 48], extra measurements of the same object and phase aberrations are taken to help condition the problem, and resolve ambiguities. In the simplest case, controlled phase aberrations are added to a second light path to create a second image, such as

$$\begin{aligned} h_1(x,y) &= \left| \mathscr{F}\left\{ A(u,v) e^{i\phi(u,v)} \right\} \right|^2 \\ h_2(x,y) &= \left| \mathscr{F}\left\{ A(u,v) e^{i(\phi(u,v)+\Delta\phi(u,v))} \right\} \right|^2 \end{aligned} \tag{1.6}$$

The most popular form of phase diversity is the quadratic wavefront $\Delta\phi(u,v) \propto u^2 + v^2$, which corresponds to a defocus. This is usually chosen for its simple implementation.

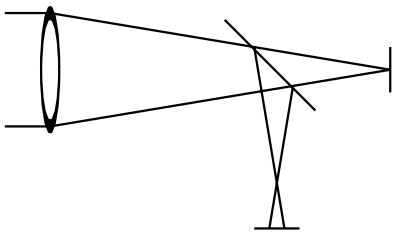


Figure 1.5 Adding a quadratic phase term using a defocus.

The extra defocused plane image directly helps to resolve the ambiguity in rotationally symmetric solutions, and often also allows iterative algorithms to converge faster. When the phase to be estimated is small, an even simpler linearised solution is possible [36].

Phase retrieval has been proposed for measuring optical misalignment in segmented telescopes [70] and even in a real-time experimental adaptive optics control system [52]. This method has also been extended to wider fields of view [34].

The defocused phase diversity arrangement is actually similar to the physical layout of the curvature sensor, which will be examined in Chapter 7. However, unlike the curvature sensor, the defocus in the phase diversity arrangement is much smaller, so that non-linear diffraction effects dominate over geometric optics. The short defocus length also results in the output signal having a higher sensitivity to the input phase.

The optimal form for the diversity wavefront remains an open question, and has been explored [57]. More generally, the extra measurements may be different from the original image in several ways. Other means of diversity can be obtained through using a different part of the light spectrum, different imaging positions, or by taking a sequence of images.

1.2.4 Conclusion

In conclusion, most image processing algorithms can run on cheap off-the-shelf hardware, but may take up too much time to be practical for real-time use. In contrast, in an adaptive optics system, the feedback loop allows for higher loop gains, potentially leading to higher performance. For certain applications, spectrography for example, the output from an adaptive optics system needs to be further processed optically, so post-processing techniques have limited uses here.

Chapter 2

Mathematical background

Linear algebra and the theory of linear systems are used heavily in optical systems for describing light propagation and image transformation. The use of transforms in linear systems theory also requires manipulation of complex numbers. We introduce the mathematical notation used in this thesis, and examine some commonly used special functions and their properties.

2.1 Vectors and matrices

Vectors, being 1D arrays of numbers, are represented with bold lower case letters $\boldsymbol{v}$. The n^{th} element of the vector is represented with a subscript $\boldsymbol{v}_n$, with the first element indexed starting from 1. Matrices can be viewed as extensions of vectors, being composed of 2D arrays of numbers. Matrices are represented with bold uppercase letters $\boldsymbol{M}$, with the element at row i and column j being $\boldsymbol{M}_{ij}$. The trace (sum of diagonal elements), transpose, inverse, and pseudo-inverse of the matrix $\boldsymbol{M}$ are denoted by $tr\{\boldsymbol{M}\}$, $\boldsymbol{M}^T$, $\boldsymbol{M}^{-1}$ and $\boldsymbol{M}^+$ respectively. Matrix multiplication is often used to describe a linear operator on sampled 1D signals represented as vectors. The precise definition for the pseudo-inverse of a matrix varies depending on the application and is defined separately for each problem.

In this thesis, we frequently encounter 2D signals in the form of images or projections. Instead of using a separate notation for linear operations on 2D signals, we continue to use 2D matrix operators and 1D vectors. The 2D signal, represented as a matrix, is stacked into a vector and multiplied with a matrix representing a linear operation on the image. A matrix $\boldsymbol{M}$ of size n by m is stacked into a vector $\boldsymbol{v}$ by

$$\boldsymbol{v}_i = \boldsymbol{M}_{i'j'} \tag{2.1}$$

for $i' = \mathrm{mod}(i,n)$ and $j' = \mathrm{ceil}\left(\frac{i-1}{n}\right)$, where the mod operation takes the remainder of i divided by n, and ceil rounds a fractional non-integer number upwards.

This reduces a 2D matrix to a 1D vector by rearranging the columns of the matrix in order, into a column vector. If applicable, the result of a linear operation $\boldsymbol{N}$ on the signal $\boldsymbol{v}$, $\boldsymbol{v'} = \boldsymbol{N}\boldsymbol{v}$ can be unstacked into an n' by m' matrix by

$$\boldsymbol{M'}_{ij} = \boldsymbol{v'}_{i+n'(j-1)} \tag{2.2}$$

In linear operations, matrices are used as a compact notation to describe weighted sums of signal components. Using a matrix representation for a problem allows results from linear algebra theory to be used. From a practical point of view, many high quality and tested numerical recipes for matrices can be reused in simulations.

2.2 Complex numbers

Complex numbers first arise as general solutions to quadratic polynomials. A complex number $a + ib$ is sum of a real and imaginary component. The imaginary component is formed from i, defined as $\sqrt{-1}$. Complex numbers, and functions of complex numbers are frequently plotted on an Argand diagram as vectors with rectangular coordinates, as shown in Figure 2.1.

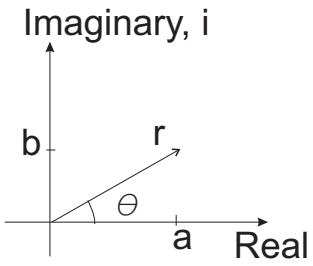


Figure 2.1 Argand diagram for the complex number $a + ib$, represented as a vector, with the real and imaginary components lying along the x and y axes. The magnitude is r and the argument is θ.

Using this geometric representation, we can also represent a complex number with its length

and orientation, using a polar coordinate. The length or magnitude of a complex number, and its orientation, measured by its angle (argument) from the x-axis, is defined by

$$\begin{aligned} r &= \sqrt{a^2+b^2} \quad a = r\cos\theta \\ \theta &= \tan^{-1}\left(\tfrac{b}{a}\right) \quad b = r\sin\theta \end{aligned} \tag{2.3}$$

The polar and rectangular forms of a complex number is linked by[1]

$$a + ib = r\cos\theta + i\sin\theta = re^{i\theta} \tag{2.4}$$

The magnitude and argument representation is commonly used to represent the magnitude and phase of a sinusoid, resulting in a complex field representation for electromagnetic waves.

The conjugate of a complex number $c = a + ib$ is defined to be $a - ib$, and is represented by $\overline{c}$.

2.3 Special functions

Some special functions are frequently used throughout the thesis, and are outlined here. They frequently have discontinuities or infinities, and are more appropriately termed generalised functions or distributions.

2.3.1 Circ function

The circ function is a 2D circular symmetric function that is useful for describing the circular aperture of telescopes, lenses and other optical components. Due to its circular symmetry, the circ function is also frequently parametrised using a single variable, as is shown here

[1] A specific form of Equation 2.4, $e^{i\pi} + 1 = 0$, is said to form the most beautiful equation in the world, since it relates many of the most important constants from the major branches of mathematics together.

$$\mathrm{circ}(r) = \begin{cases} 1 & \text{for } r < 1 \\ 0 & \text{otherwise.} \end{cases} \tag{2.5}$$

where $x^2 + y^2 = r^2$.

2.3.2 Rect function

The rect function can be used to describe rectangular aperture in optical components. In two dimensions, the rectangular function $\mathrm{rect}(x)\mathrm{rect}(y)$ is separable into the products of two 1D functions, and provide a convenient way to analyse systems by reducing the dimensionality of the problem.

$$\mathrm{rect}(x) = \begin{cases} 1 & \text{for } -\frac{1}{2} < x < \frac{1}{2} \\ 0 & \text{otherwise.} \end{cases} \tag{2.6}$$

2.3.3 Step function

The Heaviside step function is used to describe a discontinuity between two regions. This can be seen in the analysis of the knife-edge test in the pyramid wavefront sensor.

$$\mathrm{U}(x) = \begin{cases} 1 & \text{for } x \geq 0 \\ 0 & \text{for } x < 0 \end{cases} \tag{2.7}$$

The related signum function is also commonly used for the same purpose.

$$\mathrm{sgn}(x) = \begin{cases} -1 & \text{for } x < 0 \\ 0 & \text{when } x = 0 \\ 1 & \text{for } x > 0 \end{cases} \tag{2.8}$$

2.3.4 Tri function

The triangular function is also useful for describing certain functions like the optical transfer function of square lenses.

$$\text{tri}(x) = \begin{cases} 1-|x| & \text{for } |x|<1 \\ 0 & \text{everywhere else.} \end{cases} \tag{2.9}$$

2.3.5 Sinc function

The sinc function arises in the analysis of the diffraction patterns of images, and gives a convenient shorthand for dealing with the Fourier transforms of rectangular functions.

$$\text{sinc}(x) = \frac{\sin(\pi x)}{\pi x} \tag{2.10}$$

2.3.6 Bessel functions

The family of functions known as the Bessel functions are frequently encountered in problems with rotational symmetry. The zeroth order Bessel function may be variously defined to be the solution to the differential equation

$$x^2\frac{d^2y}{d^2} + x\frac{dy}{dx} + x^2y = 0 \tag{2.11}$$

or with its power series

$$J_0(x) = 1 - \frac{x^2}{4} + \frac{x^4}{64} - \frac{x^6}{2304}... \tag{2.12}$$

or as the solution to the integral

$$J_0(x) = \frac{1}{2\pi}\int_0^{2\pi} \cos(x\cos\phi)\, d\phi \tag{2.13}$$

The last integral definition provides some intuition into the nature of the Bessel function. Using a coordinate transform mapping the rectangular coordinates (x,y) to the rotated coordinates $(u,v) = (x\cos\phi + y\sin\phi, -x\sin\phi + y\cos\phi)$, Equation 2.13 can be defined only along the x-axis ($y = 0$) as

$$J_0(x) = J_0(x,0) = \frac{1}{2\pi}\int_0^{2\pi} \cos(u)\, d\phi \tag{2.14}$$

This shows that $J_0(x)$ is a sum of 2D cosinusoidal gratings over all orientations, as shown in Figure 2.2.

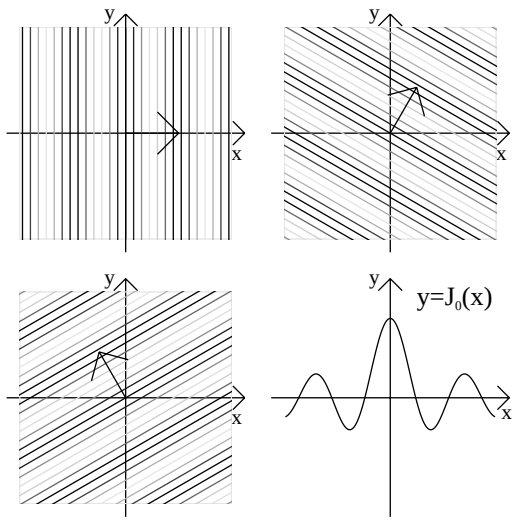


Figure 2.2 The zeroth-order Bessel function as a sum of 2D cosinusoidal waves (with the u-axes shown) rotated over all directions ϕ in the 2D plane. A 1D slice of the rotationally symmetric sum (a 2D function) is shown plotted.

More generally, other Bessel functions of the first kind, of order α are solutions to

$$x^2\frac{d^2y}{d^2} + x\frac{dy}{dx} + (x^2 - \alpha^2)y = 0 \tag{2.15}$$

with the power series representation

$$J_\alpha(x) = \sum_{m=0}^{\infty}\left[\frac{(-1)^m}{m!\Gamma(m+\alpha+1)}\left(\frac{x}{2}\right)^{2m+\alpha}\right] \tag{2.16}$$

where $\Gamma(x)$ is the Gamma function.

2.3.7 Jinc function

The Jinc function is the rotationally symmetric analogue to the sinc function.

$$\text{Jinc}(x) = \frac{J_1(\pi x)}{2x} \tag{2.17}$$

The projection of the rotationally symmetric Jinc function is a sinc.

$$\int_{-\infty}^{\infty} \text{Jinc}(\sqrt{x^2+y^2})\, dx = \int_{-\infty}^{\infty} \text{Jinc}(r)\, dx = \text{sinc}(x) \tag{2.18}$$

The first few zeros of the Jinc function are located at x = 1.220, 2.233, 3.239 ... etc.

Some useful properties of the Jinc function are shown here.

$$\text{Jinc}(0) = \frac{\pi}{4} \tag{2.19}$$

$$\int_{-\infty}^{\infty} \text{Jinc}(x)^2\, dx = \frac{2}{3} \tag{2.20}$$

$$\int_{0}^{\infty} x\text{Jinc}(x)^2 dx = \frac{1}{8} \tag{2.21}$$

Equation 2.21 is useful for finding the volume under the circularly symmetric $\text{Jinc}(x)^2$ function.

$$\int_0^{2\pi}\int_0^{\infty} \text{Jinc}(r)^2 r\, dr\, d\theta = \int_0^{2\pi} \frac{1}{8}\, d\theta = \frac{\pi}{4} \tag{2.22}$$

2.3.8 Chirp function

The chirp function describes a signal with a linearly increasing “instantaneous” frequency. Here, it is generalised to a complex exponential with quadratic phase.

$$f(x) = e^{ia_x x^2} \tag{2.23}$$

The 2D chirp function is a separable function form from the products $f(x)f(y)$. When

$f_x = f_y$, the chirp function also possesses a circular symmetry.

$$f(x,y) = e^{ia_x x^2} e^{ia_y y^2} = e^{iar^2} \tag{2.24}$$

for $r^2 = x^2 + y^2$.

The real quadratic exponential function or the Gaussian function, is a special case of the chirp function.

$$f(x) = e^{-\pi x^2} \tag{2.25}$$

This is the basic form for the normal distribution function used in statistics to describe many naturally occurring statistical distributions.

2.3.9 Delta function

The delta function is a convenient mathematical shorthand used to model sharp impulse events very with short time-scales. In images, this can model a point-source object so small that the signal is 0 everywhere except at a point, yet possesses a finite integral nonetheless.

$$\begin{aligned} \delta(x) &= \begin{cases} \text{undefined } (\infty) & \text{for } x = 0 \\ 0 & \text{everywhere else.} \end{cases} \\ \text{and } \int_{-\infty}^{\infty} \delta(x)\, dx &= 1 \end{aligned} \tag{2.26}$$

The delta function possesses the sifting property that allows us to decompose all functions into an integral sum of delta functions. It also acts as a functional that maps a function to a scalar value, namely, the value of the function at the position of the delta function x'.

$$f(x) = \int_{-\infty}^{\infty} f(x')\delta(x - x')dx' \tag{2.27}$$

for all functions $f(x)$.

2.3.10 Comb function

The comb function (also known as the Shah function) is formed from equally spaced delta functions. It is used for representing the signal sampling process.

$$\text{comb}(x) = \sum_{k=-\infty}^{\infty} \delta(x-k) \tag{2.28}$$

2.4 Linear systems

Many physical processes can be idealised as black boxes with linear properties, as shown in Figure 2.3. For all combination of inputs $f(x)$ and $g(x)$ to the black box, the output obeys the following linear superposition principles

$$\mathscr{H}\{f(x)+g(x)\} = \mathscr{H}\{f(x)\} + \mathscr{H}\{g(x)\} \tag{2.29}$$

for constants *a* and *b*, and where $\mathscr{H}\{f(x)\}$ represents the linear operation $\mathscr{H}$ on the input function $f(x)$.

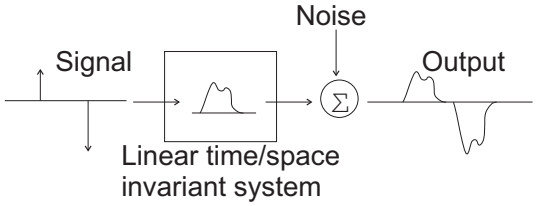


Figure 2.3 A linear system. The output for any fixed input is identical across all time. Scaling the input will also scale the output function identically.

A linear operation can be described by its response to an impulse input function or the kernel. In images, the impulse is equivalent to a point-source input, so the impulse response is also known as the point-spread-function (PSF). Let the impulse response of a system be characterised by $h(x;x')$, which represents the output due to an impulse input at x', or $\delta(x-x')$. The linearity of the operation means that any output $F(x)$ can be formed by the summed impulse response to its input $f(x)$, which can be decomposed into delta functions using the sifting property.

$$\begin{aligned} F(x) &= \mathscr{H}\{f(x)\} \\ &= \mathscr{H}\left\{\int_{-\infty}^{\infty} f(x')\delta(x-x')dx\right\} \\ &= \int_{-\infty}^{\infty} f(x')\mathscr{H}\left\{\delta(x-x')\right\}\, dx' \\ &= \int_{-\infty}^{\infty} f(x')h(x;x')\, dx' \end{aligned} \tag{2.30}$$

2.4.1 Linear shift invariant systems

In a special class of linear operations that are time or space invariant, the impulse response is

$$h(x,x') = h(x-x'),\ \forall x' \tag{2.31}$$

This shift invariance means that the output of the system at all times or spatial positions is the same, except for the shifted time or position.

$$f(x-x') \rightarrow F(x-x'),\ \text{for all functions}\, f(x) \tag{2.32}$$

In such systems, Equation 2.30 reduces to an operation known as convolution.

$$\begin{aligned} F(x) &= f(x) \odot h(x) = \int_{-\infty}^{\infty} f(x')h(x-x')\, dx' \\ F(x,y) &= f(x,y) \odot h(x,y) = \int_{-\infty}^{\infty} f(x',y')h(x-x',y-y')\, dx'\, dy' \end{aligned} \tag{2.33}$$

for either 1D - $f(x)$, or 2D signals $f(x,y)$.

Several properties of linear shift invariant systems, expressed using the convolution operator, are frequently used. They are the commutative, distributive, associative, shift-invariant, differentiation and the delta function identity properties.

Commutative

$$f(x) \odot g(x) = g(x) \odot f(x) \tag{2.34}$$

Distributive

$$[af(x) + bg(x)] \odot h(x) = a(f(x) \odot h(x)) + b(g(x) \odot h(x)) \tag{2.35}$$

Associative

$$(f(x) \odot g(x)) \odot h(x) = f(x) \odot (h(x) \odot g(x)) = f(x) \odot h(x) \odot g(x) \tag{2.36}$$

Shift-invariance

$$F(x) = f(x) \odot g(x) \rightarrow F(x - x') = f(x - x') \odot g(x) \tag{2.37}$$

Differentiation

$$\frac{d}{dx}(f(x) \odot g(x)) = \left(\frac{d}{dx} f(x)\right) \odot g(x) = f(x) \odot \left(\frac{d}{dx} g(x)\right) \tag{2.38}$$

Delta function

$$f(x) \odot \delta(x - x') = f(x - x') \tag{2.39}$$

Projection

$$\begin{aligned} \int_{-\infty}^{\infty} [a(x,y) \odot b(x,y)]\ dy &= \int_{-\infty}^{\infty} a(x,y)\ dy \odot b(x,y) \\ &= a(x,y) \odot \int_{-\infty}^{\infty} b(x,y)\ dy \end{aligned} \tag{2.40}$$

Generally, the convolution operation results in a smoothed output function. For images, this means that all imaging operations degrade the resolution of the transmitted image. In fact, in the limit, repeated convolution with random point-spread-functions results in Gaussian shaped images, a consequence quantified by the Central Limit Theorem in statistics.

Correlation

The correlation operation is mathematically similar to the convolution operation. It is used as a measure of the similarity (measured in the least-mean-square sense) between two functions, and forms the basis for matched filter designs, which "searches" for a signal template embedded within some signal. The displacement x that maximises the correlation is the position of the best match between $f(x)$ and $g(x)$.

$$f(x) \star g(x) = \int_{-\infty}^{\infty} f(x')\overline{g(x'-x)}\, dx' = f(x) \odot \overline{g(-x)} \tag{2.41}$$

The correlation operation is thus similar to the convolution operation, and is distributive and shift invariant, but not commutative or associative. The effect of the differentiation operator under correlation is

$$\frac{d}{dx}\left(f(x) \star g(x)\right) = -f(x) \star \left(\frac{d}{dx} g(x)\right) \tag{2.42}$$

2.4.2 Transforms

A special set of input functions to linear systems, known as the system eigenfunctions, have the property that they remain unchanged after being operated upon, only shifted in position and scaled in amplitude. The eigenfunctions of linear shift invariant systems are sinusoids. A linear operation can be described in terms of the amplitude and phase (position) shift imparted to sinusoids. This alternative description of the system is also known as the system transfer function $H(f)$, where

$$\mathscr{H}\{\sin(2\pi f x)\} = |H(f)| \sin(2\pi f x + \arg H(f)) \tag{2.43}$$

Using convolution, the impulse response fully describes a system. All input functions are decomposed into individual impulse functions, and passed through the system. The output of the system is the combination of all the scaled and shifted impulse responses.

Using the transfer function description, inputs to a linear system are broken down into sums of sinusoidal functions. The output from the system is the sum from the outputs of

the individual input sinusoidal components, as described by the transfer function.

$$\begin{aligned} \text{For } f(x) &= \sum_i A_i \sin(2\pi f_i x) \\ \mathscr{H}\{f(x)\} &= \mathscr{H}\left\{\sum_i A_i \sin(2\pi f_i x)\right\} \\ &= \sum_i |H(f_i)| A_i \sin(2\pi f_i x + \arg H(f_i)) \end{aligned} \tag{2.44}$$

This example illustrates an alternative description of linear systems by transforming the inputs and outputs into a different domain, presenting different views of the same data. Linear operators can also undergo transformations, and be described as operations on signals in the alternative domain. The system transfer function is the dual of the convolution operation. It is a powerful alternative for describing linear systems. The decomposition of a signal into sinusoidal waveforms is the basis of the Fourier transform.

2.4.3 Fourier transform

By decomposing a signal into its constituent frequencies, the Fourier transform converts a time or spatial waveform into a function in frequency space. The Fourier transform is similar in action to the prism in a spectrograph, which breaks down star-light into its constituent frequencies. The alternative representation provided by the transform is especially useful for understanding periodic signals.

While there is no standard notation for describing the Fourier transform, the notation in Goodman [38] is used in this thesis. For any well behaved function $g(x)$, which may be complex valued, there exists a unique Fourier transform

$$G(f_X) = \mathscr{F}\{g(x)\} = \int_{-\infty}^{\infty} g(x) e^{-i2\pi f_X x}\, dx \tag{2.45}$$

for spatial coordinates x and frequency along the x-axis f_X.

The Fourier transform can be extended to higher dimensions when transforming functions involving many variables or dimensions. They are separable into individual components along each (rectangular) axis. For example, the 2D Fourier transform can be expressed as

separate Fourier transforms along the x and y axes.

$$\begin{aligned} G(f_X, f_Y) &= \mathscr{F}\{g(x,y)\} \\ &= \int_{-\infty}^{\infty} \left(\int_{-\infty}^{\infty} g(x,y) e^{-i2\pi f_X x}\, dx \right) e^{-i2\pi f_Y y}\, dy \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x,y) e^{-i2\pi(f_X x + f_Y y)}\, dx\, dy \end{aligned} \tag{2.46}$$

with the corresponding frequency components f_X and f_Y.

The inverse Fourier transform recovers the original signal from its Fourier transform

$$\mathscr{F}^{-1}\{\mathscr{F}\{g(x,y)\}\} = \mathscr{F}\left\{\mathscr{F}^{-1}\{g(x,y)\}\right\} = g(x,y) \tag{2.47}$$

$$g(x,y) = \mathscr{F}^{-1}\{G(f_X, f_Y)\} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} G(f_X, f_Y) e^{i2\pi(f_X x + f_Y y)}\, df_X\, df_Y \tag{2.48}$$

for all continuous functions $g(x,y)$.

The forward and inverse Fourier transforms are very similar, differing only in the sign of the exponential phase term. The forward Fourier transform can thus be used instead of the inverse Fourier transform for recovering an image. In optical systems, this successive Fourier transform of an image results in inversion of the propagated image.

$$g(u,v) = \mathscr{F}\{\mathscr{F}\{g(x,y)\}\} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} G(f_X, f_Y) e^{-i2\pi(u f_X + v f_Y)}\, df_X\, df_Y = g(-x,-y) \tag{2.49}$$

Properties of the Fourier transform

By representing all signals as waves with frequency and phase, the Fourier transform is also useful for describing interference effects that commonly occur in the diffraction of light.

Various properties of the Fourier transform [38] are outlined below.

Linearity

$$\mathscr{F}\{af(x,y)+bg(x,y)\} = a\mathscr{F}\{f(x,y)\}+b\mathscr{F}\{g(x,y)\} \tag{2.50}$$

The Fourier transform is a linear transform. The addition operator in the spatial domain corresponds to the addition operator in the Fourier domain.

Scale

For $F(f_X, f_Y) = \mathscr{F}\{f(x,y)\}$

$$\mathscr{F}\{f(ax,by)\} = \frac{1}{|ab|}F(\frac{f_X}{a},\frac{f_Y}{b}) \tag{2.51}$$

Scaling the spatial coordinates results in an inverse scale of the corresponding frequency.

Shift - Exponential phase

For $F(f_X, f_Y) = \mathscr{F}\{f(x,y)\}$

$$\mathscr{F}\{f(x-a,y-b)\} = F(f_X, f_Y)e^{-i2\pi(f_X a+f_Y b)} \tag{2.52}$$

A spatial shift in the spatial domain results in an exponential phase factor in the Fourier domain.

Convolution and multiplication

$$\mathscr{F}\{f(x,y)\odot g(x,y)\} = \mathscr{F}\{f(x,y)\}\mathscr{F}\{g(x,y)\} \tag{2.53}$$

$$\mathscr{F}\{f(x,y)g(x,y)\} = \mathscr{F}\{f(x,y)\}\odot\mathscr{F}\{g(x,y)\} \tag{2.54}$$

The correspondence between convolution and multiplication, as hinted at the beginning of the section, is an important one. Equation 2.53 provides the description for the system transfer function $\mathscr{F}\{g(x,y)\}$ given the point-spread-function $g(x,y)$.

Correlation

The correlation operator is similar to a convolution, and from Equation 2.53, can be expressed as

$$\mathscr{F}\{f(x,y)\star g(x,y)\} = \mathscr{F}\left\{f(x,y)\odot\overline{g(-x,-y)}\right\} = F(f_X,f_Y)\overline{G(f_X,f_Y)} \tag{2.55}$$

The special case of auto-correlation reduces to

$$|F(f_X,f_Y)|^2 = \mathscr{F}\{f(x,y)\star f(x,y)\} \tag{2.56}$$

The squared magnitude of the Fourier transform of a function is also known as the power spectrum or the spectral density of the function. The power spectrum is a real quantity which shows the breakdown of the signal power within each frequency.

Rayleigh and Parseval's Theorem (Conservation of Energy)

In a new twist to Pythagoras' theorem, the total energy in a signal is preserved during the Fourier transform. In physical situations, $|f(x,y)|^2$ might represent the power density within a telescope aperture (integrated to give the total power or intensity), while $|F(f_X,f_Y)|^2$ would represent the propagating power density spread over various directions.

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}|f(x,y)|^2\,dx\,dy = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}|F(f_X,f_Y)|^2\,df_X\,df_Y \tag{2.57}$$

Differentiation

$$\mathscr{F}\left\{\frac{d}{dx}f(x)\right\} = i2\pi f_X F(f_X) \tag{2.58}$$

Under differentiation, the Fourier transform of a function is multiplied by the frequency.

Rotational symmetry in the Fourier transform

For the special case of rotationally symmetric functions, the Fourier transform exhibits some surprising and useful properties. Consider a rotationally symmetric signal $f(x,y)$ that only has a radial r dependence,

$$f(x,y) = f(\sqrt{x^2+y^2}) = f(r) \tag{2.59}$$

Due to the rotational symmetry of its Fourier transform, we employ a rectangular to polar coordinates transform in the spatial and frequency domain, that is, from (x,y) and (f_X, f_Y) to (r,θ) and (ρ,ϕ). The Fourier transform is

$$\begin{aligned}
F(f_X, f_Y) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)e^{-i2\pi(f_X x + f_Y y)}\,dx\,dy \\
F(\rho,\phi) &= \int_0^{2\pi}\int_0^{\infty} f(r)e^{-i2\pi(\rho\cos\phi r\cos\theta + \rho\sin\phi r\sin\theta)} r\,dr\,d\theta \\
&= \int_0^{\infty} r f(r)\int_0^{2\pi} e^{-i2\pi\rho r\cos(\phi-\theta)}\,d\theta\,dr \\
&= \int_0^{\infty} r f(r)\int_0^{2\pi} \cos(2\pi\rho r\cos(\phi-\theta)) - i\sin(2\pi\rho r\cos(\phi-\theta))\,d\theta\,dr \\
F(\rho) &= \int_0^{\infty} 2\pi r f(r) J_0(2\pi r\rho)\,dr
\end{aligned} \tag{2.60}$$

using the identities

$$\int_0^{2\pi} \cos(2\pi\rho r\cos(\phi-\theta))\,d\theta = 2\pi J_0(2\pi r\rho) \tag{2.61}$$

from Equation 2.13, and

$$\int_0^{2\pi} \sin(2\pi\rho r\cos(\phi-\theta))\,d\theta = 0 \tag{2.62}$$

due to the odd-symmetry of the sine function.

Conveniently, the 2D Fourier transform can be reduced to a 1D transform with the zeroth order Bessel function of the first kind as a kernel. Known as the Hankel transform or the Fourier-Bessel transform, the rotationally symmetric Fourier transform inherits some of properties of the 2D Fourier transform (subject to the symmetry constraint). Defining the Hankel transform as

$$\mathscr{H}\{f(r)\} = F(\rho) = 2\pi \int_0^\infty r f(r) J_0(2\pi r\rho)\, dr \tag{2.63}$$

we obtain the following properties.

$f(r)$	$F(\rho) = \mathscr{H}\{f(r)\}$
$f(ar)$	$\frac{1}{a^2}F(\frac{\rho}{a})$
$f(r) \odot g(r)$	$F(\rho)G(\rho)$
$r^2 f(r)$	$-\boldsymbol{\nabla}^2 F(\rho)$

Table 2.1 Properties of the Hankel Transform.

The Jinc function

In this thesis, the Fourier transform of $\mathrm{circ}(x,y)$ is often required. Being a circularly symmetric function, we can use the Hankel transform to simplify the problem.

$$\begin{aligned} \mathscr{F}\{\mathrm{circ}(x,y)\} &= \mathscr{H}\{\mathrm{rect}(r)\} \\ &= 2\pi \int_0^R r J_0(2\pi r\rho)\, dr = \frac{R J_1(2\pi\rho R)}{\rho} \\ &= 4R^2 \mathrm{Jinc}(2R\rho) \end{aligned} \tag{2.64}$$

where $J_1(x)$ and $\mathrm{Jinc}(x)$ are defined in Equation 2.16 and Equation 2.17, and the radius $R = 1$. Here, we also used the identity $\int_0^x x' J_0(x')\, dx' = x J_1(x)$.

The result from Equation 2.64 actually corresponds to the equation used to describe the optical field in the imaging plane of a telescope with an un-aberrated, circularly symmetric aperture[2].

[2]The propagation of light can be described with a Fourier transform. The optical properties of the Fourier

Fourier transforms of common functions

The Fourier transforms of some commonly used functions have some useful properties, and merit some attention.

$f(x)$	$F(f_X) = \int_{-\infty}^{\infty} f(x)e^{-i2\pi f_X x}\,dx$
$\delta(x)$	1
$\mathrm{Rect}(x)$	$\mathrm{Sinc}(f_X)$
$\mathrm{U}(x)$	$\frac{1}{i2\pi f_X}$
$\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{x^2}{2\sigma^2}}$	$e^{-2\pi^2\sigma^2 f_X^2}$
$\cos(2\pi f_0 x)$	$0.5\delta(f_X + f_0) + 0.5\delta(f_X - f_0)$
$\mathrm{circ}(\sqrt{x^2+y^2})$	$4\mathrm{Jinc}(\frac{\sqrt{f_X^2+f_Y^2}}{2})$

Table 2.2 Table of Fourier Transform pairs of commonly used functions.

Interestingly, the transform of the Gaussian function, is also a Gaussian function[3]. The Gaussian function is simple to specify and intuitively satisfying as a blurring function in images. Additionally Fourier analysis of images is helped by both the function and transform being real. The Fourier Transform of a Gaussian can be derived from the identity

$$\int_{-\infty}^{\infty} e^{-cx^2}\,dx = \sqrt{\frac{\pi}{c}} \tag{2.65}$$

the Fourier transform of the function e^{-cx^2} is

transform will be examined in Section 3.3.2.

[3]An easy way to account for the scale factors is to consider $\mathscr{F}\left\{\mathrm{Gaussian}\{0,\sigma^2\}(x)\right\} \propto \mathrm{Gaussian}\{0,\frac{1}{\sigma^2}\}(u) = ke^{-\frac{u^2}{\left(2\frac{1}{\sigma^2}\right)}}$ for $u = 2\pi f$ (radians). The Fourier transform must equal 1 at $u = 0$ (the DC term), so the scale factor, k, for the exponential, must be 1.

$$\begin{aligned}\mathscr{F}\left\{e^{-cx^2}\right\} &= \int_{-\infty}^{\infty} e^{-cx^2} e^{-i2\pi fx}\, dx \\ &= e^{\left(\frac{i\pi f}{\sqrt{c}}\right)^2} \int_{-\infty}^{\infty} e^{\left(\sqrt{c}x+\frac{i\pi f}{\sqrt{c}}\right)^2}\, dx \\ &= \sqrt{\frac{\pi}{c}} e^{-\frac{\pi^2 f^2}{c}} \end{aligned} \tag{2.66}$$

Signal representation and the Discrete Fourier Transform

The continuous function transforms are useful as a mathematical aid in the analysis of continuous signals. In practice, signals are frequently measured or sampled at discrete times and recorded or quantised as discrete values. For this, the continuous Fourier transform is re-framed as a discrete transform. We must first examine the properties of discrete signals.

Sampling

In imaging applications, a square array of intensity detectors, such as the CCD or CMOS detector, records intensity falling on the detectors at regular intervals. Each sample of the signal is measured over the area covered by each detector. A convenient approximation for sampled signals assumes that the original signal is sampled point-wise by multiplication with a regularly spaced array of delta functions. For a 1D signal, this is

$$\begin{aligned} f_S(x) &= \sum_{n=-\infty}^{\infty} f(x)\delta(x-n\Delta x) \\ &= f(x)\mathrm{comb}_{\Delta x}(x) \end{aligned} \tag{2.67}$$

This is in fact a notational convenience, and represents a simplification of the more rigorous representation $f_S(n\Delta x) = \int_{-\infty}^{\infty} f(x)\delta(x-n\Delta x)\, dx$.

As an example, consider the class of all sinusoids sampled at intervals of Δx. Any signal can be recovered from its samples exactly by fitting a sinusoid to the sampled points. However, multiple solutions are possible - for $sin(2\pi fx)$ sampled at f_S, there are an infinite number of solutions of the form

$$sin(2\pi f'x) \text{where} f' = f \pm n f_s \tag{2.68}$$

for all integers n.

In general, the effect of sampling on a signal's spectrum can be found (using the property that the comb function is self-similar under the Fourier transform).

$$\begin{aligned} \mathscr{F}\{f_S(x)\} &= \mathscr{F}\{f(x)\text{comb}_{\Delta x}(x)\} \\ &= F(f_X) \odot \text{comb}_{F_s}(f_X) \end{aligned} \tag{2.69}$$

The convolution of the signal transform with the periodic array of deltas is shown for a band-limited signal in Figure 2.4. The sampling frequency is inversely proportional to the spacing of the samples $F_s = \frac{1}{\Delta x}$. Here, each "island" of spectra is an exact copy of the next, and provides no additional information. At lower frequencies, the spectra of the sampled signal may start to overlap, resulting in aliasing, which interferes with interpretation of the signal. Provided the sampling frequency is high enough, a good representation of the original signal is recorded, and no information is lost.

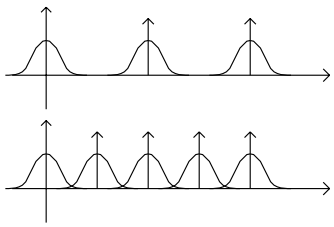


Figure 2.4 The effect of sampling (with frequency F_s) on a band-limited signal. At lower sampling frequencies, some parts of the signal spectra may overlap.

Nyquist sampling criterion

The sampling frequency required to sample a signal without aliasing depends on the signal to be sampled. Rapidly changing signals need to be sampled at a higher rate compared to slowly changing signals. From Equation 2.68 and Figure 2.4, the lowest sampling frequency, known as the Nyquist frequency, has to be two times the highest frequency present in the signal.

Discrete Fourier Transform

Having defined a representation for discrete signals, we can define the Discrete Fourier Transform (DFT). The DFT of a signal $f[n] = f_S(n\Delta x)$ with N total samples is given by

$$F[k] = \mathscr{F}\{f[n]\} = \frac{1}{N}\sum_{n=0}^{N-1} f[n]e^{-i2\pi k\frac{n}{N}} \tag{2.70}$$

for integers $0 \leq k, n < N$.

The equivalent matrix formulation, representing with vectors $\boldsymbol{F}_n = F[n]$ and $\boldsymbol{f}_n = f[n]$, is

$$\boldsymbol{F} = \boldsymbol{M}\boldsymbol{f} \tag{2.71}$$

where $\boldsymbol{M}_{nm} = e^{-i\frac{2\pi nm}{N}}$

The basis vectors in $\boldsymbol{M}$ are orthogonal with respect to each other, and normalised. From the properties of orthogonal matrices [8], the inverse Fourier transform matrix $\boldsymbol{M}^{-1} = \boldsymbol{M}^* = \overline{\boldsymbol{M}^T} = \overline{M}$, or

$$f[n] = \sum_{k=0}^{N-1} F[k]e^{i2\pi n\frac{k}{N}} \tag{2.72}$$

The signal representation in both the time and frequency domain is discrete and finite. Aside from the discreteness of the signals, the properties of the DFT (Section 2.4.3) are similar to the continuous Fourier Transform. However, the signal and spectra are additionally implicitly assumed to be periodic, so $f[n+N] = f[n]$ and $F[k+N] = F[k]$. In practice, this periodicity assumption leads to discontinuities between the beginning and end of sampled signals, as shown in Figure 2.5.

Figure 2.5 The assumption of periodicity leads to discontinuities in sampled signals.

This periodicity also affects the discrete convolution operation. To extend the convolution operation to discrete signals, we require the discrete convolution operation to be

$$h[n] = f[n] \odot g[n] = \sum_{n'=0}^{2N-1} f[n']g[n-n'] \tag{2.73}$$

for $0 \leq n < 2N$, with $g[n-n'] = 0$ when $n-n' < 0$.

The indirect convolution operation, where $f[n]$ and $g[n]$ are transformed into the frequency domain, multiplied (Equation 2.53), and inverse transformed back to the time domain again, results in the circular convolution operation

$$h[n] = f[n] \odot g[n] = \sum_{n'=0}^{N-1} f[n']g[n-n'] \tag{2.74}$$

for $0 \leq n < N$, and with wrap around (due to periodicity) $g[n-n'] = g[N+n-n']$ when $n-n' < 0$.

To obtain the more useful convolution defined in Equation 2.73, the signals $f[n]$ and $g[n]$ should in general be zero-padded to double their original sizes. The effect of "circularity" from convolution in the Fourier domain is still present, but the separation between the periodic signals now removes any overlap when convolving. This effect is the dual of the aliasing problem when the repeated (periodic) signal spectra of under-sampled signals overlap. This requirement to zeropad signals also applies when measuring a signal's spectral density, since the squared magnitude requires multiplication in the Fourier domain of a signal with itself, and corresponds to a correlation operation in the spatial domain.

In imaging application, for 2D images sampled over a square grid array, the 2D DFT is separable into 2 1D transforms and is straight-forward to compute given the 1D DFT. Other sampling strategies are also available in 2D, (for example rectangular grids, or hexagonal patterns) but are not considered in this thesis.

Fast Fourier Transform

For an N-point signal, the DFT is formed from an NXN matrix multiplication, and requires N^2 operations. For signals sampled over a long time (large N), the computational costs of the DFT become prohibitive. An optimisation, called the Fast Fourier transform (FFT) [15] is available for speeding up calculations of the discrete signal spectrum. The FFT is strictly a computational optimisation, and otherwise produces identical results to the DFT. First, the signal is split into two half-period components

$$\begin{aligned}\sum_{n=0}^{N-1} f[n]e^{-i2\pi k\frac{n}{N}} &= \sum_{n=0}^{\frac{N}{2}-1} f[n]e^{-i2\pi k\frac{n}{N}} + \sum_{n=\frac{N}{2}}^{N-1} f[n]e^{-i2\pi k\frac{n}{N}} \\ &= \sum_{n=0}^{M-1} f[n]e^{-i2\pi\frac{k}{2}\frac{n}{M}} + \sum_{m=0}^{M-1} f[m+M]e^{-i2\pi\frac{k}{2}(\frac{m+M}{M})} \\ &= \sum_{p=0}^{M-1} \left(f[p] + e^{-i2\pi\frac{k}{2}} f[p+M]\right) e^{-i2\pi\frac{k}{2}\frac{p}{M}} \end{aligned} \tag{2.75}$$

The DFT of the N-point signal can now be decomposed into 2 DFT's of 2 $\frac{N}{2}$-point signals. For integer $0 \leq k' < M$ and for even $k = 2k'$

$$F[k] = \sum_{p=0}^{M-1} (f[p] + f[p+M])\, e^{-i2\pi k'\frac{p}{M}} \tag{2.76}$$

whereas for odd $k = 2k'+1$

$$F[k] = \sum_{p=0}^{M-1} \left((f[p] - f[p+M])e^{i2\pi\frac{p}{M}}\right) e^{-i2\pi k'\frac{p}{M}} \tag{2.77}$$

This division of a problem in two smaller sub-problems after N steps results in an algorithmic complexity of $N\log N$ compared to the N^2 of the naive matrix multiplication method. With the discovery of the FFT, Fourier analysis became a convenient and practical tool that found widespread use.

Aside from the complex exponential basis functions, other similar basis functions like the Hadamard basis functions and the discrete cosine functions (used in the jpeg image encoding standard) may also be used for representing discrete transforms of signals. Other more general transforms like the wavelet transform and the Gabor transform are commonly used in image processing, but have found no application in this thesis.

2.4.4 Zernike polynomials

Depending on the particular geometry of the functions being transformed, a different set of bases functions may be used. The Zernike polynomials are a set of functions defined on

a circle of radius 1 [65]. Due to their rotational invariance and circular support, they are traditionally used for describing optical aberrations in optical instruments. More recently, they are used to describe aberrations in the human eye, and also those resulting from atmospheric turbulence. Their use in digital watermarking of images has also been suggested.

They are defined in polar coordinates as products of radial $R_n^m(r)$ (where r is the radius) and angular functions (sin and cos terms of the azimuthal angle θ). The Zernike polynomials on the unit circle ($r \leq 1$) are defined as

$$Z_i(r,\theta) = \begin{cases} \sqrt{n+1}R_n^0(r) & \text{if m = 0,} \\ \sqrt{n+1}R_n^m(r)\sqrt{2}\cos(m\theta) & \text{if } m \neq 0\text{, and i is even,} \\ \sqrt{n+1}R_n^m(r)\sqrt{2}\sin(m\theta) & \text{if } m \neq 0\text{, and i is odd,} \end{cases} \tag{2.78}$$

where

$$R_n^m(r) = \sum_{s=0}^{\frac{n-m}{2}} \frac{(-1)^s(n-s)!}{s![\frac{n+m}{2}-s]![\frac{n-m}{2}-s]!} r^{n-2s} \tag{2.79}$$

for non-negative integral values of n and m, with $m \leq n$ and $n-|m|$ being even. i represents the mode ordering number for the polynomials, and follows the numbering convention used by Noll [65].

The lower order Zernike polynomials loosely corresponds to the classical Seidel aberrations for describing imperfect optical systems. These are shown in Figure 2.6 with their corresponding names.

Unlike the Seidel aberrations, the Zernike polynomials form a complete set of orthogonal bases functions over the unit circle.

$$\int_0^{2\pi}\int_0^{\infty} Z_i(r,\theta)Z_j(r,\theta)A(r,\theta)\, r dr\, d\theta = \begin{cases} 0\ \forall i \neq j \\ 1\ \forall i = j \end{cases} \tag{2.80}$$

where $A(r,\theta)$ is the aperture weighting function ($\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} A(r,\theta)dx\,dy = 1$), being $A(r,\theta) = \frac{1}{\pi}$ within the unit circle, and 0 everywhere else.

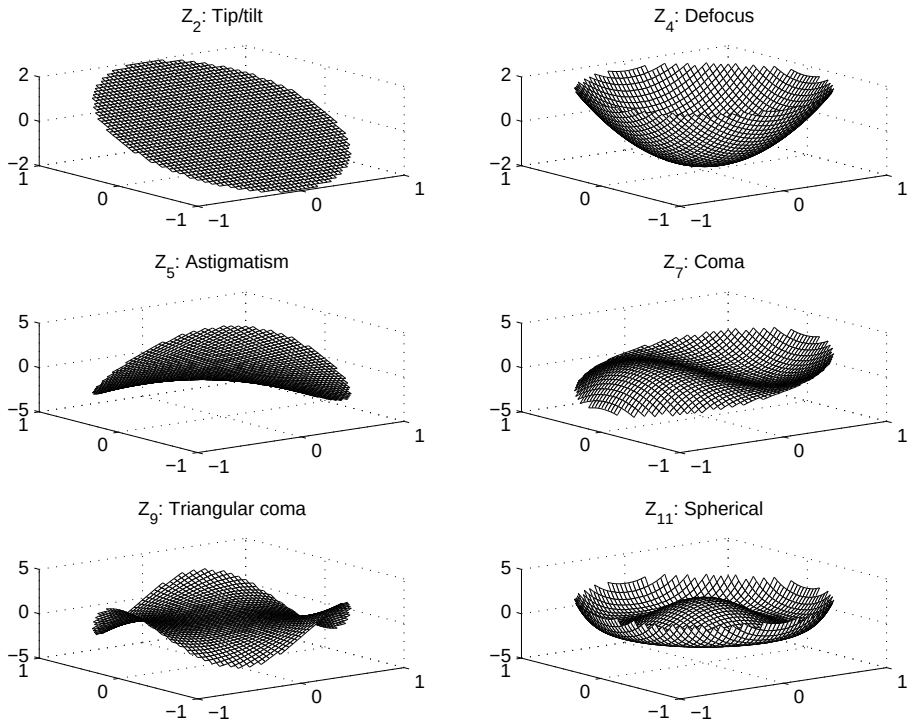


Figure 2.6 The Zernike polynomials and their closest corresponding Seidel aberrations.

The property of orthogonality is convenient for treating the modes separately without having to balance the aberration terms, as for the Seidel aberrations in classical optics. Since they form a complete set, the Zernike polynomials can represent any arbitrary phase function over a unit circle with a weighted sum

$$\phi(r,\theta) = \sum_{i=1}^{\infty} \alpha_i Z_i(r,\theta) \tag{2.81}$$

with $\boldsymbol{\alpha}$, the vectorised form of all coefficients α_i, being sufficient to describe the phase.

The orthogonality of the Zernike polynomials conserves the energy of the phase in the weighted sum representation.

$$\int_0^{2\pi} \int_0^{\infty} \phi(r,\theta)^2 A(r,\theta)\, r dr\, d\theta = \sum_{i=1}^{\infty} \alpha_i^2 \tag{2.82}$$

The Zernike polynomials also possess the property of rotational invariance. As represented

using Equation 2.81, rotating any arbitrary function preserves the energy in the Zernike modes at each radial order and azimuthal frequency. Equation 2.85 shows that after rotating through any arbitrary angle ψ, the energy present in each radial order and azimuthal frequency remain constant.

If

$$\phi(r,\theta) = \sum_{i=0}^{\infty} \alpha_i Z_i(r,\theta) \tag{2.83}$$

and

$$\phi(r,\theta+\psi) = \sum_{i=0}^{\infty} \alpha_i' Z_i(r,\theta) \tag{2.84}$$

then the coefficients are the "same", in the sense that pairs of the Zernike coefficients within the same radial order contain the same amount of energy

$$\sum_{i\in S_{n,m}} \alpha_i^2 = \sum_{i\in S_{n,m}} \alpha_i'^2 \;\; \forall n,m \tag{2.85}$$

where $S_{n,m}$ refers to the set of all Zernike modes with radial order n and azimuthal frequency m.

That is, for a fixed n and m, $R_n^m(r)cos(m(\theta+\psi))$, the sine and cosine terms are

$$\begin{aligned} Z_c(r,\theta) &= R_n^m(r)\cos(m(\theta+\psi)), \text{and} \\ Z_s(r,\theta) &= R_n^m(r)\sin(m(\theta+\psi)) \end{aligned} \tag{2.86}$$

Their corresponding coefficients

$$\alpha_s^2 + \alpha_c^2 = \text{const} \tag{2.87}$$

for any arbitrary rotation ψ, holding n and m constant.

For numerical simulations, discrete versions of the Zernike polynomials are required. However, unlike well-known transforms like the Discrete Fourier Transform or the Discrete Cosine Transform, there are no discrete orthogonal basis functions to represent the Zernike polynomials. We are therefore limited to a discrete approximation of the Zernike polynomials.

2.5 Probability and statistics

The field of optical imaging inherently deals with random statistical phenomena. From the unknown light source, through the random transmission medium, to the detection and measurement of light, a statistical treatment is required to quantify the randomness and uncertainty of the whole system. We shall describe the optical imaging problem using a probabilistic framework.

Probability

Probability is used to describe chance or random events. The Theory of Probability was given a mathematical foundation in the mid-17th century by correspondences between the mathematicians Blaise Pascal and Pierre Fermat. The probability or likelihood of an event is measured using a real number ranging from 0 to describe events that will not occur to 1 to describe events that are certain to occur. In addition to the law governing mutually exclusive events, these three axioms form the fundamental basis for probability

$$\begin{aligned} &0 \leq P(A) \leq 1 \\ &P(S) = 1 \implies \text{S is certain to occur} \\ &P(A_1 \cup A_2) = P(A_1) + P(A_2) \text{ for mutually exclusive events } A_1 \text{ and } A_2 \end{aligned} \tag{2.88}$$

As an example, consider the probability of obtaining a certain face up when throwing a 6 sided die. If each face is *just as likely* as any other face to appear facing upwards when the die is thrown, the probability of a successful throw is $\frac{1}{6}$.

$$P(A) = \lim_{N \to \infty} \frac{n}{N} \tag{2.89}$$

This assigns a numerical value to events in terms of their frequency of appearance in the long run. It is intuitively satisfying, and also obeys the basic axioms of probability.

Another example involving discrete probabilities is the photon count measurement in an imaging process. The behaviour of photon arrival obeys Poisson statistics, and this phenomenon is particularly significant at low light levels. This is an example of how probabilistic frameworks are used for describing measurement uncertainty, noise or random signals. The probability of obtaining a photon count x for a light detector is given by

$$P(x) = \frac{e^{-\mu}\mu^x}{x!} \tag{2.90}$$

where μ is the expected (average) photon count for the detector over many experiments. A Poisson distribution with a high mean value can be approximated using Gaussian white noise for analysis purposes.

The probability distribution functions for these two different types of random phenomena are shown in Figure 2.7.

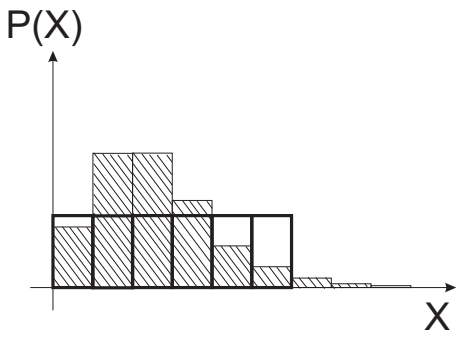


Figure 2.7 Two different types of probability distribution functions taking discrete values.

The concept of probability also extends to continuous variables. In this thesis, the wavefront slope of atmospheric turbulence is assumed to take on random values over time, averaging around 0. In fact, the probability distribution function for the wavefront slope is Gaussian, so the probability density of the wavefront slope being x is

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{x^2}{2\sigma^2}} \tag{2.91}$$

where the variance σ^2 quantifies the spread (width) of the distribution.

The distribution function is plotted in Figure 2.8. When describing probabilities of continuous variables, the probability of any specific slope x is a density value. Integrating the density function over a range of values provides a numerical probability value, so we measure probabilities over a range of slopes instead. The shaded area under the curve in Figure 2.8 is the probability that an observed wavefront has a slope that lies within the shaded range.

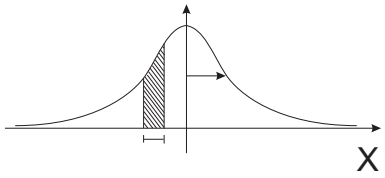


Figure 2.8 The bell-shaped Gaussian or normal probability distribution function.

Moments of a distribution

Often, when describing a probability distribution function, instead of providing the whole probability distribution function in minute detail, we are only interested in a few of the more important features, like the general shape or position of the distribution. The moments of a distribution often provide a concise and mathematically convenient description of the distribution. The first moment of a distribution is the mean of the distribution.

$$\langle X \rangle = \int_{-\infty}^{\infty} x p(x)\, dx \tag{2.92}$$

where $\langle X \rangle$ is a shorthand for the expected value of the random variable X.

The variance is described by the second moment of a distribution. The higher moments are given by

$$\langle X^n \rangle = \int_{-\infty}^{\infty} x^n p(x)\, dx \tag{2.93}$$

To fully specify many distributions, only the lowest moments are required. For example, the Gaussian distribution is specified by its mean and variance, and the Poisson distribution is specified by its mean.

Characteristic functions and Fourier Transforms

Another common transformation of the probability distribution function is taking its characteristic function,

$$\phi_X(v) = \left\langle e^{ivX} \right\rangle \tag{2.94}$$

This is the similar to the Fourier transform, and in fact, represents an alternative representation of the PDF in a different domain. The exponential (on the right hand side) is a sum of all powers of X, so the characteristic function is effectively a weighted sum of all the moments of the distribution.

The properties of the Fourier transform apply to the characteristic function. For example, the distribution of the sum of two random variables X and Y is the convolution of their respective distribution functions. The characteristic function is the product of their individual characteristic functions.

$$\phi_{X+Y}(v) = \phi_X(v)\phi_Y(v) \tag{2.95}$$

As another example, knowing the Fourier transform of the Gaussian function, we can find the characteristic function of a normal distribution. Given the distribution

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{x^2}{2\sigma^2}} \tag{2.96}$$

and its Fourier transform (see Equation 2.66)

$$P(f) = e^{-2\pi^2 f^2 \sigma^2} \tag{2.97}$$

the characteristic function of a Gaussian probability distribution function is given by

$$\begin{aligned}\left\langle e^{icX}\right\rangle &= \int_{-\infty}^{\infty} p(x)e^{icx}\,dx \\ &= P\left(f=-\frac{c}{2\pi}\right) \\ &= e^{-\frac{c^2\sigma^2}{2}}\end{aligned} \tag{2.98}$$

This identity is also useful for expressing various quantities like the Strehl ratio or the telescope optical transfer function in terms of the phase structure function. These quantities will be examined later in Section 4.1 and Chapter 4.

Distributions of multiple variables

When dealing with multiple random variables, the moments of a probability distribution can be extended to describe the interaction between variables - how do two variables change together (does one increase while another decreases?). The most used measure of the relationship between a pair of linear variables is their correlation. The correlation coefficient between two variables X and Y with joint distribution $p(x,y)$ is given by

$$\frac{\langle (X-\bar{X})(Y-\bar{Y})\rangle}{\sigma_X\sigma_Y} = \frac{\langle XY\rangle - \bar{X}\bar{Y}}{\sigma_X\sigma_Y} \tag{2.99}$$

The numerator, $\langle (X-\bar{X})(Y-\bar{Y})\rangle$, known as the covariance, is an extension of the variance measure of a single variable. These quantities are multivariate extensions of moments as defined by

$$\langle XY\rangle = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xyp(x,y)\,dx\,dy \tag{2.100}$$

A correlation coefficient of 1 describes a linear increasing relationship between two variables, while -1 describes a decreasing relationship. If the two variables are independent, then their correlation coefficient is 0 (however, if two variables have a correlation coefficient of 0, no conclusion on their independence may be drawn).

The joint probability distribution function of independent variables is the product of their individual (marginal) probability distribution functions. The multivariate Gaussian distri-

bution is useful for illustrating the case when variables may not be independent.

Multivariate Gaussian distributions

A direct extension of Equation 2.91 to problems involving multiple independent and identically distributed Gaussian variables is given in Equation 2.101.

$$p(\boldsymbol{x}) = \frac{1}{\sqrt{2\pi\sigma^2}^N} e^{-\frac{1}{2\sigma^2}\boldsymbol{x}^T\boldsymbol{x}} \tag{2.101}$$

Here, $p(\boldsymbol{x})$ is a single-valued probability distribution function, which is dependent on many input variables, here represented as a vector $\boldsymbol{x}$. It is a product of the marginal distributions of all the individual variables. In general, these variables might not be independent, nor would they be identically described by the same mean and variance. In such cases, the more general expression for a multi-variate Gaussian distribution is

$$p(\boldsymbol{x}) = \frac{1}{\sqrt{(2\pi)^N |\boldsymbol{C}|}} e^{-\frac{1}{2}\boldsymbol{x}^T\boldsymbol{C}^{-1}\boldsymbol{x}} \tag{2.102}$$

where $\boldsymbol{C} = \left\langle \boldsymbol{x}\boldsymbol{x}^T \right\rangle$, and $|\boldsymbol{C}|$ is its determinant. Without loss of generality, we have also assumed that the mean of all variables are 0.

The covariance matrix $\boldsymbol{C}$ describes the correlation between the variables, and can be diagonalised with a singular value decomposition. This corresponds to a coordinate transformation of the $\boldsymbol{x}$ vector, so the new coordinate axes now represent independent variables.

2.5.1 Random signals and random processes

A random signal is sequence of random variables over time or space. A random or stochastic process describes a set of (or an ensemble of) space/time varying signals. Random processes are random in the sense that repeated experiments will give rise to different outcomes - a signal taking on random temporal or spatial values. A probability distribution is defined to describe the chance of observing any function from the sample space.

The theory of random processes can be used to model the wavefront aberrations caused by the atmosphere. In the absence of any prior knowledge about the atmosphere, the pressure, temperature and humidity in the atmosphere can be modelled as a random function that

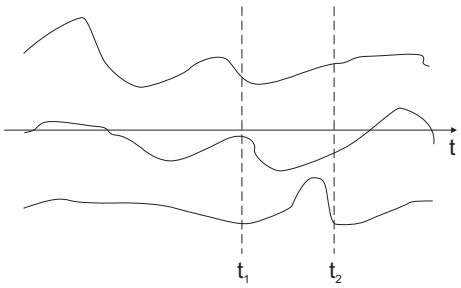


Figure 2.9 The values taken by these random functions at times t_1 and t_2 are described by random variables. Just like random variables, we can examined their statistics and correlation with each other over time.

changes over space and time. The resulting optical aberrations are the result of combining many random processes. As long as the underlying random processes have finite variance, the final statistical behaviour of their sum obeys the Gaussian distribution. However, in adaptive optics, the variance of the phase piston term caused by atmospheric turbulence has an infinite variance. Fortunately, the piston term is not measurable and is usually removed during calculations, so the phase statistics can be modelled using Gaussian distributions.

Stationary and non-stationary signals

A random signal may have signal statistics that remain constant over time. This is referred to as strict sense stationarity. The mean and variance of the signal value at all times is a constant. A looser restriction, that the signal has a constant mean, and auto-correlation that is dependent only on the time/position difference, gives us the larger set of wide sense stationary processes.

The covariance function of the signal is defined to be

$$B(t,t') = \big\langle (f(t) - \langle f \rangle)(f(t+t') - \langle f \rangle) \big\rangle \tag{2.103}$$

where $\langle f \rangle$ is the mean signal value (time independent).

For stationary signals, there is no t dependence, and for wide sense stationary signals, only a t' dependence, $B(t,t') = B(t')$. The signal variance at time t corresponds to $B(t,0)$. When the mean $\langle f \rangle$ is 0, we have the auto-correlation function $B(t,t') = \langle f(t) f(t+t') \rangle$.

The statistics of atmospheric turbulence change wildly over large distances or time scales, but are approximately stationary over smaller distances and time scales. It can be described

using the model of wide sense stationary signals.

Structure function

The covariance function is undefined for some functions. For example, there is no meaningful absolute value for the aberration phase function, which may be infinite depending on the optical model used. The atmospheric phase structure function, which uses a relative phase difference, is substituted instead. It is defined as

$$D_\phi(\boldsymbol{x}') = \left\langle (\phi(\boldsymbol{x}) - \phi(\boldsymbol{x}+\boldsymbol{x}'))^2 \right\rangle \tag{2.104}$$

This phase structure function is frequently used as a placeholder for the mathematical manipulation of the phase covariance function using

$$D_\phi(\boldsymbol{x}') = 2B_\phi(\boldsymbol{0}) - 2B_\phi(\boldsymbol{x}') \tag{2.105}$$

Power spectra of random signals

As shown in Equation 2.56, the power spectral density of a function is given by the Fourier transform of its auto-correlation function. More generally, for wide sense stationary random processes, which may not be square integrable (undefined Fourier transform), the same relationship exists. This is known as the Wiener-Khintchine or the Khintchine-Kolmogorov theorem. We can use this to analyse of the power spectra of atmospheric turbulence, which is a random process with fractal-like properties.

Power densities of fractals

Using the Wiener-Khintchine theorem, random fractals can have a defined power spectra. The self-similarity or scaling of fractals means that their spectra must possess certain properties. Consider a random fractal process $f(x)$ which is self-similar to $\frac{1}{r^H} f(rx)$ when scaled by r, with $(0 < H < 1)$ being the fractal Hurst dimension, which is a measure of the self-similarity of fractals [71]. The power spectral density (from its Fourier transform) is also self-similar under scaling. Defining

$$\begin{aligned} F(f_X) &= \mathscr{F}\{f(x)\} \\ \mathscr{F}\{g(x)\} = \mathscr{F}\left\{\frac{1}{r^H} f(rx)\right\} &= \frac{1}{r^{H+1}} F\left(\frac{f_X}{r}\right) \end{aligned} \tag{2.106}$$

Being the same fractal, their power spectra (with appropriate matching of scale) are equal

$$\begin{aligned} \frac{1}{r} P_g(f_X) &= P_f(f_X) \\ r\frac{1}{r^{2H+2}} \left| F\left(\frac{f_X}{r}\right)\right|^2 &= |F(f_X)|^2 \\ \frac{1}{r^{2H+1}} P_f\left(\frac{f_X}{r}\right) &= P_f(f_X) \end{aligned} \tag{2.107}$$

The power law obeyed by $P_f(f_X) \propto (f_X)^k$ solves to $k = -(2H+1)$. It is interesting to compare this to the Kolmogorov power law (to be explained in Chapter 4) which exhibits $k = -\frac{11}{3}$ so its Hurst dimension is $H = \frac{4}{3}$.

2.5.2 Bayesian estimation

Conditional probability

The conditional probability of an event A given that another event B has occurred is denoted by $P(A|B)$. For example, when throwing two dice, the *a priori* probability of obtaining a sum of 4 is $\frac{1}{12}$. However, if we know that one of the dice has landed with a 2 facing up, then the probability of obtaining a sum of 4, that is, of obtaining a 2 on the second dice, becomes $\frac{1}{6}$. Had we obtained a 5 on one die, we would have been able to say that regardless of the outcome of the second dice, the probability of obtaining a sum of 4 is 0 (not possible). Knowledge of the outcome of one event sometimes allows us to make better estimates of the probability of a second related event.

The relationship between the conditional probability and joint probability of two events are given by

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \tag{2.108}$$

In our previous example, let A represents the event "obtaining a sum of 4", B_1 the event "obtaining 2 on the first die", and B_2 the event "obtaining a 5 on the second die".

$$P(A|B_1) = \frac{P(A \cap B_1)}{P(B_1)} = \frac{P(X)P(B_1)}{P(B_1)} = \frac{\frac{1}{36}}{\frac{1}{6}} = \frac{1}{6} \tag{2.109}$$

where X refers to the event "obtaining a 2 on the second die", with the outcome of the first and the second die being independent events. Similarly, $P(A \cap B_2) = 0$, so $P(A|B_2) = 0$.

Reversed conditional probability

Reversing the example, if we are given A (sum of dice = 4), and need to determine the probabilities of each outcome on the second die (X) without any prior knowledge of B (the outcome of the first die), we will need

$$\begin{aligned} P(X|A) &= \frac{P(X \cap A)}{P(A)} = \frac{P(X)P(B = A - X)}{P(A)} \\ &= \begin{cases} \frac{\frac{1}{6} \times \frac{1}{6}}{\frac{1}{12}} & \text{for } x = 1,2,3 \\ 0 & \text{for } x = 4,5,6 \end{cases} \end{aligned} \tag{2.110}$$

Often, the "reversed" conditional probability of Equation 2.110 is easier to derive from the "forward" conditional probability $P(A|X)$ using

$$P(X|A) = \frac{P(X \cap A)}{P(A)} = \frac{P(X \cap A)}{P(X)} \frac{P(X)}{P(A)} = \frac{P(A|X)P(X)}{P(A)} \tag{2.111}$$

Extended to continuous random functions, this forms the basis for Bayesian estimation using noisy measurements.

Maximum likelihood and Maximum A Posteriori estimation

Bayesian estimation is used for estimation from noisy measurements by taking into account noise statistics. A common example in this thesis is the linear problem

$$\boldsymbol{d} = \boldsymbol{H}\boldsymbol{\alpha} + \boldsymbol{n} \tag{2.112}$$

where $\boldsymbol{\alpha}$ is some quantity (to be estimated) producing a noisy signal $\boldsymbol{d}$ through the linear process $\boldsymbol{H}$. The noise statistics is known in advance, and frequently represent either white noise $P(n_i) = \frac{1}{\sqrt{2\pi}\sigma_{n_i}} e^{-\frac{n_i^2}{2\sigma_{n_i}^2}}$, or photon noise, in which case $P(d_i) = \frac{e^{-\mu_i}\mu_i^{d_i}}{d_i!}$, for the Poisson mean and variance μ being the expected value of $\boldsymbol{d}$, and $\boldsymbol{d}_i = (\boldsymbol{H}\boldsymbol{\alpha})_i$.

The estimate for $\boldsymbol{\alpha}$ is denoted by $\hat{\boldsymbol{\alpha}}$, and under maximum likelihood estimation, is found by maximising the likelihood

$$\ln P(\hat{\boldsymbol{\alpha}}|\boldsymbol{d}) = \ln P(\boldsymbol{d}|\hat{\boldsymbol{\alpha}}) + \ln P(\hat{\boldsymbol{\alpha}}) - \ln P(\boldsymbol{d}) \tag{2.113}$$

The reversed form of the conditional probability is often easier to derive from the statistics of the noise. For uncorrelated white noise

$$\begin{aligned}
\ln P(\boldsymbol{d}|\hat{\boldsymbol{\alpha}}) &= \ln P(\boldsymbol{n} = \boldsymbol{d} - \boldsymbol{H}\hat{\boldsymbol{\alpha}}) \\
&= \ln\left(\Pi_i \frac{1}{\sqrt{2\pi}\sigma_{n_i}} e^{-\frac{n_i^2}{2\sigma_{n_i}^2}}\right) \\
&= \sum_i \ln\frac{1}{\sqrt{2\pi}\sigma_{n_i}} - \frac{n_i^2}{2\sigma_{n_i}^2} \\
\Longrightarrow &\sum_i -\frac{(d_i - (\boldsymbol{H}\hat{\boldsymbol{\alpha}})_i)^2}{2\sigma_{n_i}^2}
\end{aligned} \tag{2.114}$$

The *a priori* likelihood function $\ln P(\hat{\boldsymbol{\alpha}})$ describes our prior estimate for the likelihood of the quantities to be estimated. Frequently, no prior assumption of the likelihood of any particular solution is made (uniform distribution). This corresponds to the maximum likelihood solution, where only the first term of Equation 2.113, as shown in Equation 2.114,

is used. The *a priori* likelihood of $\hat{\boldsymbol{\alpha}}$ is independent of $\hat{\boldsymbol{\alpha}}$ and can be ignored.

The third term of Equation 2.113 is always ignored, as it has no dependence on $\hat{\boldsymbol{\alpha}}$.

For the specific case of random Gaussian noise, Gaussian priors, and vector valued quantities of N measurements, the prior distribution is (refer Equation 2.102)

$$P(\boldsymbol{\alpha}) = \frac{1}{\sqrt{(2\pi)^N |\boldsymbol{C_\alpha}|}} e^{-\frac{1}{2}\boldsymbol{\alpha}^T \boldsymbol{C_\alpha}^{-1} \boldsymbol{\alpha}} \tag{2.115}$$

the noise is

$$P(\boldsymbol{n}) = \frac{1}{\sqrt{(2\pi)^N |\boldsymbol{C_n}|}} e^{-\frac{1}{2}\boldsymbol{n}^T \boldsymbol{C_n}^{-1} \boldsymbol{n}} \tag{2.116}$$

and the solution is

$$\hat{\boldsymbol{\alpha}} = (\boldsymbol{H}^T \boldsymbol{C_n}^{-1} \boldsymbol{H} + \boldsymbol{C_\alpha}^{-1})^{-1} \boldsymbol{H}^T \boldsymbol{C_n}^{-1} \boldsymbol{d} \tag{2.117}$$

The inverse is more conveniently represented[4] with fewer matrix inversions [56] as

$$(\boldsymbol{H}^T \boldsymbol{C_n}^{-1} \boldsymbol{H} + \boldsymbol{C_\alpha}^{-1})^{-1} \boldsymbol{H}^T \boldsymbol{C_n}^{-1} = \boldsymbol{C_\alpha} \boldsymbol{H}^T (\boldsymbol{H} C_\alpha \boldsymbol{H}^T + C_n)^{-1} \tag{2.118}$$

The maximum likelihood solution is a special case, where $\boldsymbol{C_\alpha}$ is ignored because it has no effect on the solution. When the noise covariance is the identity matrix (independent and identically distributed across all measurements), the solution is

$$\hat{\boldsymbol{\alpha}} = (\boldsymbol{H}^T \boldsymbol{H})^{-1} \boldsymbol{H}^T \boldsymbol{d} \tag{2.119}$$

This corresponds to the least squares error minimisation problem. Using a Bayesian framework, we see that the intuitive notion of least squares data fitting is based on several as-

[4] Although this equivalence identity requires great revelation to infer, its proof, with hindsight, is simple. Pre-multiplying by $\boldsymbol{H}^T \boldsymbol{C_n}^{-1} \boldsymbol{H} + \boldsymbol{C_\alpha}^{-1}$ and post-multiplying by $\boldsymbol{H} C_\alpha \boldsymbol{H}^T + \boldsymbol{C_n}$ on both sides result in equivalence.

sumptions (prior information, noise model) that are otherwise implicit. Aside from image processing, the Bayesian reasoning technique is also used in a wide range of statistics based problems like belief and inference systems, control theory, and modelling. It is an intuitive yet formal and practical tool for reasoning with randomness or uncertainty.

2.5.3 Information Theory

In 1948, Claude Shannon [90] proposed a quantity that he termed entropy for measuring the "rate" of information production. A random source of information is assumed to produce N discrete symbols with probabilities p_i for $1 < i < N$. The entropy measure, H, of this source has to satisfy three conditions.

1. H is continuous in p_i,

2. When all p_i's are identical, H increases monotonically with increasing N,

3. If the information source is combined from multiple sources S_i, then the total entropy is a weighted sum of the individual entropies of the information sources. The weights are proportional to the probability of obtaining each subset of symbols, $H = \sum P(S_i)H_i$.

Shannon showed that the only valid formula for H is proportional to $\sum p_i \ln p_i$. However, entropy is not the only possible formulation for measuring information. In the field of statistical estimation, another quantity known as the Fisher information [30, 51] is used to measure the information content of continuous random distributions.

Fisher information

To understand the Fisher information of a random distribution, we begin with the parameter estimation problem. We are often interested in the mean, variance or some other parameter characterising a probability distribution.

For a probability distribution $p_\theta(\boldsymbol{x})$ or equivalently $p(\boldsymbol{x}|\theta)$ parametrised by an unknown θ, an estimate $\hat{\theta}(\boldsymbol{x})$ is obtained by observing the outcomes $\boldsymbol{x}$ drawn from the distribution. This estimate can be considered to be a random variable. Its mean and variance are given by

$$\left\langle \hat{\theta}(\boldsymbol{x}) \right\rangle = \int_{-\infty}^{\infty} \hat{\theta}(\boldsymbol{x}) p(\boldsymbol{x})\, d\boldsymbol{x} \tag{2.120}$$

and

$$\mathrm{var}\left\{\hat{\theta}(\boldsymbol{x})\right\} = \left\langle (\hat{\theta}(\boldsymbol{x}) - \theta)^2 \right\rangle \tag{2.121}$$

Estimators with low variances are generally better than those with higher variances. When the estimator is an unbiased estimator, $\left\langle \hat{\theta}(\boldsymbol{x}) \right\rangle = \theta$. The minimum variance unbiased estimator, or MVU, is frequently used as an optimality criterion in statistical estimation. The minimum lower bound on the variance of unbiased estimators is given by the Cramer-Rao lower bound (CRLB)[5]. Using the Cauchy-Schwarz inequality,

$$\begin{aligned}
\left\langle (\hat{\theta} - \theta)\frac{\partial}{\partial \theta} \ln p(\boldsymbol{x}|\theta) \right\rangle &\leq \left\langle (\hat{\theta} - \theta)^2 \right\rangle \left\langle \left[\frac{\partial}{\partial \theta} \ln p(\boldsymbol{x}|\theta)\right]^2 \right\rangle \\
\left\langle \hat{\theta}\frac{\partial}{\partial \theta} \ln p(\boldsymbol{x}|\theta) \right\rangle - \theta \left\langle \frac{\partial}{\partial \theta} \ln p(\boldsymbol{x}|\theta) \right\rangle &\leq \mathrm{var}\left\{\hat{\theta}\right\} J \\
\mathrm{var}\left\{\hat{\theta}\right\} &\geq \frac{\left\langle \hat{\theta}\frac{\partial}{\partial \theta} \ln p(\boldsymbol{x}|\theta) \right\rangle - 0}{J} = \frac{1}{J}
\end{aligned} \tag{2.122}$$

where J is the Fisher information[6]. Two equivalent forms for J are

$$J = \left\langle \left[\frac{\partial}{\partial \theta} \ln p(\boldsymbol{x}|\theta)\right]^2 \right\rangle = -\left\langle \frac{\partial^2}{\partial \theta^2} \ln f(\boldsymbol{x}|\theta) \right\rangle \tag{2.123}$$

The Fisher information is a measure of the "spread" of the probability distribution function. The larger the spread in the distribution function, the more variable the outcomes of the random process, and subsequently, the higher the variance of the estimator. The Fisher information can also be interpreted as a measure of how much information is obtained from each observation of a random event.

In general, there is no known mechanical procedure for deriving minimum variance unbiased estimators. However, in linear processes, an efficient estimator (one that achieves the

[5]The CRLB only applies to unbiased estimators. Biased estimators can potentially achieve lower estimator variances.

[6]$s = \frac{\partial}{\partial \theta} \ln p(x|\theta)$ is also known as the score function and is similar in form to entropy and the log-likelihood function.

CRLB) is often straight-forward to derive. For some problems, an efficient estimator may not exist so the minimum variance unbiased estimator does not achieve the CRLB.

Multiple parameter estimation

The Cramer-Rao Lower Bound for estimators with multiple (vector) parameters $\boldsymbol{\theta}$ is similar to the scalar case

$$\boldsymbol{J} = \left\langle \boldsymbol{s}(\boldsymbol{\theta},\boldsymbol{X})\boldsymbol{s}^T(\boldsymbol{\theta},\boldsymbol{X}) \right\rangle = -\left\langle \frac{\partial}{\partial \boldsymbol{\theta}} \left(\frac{\partial}{\partial \boldsymbol{\theta}} \ln f(\boldsymbol{x}|\boldsymbol{\theta}) \right) \right\rangle \tag{2.124}$$

or

$$J_{ij} = \left\langle \frac{\partial \ln p(\boldsymbol{x}|\boldsymbol{\theta})}{\partial \theta_i} \frac{\partial \ln p(\boldsymbol{x}|\boldsymbol{\theta})}{\partial \theta_j} \right\rangle = -\left\langle \frac{\partial^2 \ln p(\boldsymbol{x}|\boldsymbol{\theta})}{\partial \theta_i \partial \theta_j} \right\rangle \tag{2.125}$$

The variance of the *i*-th parameter is given by

$$\mathrm{var}\left\{ \hat{\boldsymbol{\theta}}_i \right\} \geq \left(\boldsymbol{J}^{-1} \right)_{ii} \tag{2.126}$$

Chapter 3

Optics

Today, optical systems like telescopes, microscopes and spectrographs are commonly used for scientific observations and measurements. Their invention arose from the needs of astronomical observations, and experiments by Galileo, Newton, Huygens, Hooke, and others on the nature of colour and light in the 16^{th} to 17^{th} century.

From everyday experience, it is obvious that light rays travel in straight lines, and upon meeting an obstruction, will cast a shadow. The path of these light rays can be modified by shaped and optically active materials like mirrors, prisms and lenses, to form telescopes and microscopes.

However, light had also been observed to possess wave-like properties. Hooke had suggested a wave theory of light as early as 1665, while Huygens published a description on the propagation of wavefronts in 1678. In 1803, Young provided conclusive evidence of interference in light, demonstrating in sunlight, with “a slip of card”, the light and dark fringes resulting from light cancellation.

In this section, we introduce the theory of light and provide some examples of optical systems and their usage.

3.1 Geometric optics

The theory of geometric optics assumes that light travels in a straight line. Light travels from a light source in a straight line, and stops when absorbed by any object in their path,

leaving dark regions of shadows behind the object. The direction of the light rays can be changed using mirrors, lenses, and prisms.

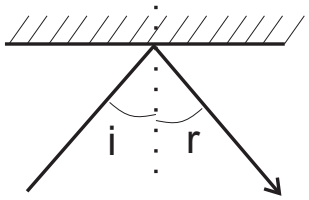


Figure 3.1 Reflection of a light ray along the plane of propagation (plane of incidence).

Figure 3.1 shows a light ray reflecting off a mirror at the same angle r as the incident angle i. The incident and reflected angles are usually defined with respect to the mirror normal, which is the dotted line perpendicular to the mirror surface at the point of reflection. The angle of reflection rule also applies to curved mirror surfaces, where the normal is perpendicular to the mirror surface.

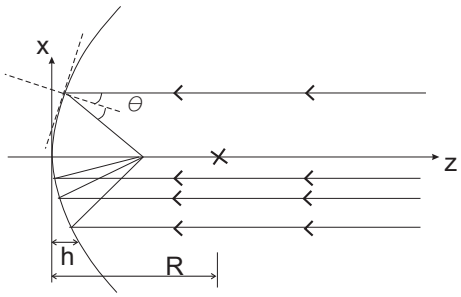


Figure 3.2 Light from a distant object reflecting off a curved mirror surface. The mirror curvature (and the corresponding shorter focal distance) is shown exaggerated here for illustration.

This is shown in Figure 3.2, where a spherical mirror focuses parallel light rays from a distant object onto a point (the focus) at the optical axis. For a mirror with radius of curvature R, the height of the mirror surface $h(x)$ and its slope $h_x(x)$ are

$$\begin{aligned} h(x) &= R - \sqrt{R^2 - x^2} \\ \tan\theta &= h_x(x) = \frac{x}{\sqrt{R^2 - x^2}} \end{aligned} \tag{3.1}$$

where θ is the angle (in radians) between the mirror normal to the horizontal (and light ray).

In optical systems analysis, the paraxial approximation for ray tracing is commonly used. It assumes light rays that are close to the optical axes throughout the optical system and small light ray angles. The small ray angles can then be approximated to first order by

$$\begin{aligned} \sin\theta &\approx \theta \\ \cos\theta &\approx 1 \\ \tan\theta \approx \theta &\approx \frac{x}{R} \end{aligned} \tag{3.2}$$

where θ is the ray angle (in radians) with the paraxial axis.

Every ray intersects the optical axis at

$$f = h(x) + x\tan(2\theta) \approx h(x) + \frac{x}{2\tan\theta} \approx h(x) + \frac{R}{2} \approx \frac{R}{2} \tag{3.3}$$

By concentrating diffused light rays onto a single point, the spherical mirror forms an image of the distant object at the focus. In practice, image detectors need to be placed out of the way of the incoming light, so additional reflectors are used to redirect the light.

An alternative to imaging using reflection from mirrors, is to use refraction through transparent materials. Refraction occurs when light passes from one medium into another medium, changing its speed, and direction.

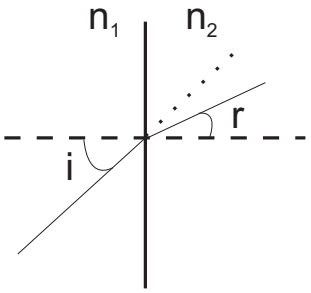


Figure 3.3 Refraction of a light ray at the boundary of two transparent materials.

Here, the incident and refracted angles are defined with respect to the normal at the boundary between the two media. Figure 3.3 shows a light ray changing its direction after entering the second medium. The change in direction is given by Snell's law,

$$\begin{aligned} \frac{\sin\theta_1}{v_1} &= \frac{\sin\theta_2}{v_2} \\ n_1 \sin\theta_1 &= n_2 \sin\theta_2 \end{aligned} \tag{3.4}$$

where n_i, the refractive index for a medium, is defined to be $\frac{c}{v_i}$, the ratio between the speed of light in vacuum to the speed of light in medium i.

A related property of refraction is dispersion, which results from wavelength-dependence in the refractive index. Light from different wavelengths or colours is refracted by different amounts, separating the components of light. This creates the colours in rainbows, and is used in prisms for spectrography.

Refraction in lenses

Similar to mirrors, lenses are used to create an inverted and scaled image of distant objects using refraction. Unlike mirrors, lenses transmit light, so the optical axis is not folded or mirrored, allowing images to be formed along the optical axis without obscuration of the aperture from other optical elements.

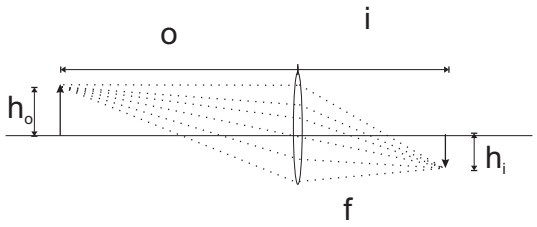


Figure 3.4 Imaging an object at o with a lens of focal length f. The real image i is rotated and scaled by the imaging operation.

Figure 3.4 shows the transmissive lens imaging an object at distance o (in contrast, Figure 3.2 is equivalent to imaging an object at infinity). The distances of the object and image from the lens are determined by the thin lens approximation equation. Aside from sign changes, this equation applies identically to the optical analysis of both reflective mirrors and transmissive lenses.

$$\frac{1}{f} = \frac{1}{i} + \frac{1}{o} \tag{3.5}$$

The image magnification $M = \frac{h_i}{h_o}$ is a function of the object distance and lens focal length

$M = \frac{f}{o-f}$. The ratio of the focal length to the lens diameter (not shown in the figure), is also known as the F-number or the F-ratio.

$$F = \frac{f}{D} \tag{3.6}$$

The F-number is a measure of the effect of the optical system on light. Larger F-numbers indicate that light is bent more passing through the system, and as a rule of thumb, suffers from more aberrations. Aberrations in optical systems are also caused by imperfections in the lens shape, lens surface, and off-axis imaging, and is explained more in subsequent sections.

3.1.1 Optical path length

The speed of light in any media is slower than speed of light in vacuum, so the refractive index is always greater than 1. Because light can travel at different speeds across different media, it is convenient to measure the path length travelled as an equivalent distance in vacuum. The optical path length through a medium is the same distance travelled in vacuum in the same time period. Figure 3.5 shows 3 different media with different refractive indices.

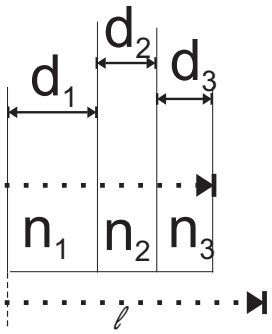


Figure 3.5 The light path length for 3 different transparent media compared to the path length of light in a vacuum.

The total time taken by light to travel through all three layers are given by $c\left(\frac{d_1}{v_1} + \frac{d_2}{v_2} + \frac{d_3}{v_3}\right)$, or $\sum_{i=1}^{3} n_i d_i$. In general the optical path length for any media is found by integrating the refractive index along the light path

$$\int_0^L n(l)dl \tag{3.7}$$

starting from 0 and ending at L.

The optical path length is the uniform measure of distance for light in different types of media. The laws of reflection and refraction can be derived from the principle of least action, or in optics, the principle of shortest optical path. Given a set of paths between two points, that path taken by the light ray is the path that takes the least amount of time (has the shortest optical path). Figure 3.6 illustrates the principle for a straight path, a reflected path (off the mirror), and on the right, a refracted path through two different media.

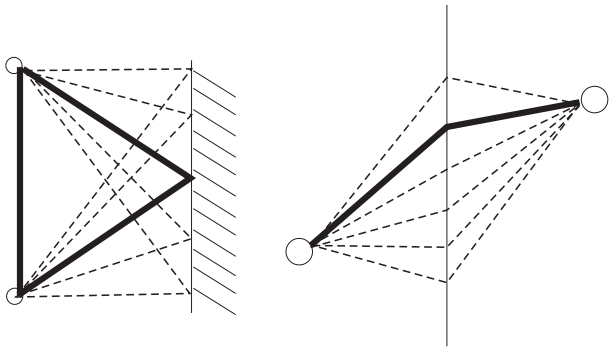


Figure 3.6 The path between two points "chosen" by a beam of light is the one with the shortest total optical path length.

3.1.2 Wavefront

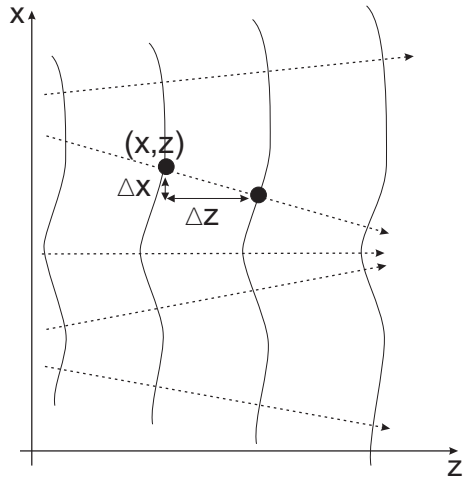


Figure 3.7 Propagation of wavefront along the z-axis.

The propagation of light rays away from an object can also be described using light wavefronts. The wavefront is a surface of constant optical path length from a common source. At any point, the direction of wavefront propagation is perpendicular to the wavefront slope at that point, as shown in Figure 3.7. Here, an analogy can be drawn with surface water

waves, where the direction of wave travel corresponds to light rays, and ripples correspond to wavefronts.

At any point (x,z) within the propagation region, the phase of the complex field $\phi(x,z)$ can be found from the wavefront, $\phi(x,z) = kW(x,z)$. This wavefront function represents the optical path length (in distance units) from the $z = 0$ plane. The relative advance or retardation of the different light rays across the plane of propagation is found from their wavefront differences.

Given a wavefront $W(x,z)$ propagating in the z-direction, the Wavefront Transport Equation is found from

$$\begin{aligned} W(x+\Delta x, z+\Delta z) &= W(x,z) + \sqrt{\Delta x^2 + \Delta z^2} \\ W(x,z) + W_x(x,z)\Delta x + W_z(x,z)\Delta z &\approx W(x,z) + \Delta z\left(1 + \frac{1}{2}\left(\frac{\Delta x}{\Delta z}\right)^2\right) \\ W_z(x,z) &= 1 - \frac{1}{2}W_x(x,z)^2 \end{aligned} \tag{3.8}$$

where $W_z(x,z)$ and $W_x(x,z) = \frac{\Delta x}{\Delta z} \ll 1$ are the wavefront derivatives along the z and x-axes.

The first term of Equation 3.8 is due to the increasing optical path length as the wavefront travels, and doesn't affect the direction that light travels in. Changes in direction are caused by the second wavefront slope term. The effects of diffraction on the transport equation are given in Section 3.4,

Using the concept of wavefronts, the effects of reflection and refraction, in changing the direction of light, can be described as modifications to the wavefront. Active optical surfaces modify the direction of light rays, so the equivalent changes to the wavefront, as shown in Figure 3.8, can be inferred.

Similar to Equation 3.1, the presence of a quadratic term in the wavefront corresponds to a focusing action. Imperfections in optical systems result in deviations in the wavefront from the quadratic shape. These imperfections, known as optical aberrations, cause blurring in images. Aberrations cannot be avoided completely, but through good design, can be minimised. The analysis of optical systems involves the adjustment of the shapes and positions of lenses to optimise for the conflicting requirements for image position and magnifica-

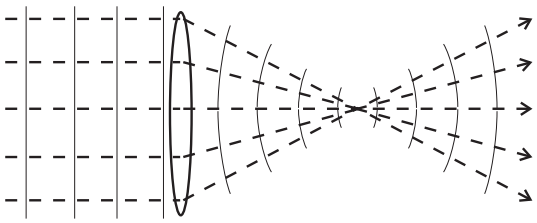


Figure 3.8 Effect of optical lenses on the wavefront.

tion, width of the field of view, aperture size (brightness), and minimisation of aberrations (image blurring).

3.2 Optical analysis

It is conventional to use right-handed coordinate axes in an optical system and shown in Figure 3.9. The wavefront shown in Figure 3.9 also has a negative curvature.

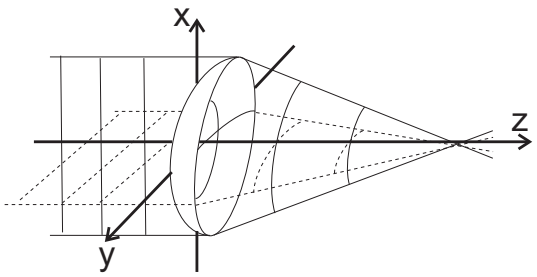


Figure 3.9 The right-handed Cartesian coordinates conventionally used in optical analysis. Light is shown travelling from the left to the right along the optical (z) axis.

3.2.1 Geometric optics

The ray-tracing equations of Equation 3.9 and Equation 3.10 [83] use geometric optics to describe light rays. They always travel in a direction perpendicular to the local wavefront slope.

$$x'(x,y) = x + zW_x(x,y,0) \tag{3.9}$$

$$y'(x,y) = y + zW_y(x,y,0) \tag{3.10}$$

where (x', y', z) represents the location of the ray within the x-y plane at z, from the ray starting at $(x, y, 0)$.

The intensity at any point in the propagation path is given by the light ray density through that point. Figure 3.10 shows the propagation of a wavefront with a uniform negative curvature (constant $W_{xx} = -2a$, $W_{yy} = -2b$). The intensity along the optical axis is inversely related to the cross-sectional area (shown in rectangles) of the light beam. Relative to the intensity before propagation, $I(z = 0)$,

$$\begin{aligned} I(z)A(z) &= I(0)A(0) = ID_xD_y \\ I(z) &= \frac{ID_xD_y}{d_x(z)d_y(z)} \\ &= \frac{I}{1 + z(W_{xx} + W_{yy}) + z^2W_{xx}W_{yy}} \end{aligned} \tag{3.11}$$

where $A(z)$ is the cross-sectional area of the propagating beam, with dimensions $d_x(z)$ and $d_y(z)$ ($D_x = d_x(0)$, $D_y = d_y(0)$) as shown in Figure 3.10.

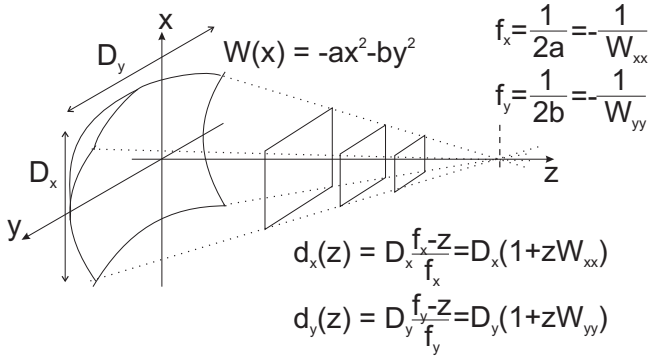


Figure 3.10 Changes in intensity over distance due to a wavefront curvature. The presence of a negative curvature ($a \geq 0$, and $b \geq 0$) focuses incoming light rays, resulting in a brightening in the intensity.

Since the intensity is determined by the local wavefront curvature, we can estimate the wavefront curvature from changes in the intensity after propagating the wavefront. In fact, this method forms the basis for the curvature wavefront sensor, to be explained in later sections.

In the general case, for wavefronts with different curvature-axis orientations, the intensity is given by Equation 3.12.

$$I(x',y',z+\Delta z) = \frac{I(x,y,z)}{1+H(x,y,z)\Delta z + K(x,y,z)\Delta z^2} \tag{3.12}$$

for $H(x,y,z) = \nabla^2 W(x,y,z) = W_{xx}(x,y,z) + W_{yy}(x,y,z)$, the Laplacian or the mean curvature of the wavefront and $K(x,y,z) = W_{xx}(x,y,z)W_{yy}(x,y,z) - W_{xy}(x,y,z)^2$, the Gaussian curvature of the wavefront. The mean and Gaussian curvatures are defined as the mean (up to a scale factor) and product of the principal curvatures W_{uu} and W_{vv} for u and v lying along the axes of the principal curvatures.

3.2.2 Seidel aberrations

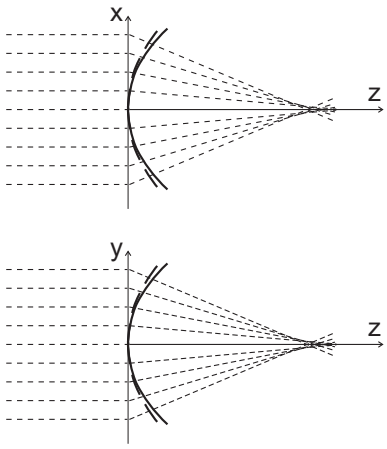
Traditionally, optical aberrations are classified according to their polynomial expansion. The classical Seidel aberrations are third order approximations to wavefronts, with five known aberrations, namely spherical, coma, astigmatism, curvature of field, and distortion. The wavefront shape and corresponding effect on image is shown in Figure 3.11 as ray-intercept diagrams, another tool commonly used to describe aberrations.

3.3 Diffraction

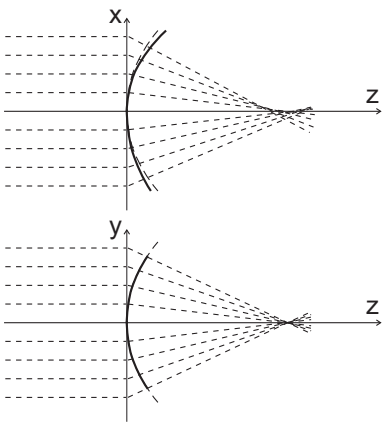
Under geometric optics, a perfect lens would focus light from distant point sources down to a point. This image is infinitesimally small, and infinitely bright. Clearly, this is impossible, and shows that geometric optics is merely an approximation. In fact, there is a lower limit to the size of the point source image, determined by diffraction effects. Diffraction refers to the behaviour of light not predicted by geometric optics.

For a full description of light, we begin with the foundation for electromagnetism, Maxwell's Equations. These four equations unify electric and magnetic field theory.

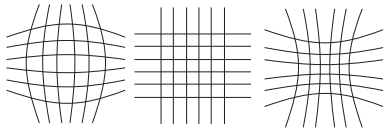
$$\begin{aligned} \nabla \times \boldsymbol{E} &= -\mu \frac{\partial \boldsymbol{H}}{\partial t} \\ \nabla \times \boldsymbol{H} &= \varepsilon \frac{\partial \boldsymbol{E}}{\partial t} \\ \nabla \cdot \varepsilon \boldsymbol{E} &= 0 \\ \nabla \cdot \mu \boldsymbol{H} &= 0 \end{aligned} \tag{3.13}$$



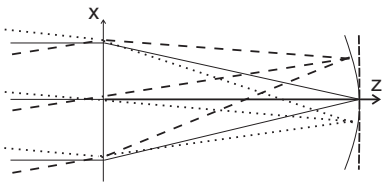
(a) Rays with different heights come to focus at varying distance(spherical)



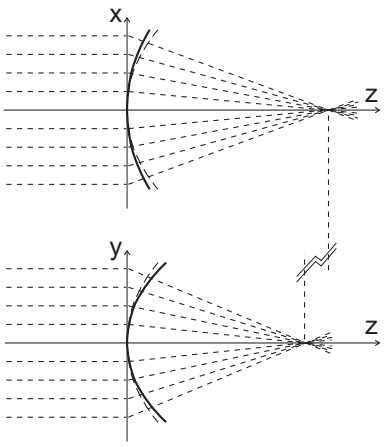
(b) Rays from different sides come to focus at different distances(coma)



(c) Position dependent deformation of a grid-line image.(distortion)



(d) Curved imaging plane.(field/Petzval curvature)



(e) Rays at different orientations come to focus at different focal distances.(astigmatism)

Figure 3.11 The Seidel aberrations.

where E and H are the electric and magnetic field vectors respectively, and ε and μ are the medium permittivity and permeability respectively (in vacuum, they are denoted by ε_0 and μ_0). The permittivity and permeability of free space (vacuum) is linked to the speed of light in a vacuum by $c = \frac{1}{\sqrt{\mu_0 \varepsilon_0}}$.

For a linear, isotropic, homogeneous and non-dispersive propagation medium, E and H have identical forms.

$$\begin{aligned} \nabla^2 \boldsymbol{E} - \frac{n^2}{c^2}\frac{\partial^2 \boldsymbol{E}}{\partial t^2} = 0 \\ \nabla^2 \boldsymbol{H} - \frac{n^2}{c^2}\frac{\partial^2 \boldsymbol{H}}{\partial t^2} = 0 \end{aligned} \tag{3.14}$$

where $n = \sqrt{\frac{\varepsilon}{\varepsilon_0}}$ is the refractive index.

Both E and H are symmetrical in all vector components, so only a single scalar equation suffices for expressing all components.

$$\nabla^2 u - \frac{n^2}{c^2}\frac{\partial^2 u}{\partial t^2} = 0 \tag{3.15}$$

where u may be any of $\boldsymbol{E}_x$, $\boldsymbol{E}_y$, $\boldsymbol{E}_z$, $\boldsymbol{H}_x$, $\boldsymbol{H}_y$, or $\boldsymbol{H}_z$.

Although this breaks down when the medium is inhomogeneous, or anisotropic (for example, boundary conditions imposed by obstructions), the scalar diffraction approximation remains useful and accurate when the diffracting structures are large, and diffraction angles are kept small.

3.3.1 Scalar diffraction theory

The scalar field of Equation 3.15 is a space and time varying quantity

$$u(x,y,z,t) = A(x,y,z)cos(2\pi ft + \phi(x,y,z)) \tag{3.16}$$

where f, A and ϕ are the wave frequency, amplitude and phase respectively.

It can be also be represented in a phasor or complex form.

$$U(x,y,z) = A(x,y,z)e^{i\phi(x,y,z)} \tag{3.17}$$

so $u(x,y,z,t) = U(x,y,z)e^{-i2\pi ft}$.

Using the phasor representation, Equation 3.15 becomes the Helmholtz equation [38],

$$\nabla^2 U + k^2 U = 0 \tag{3.18}$$

for $k = \frac{2\pi}{\lambda}$, where the wavelength $\lambda = \frac{c}{nf}$.

Using Green's Theorem from calculus, and the Green's function $G(r) = \frac{e^{ikr}}{r}$, a few formulations for the diffraction equation have been proposed.

$$\iiint_V \left[U\nabla^2 G - G\nabla^2 U\right] dv = \iint_S \left[U\frac{\partial G}{\partial n} - G\frac{\partial U}{\partial n}\right] ds \tag{3.19}$$

for a volume V and surface S, where $\frac{\partial}{\partial n}$ is a partial derivative on the surface in the normal outward direction.

When both U and G obey the Helmholtz equation,

$$\iint_S \left[U\frac{\partial G}{\partial n} - G\frac{\partial U}{\partial n}\right] ds = \text{const} \tag{3.20}$$

taking the limit in the volume around the point of interest (ξ,η,z) allows us to find the field there in terms of the field specified by an enclosing surface. This is the integral theorem of Helmholtz and Kirchhoff.

$$4\pi U(\xi,\eta,z) = \iint_{S_P} \left[U\frac{\partial G}{\partial n} - G\frac{\partial U}{\partial n}\right] ds = \iint_S \left[U\frac{\partial G}{\partial n} - G\frac{\partial U}{\partial n}\right] ds \tag{3.21}$$

for S_P being the surface enclosing the point (ξ,η,z), and S being the surface containing some input field.

The Rayleigh-Sommerfeld diffraction equations are derived from different choices for Green's

function in the integral. The first and second Rayleigh-Sommerfeld solutions are

$$U(\xi,\eta) = \frac{1}{i\lambda}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} U(x,y)\frac{e^{ikr}}{r}\cos\theta\; dx\, dy \tag{3.22}$$

$$U(\xi,\eta) = \frac{1}{2\pi}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \frac{\partial U(x,y)}{\partial n}\frac{e^{ikr}}{r}\; dx\, dy \tag{3.23}$$

The Rayleigh-Sommerfeld diffraction formula is given by

$$\boldsymbol{U}(\xi,\eta,z) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \boldsymbol{U}(x,y,0)h(x,y;\xi,\eta)\; dx\, dy \tag{3.24}$$

The first Rayleigh-Sommerfeld solution will be used in all subsequent calculations due to its simplicity.

3.3.2 Fourier optics

A few convenient approximations can be used in the typical aperture diffraction problem. The imaging distance is usually much larger than the diffracting aperture, $z \gg x$ and $z \gg y$. The large distance allows us to approximate r by [38, 68]

$$\begin{aligned} r &= \sqrt{z^2+(x-\xi)^2+(y-\eta)^2} \\ &= z\sqrt{1+\left(\frac{x-\xi}{z}\right)^2+\left(\frac{y-\eta}{z}\right)^2} \\ &\approx z\left(1+\frac{1}{2}\left(\frac{x-\xi}{z}\right)^2+\frac{1}{2}\left(\frac{y-\xi}{z}\right)^2\right) \end{aligned} \tag{3.25}$$

Using this first order approximation for r, and assuming that the diffraction angle of interested is very small as is usually the case, $\cos\theta = 1$, the Rayleigh-Sommerfeld kernel can be reduced to

$$\begin{aligned} h(x,y;\xi,\eta) &= \frac{1}{i\lambda r}e^{ikr}\cos\theta \\ &\approx \frac{1}{i\lambda z}e^{ikz\left(1+\frac{1}{2}\left(\frac{x-\xi}{z}\right)^2+\frac{1}{2}\left(\frac{y-\xi}{z}\right)^2\right)} \\ &= \frac{e^{ikz}}{i\lambda z}e^{i\frac{k}{2z}\left((x-\xi)^2+(y-\eta)^2\right)} \end{aligned} \tag{3.26}$$

leading to the Fresnel approximation

$$U(\xi,\eta,z) = \frac{e^{ikz}}{i\lambda z}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\boldsymbol{U}(x,y,0)e^{i\frac{k}{2z}\left((x-\xi)^2+(y-\eta)^2\right)}\,dx\,dy \tag{3.27}$$

Equation 3.27 is also known as the near field equation. In the far field, when z is much larger, we can further approximate $e^{i\frac{k}{2z}(x^2+y^2)}$ by 1. This leads to the Fraunhofer diffraction equation, which has the same form as a Fourier transform!

$$\begin{aligned} U(\xi,\eta,z) &= \frac{e^{ikz}}{i\lambda z}e^{i\frac{k}{2z}(\xi^2+\eta^2)}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\boldsymbol{U}(x,y,0)e^{i\frac{k}{2z}(x^2+y^2)}e^{-i\frac{2\pi}{\lambda z}(x\xi+y\eta)}\,dx\,dy \\ &\approx \frac{e^{ikz}}{i\lambda z}e^{i\frac{k}{2z}(\xi^2+\eta^2)}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\boldsymbol{U}(x,y,0)e^{-i\frac{2\pi}{\lambda z}(x\xi+y\eta)}\,dx\,dy \end{aligned} \tag{3.28}$$

Using the Fourier transform, the wave-like interference properties of the imaging process can be decomposed into its component angular spectra. The properties of the Fourier transform like linearity, the scaling property, or more usefully, the convolution-multiplication law, also corresponds to various optical imaging operations.

From the linearity of the Fourier transform, brighter apertures fields result in proportionately brighter angular spectra. Additionally, the individual angular spectra of different subapertures sum. From Parseval's theorem, the total intensity in the angular spectrum is conserved. Due to scaling, larger apertures result in narrower angular spectra, and conversely, smaller apertures result in angular spectra that is more spread out.

In the Fourier displacement or shift property, displacements in the aperture result in linear phase shifts in the angular spectra. The converse, a more relevant property, is that a phase shift at the aperture results in a displacement of the angular spectra. Finally, from the convolution-multiplication theorem, we discover a new class of optical transformations. Fourier optical image processing in the frequency domain (focal plane) represent a new and powerful class of techniques that sometimes cannot be done purely in the spatial domain (aperture plane).

3.3.3 Fourier imaging with lenses

The effect of lenses on light can also be described using Fourier optics. The curved surface, or gradient in the refractive index of a transparent optical material, and the corresponding optical path differences, adds additional phase terms to the transmitted light. The wavefront added by a convex lens of uniform refractive index is proportional to the thickness of the lens.

For a spherical lens with radii of curvatures R_1 and R_2, the thin lens approximation (called the Lensmaker's equation) for its thickness is

$$\Delta(x,y) = \Delta_0 - \frac{x^2+y^2}{2}\left(\frac{1}{R_1} - \frac{1}{R_2}\right) = \Delta_0 - \frac{x^2+y^2}{2(n-1)f} \tag{3.29}$$

with a lens refractive index of n and resultant focal length of f for the lens.

The added phase term from the lens curvature, ignoring the constant phase terms due to Δ_0, is

$$e^{ik(n-1)(\Delta(x,y)-\Delta_0)} = e^{-ik(n-1)\frac{x^2+y^2}{2}\left(\frac{1}{R_1}-\frac{1}{R_2}\right)} = e^{-i\frac{k}{2f}(x^2+y^2)} \tag{3.30}$$

Adding the thin lens phase term to the aperture field in Equation 3.27, the complex field after passing through the lens (Fresnel propagation) is

$$\begin{aligned} \boldsymbol{U}(\xi,\eta,z) &= \frac{e^{ikz}}{i\lambda z} e^{i\frac{k}{2z}(\xi^2+\eta^2)} \\ &\quad \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \boldsymbol{U}(x,y,0) e^{-i\frac{k}{2f}(x^2+y^2)} e^{i\frac{k}{2z}(x^2+y^2)} e^{-i\frac{2\pi}{\lambda z}(x\xi+y\eta)}\, dx\, dy \end{aligned} \tag{3.31}$$

At the focal plane, where $z = f$, the exponential phase terms cancel, leaving

$$\frac{e^{ikz}}{i\lambda z} e^{i\frac{k}{2z}(\xi^2+\eta^2)} \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \boldsymbol{U}(x,y,0) e^{-i\frac{2\pi}{\lambda z}(x\xi+y\eta)}\, dx\, dy \tag{3.32}$$

This is the same form as the Fourier or Fraunhofer far-field, except it is at a convenient finite distance, readily setup in laboratory experiments.

3.4 Transport equations

The Parabolic Equation (as explained by Teague [96]) is an approximation to the scalar wave equation (Equation 3.15) and is an alternative representation of the Fresnel propagation equation (Equation 3.27) [97].

$$i\frac{\partial}{\partial z} u(\boldsymbol{r}) + \frac{\boldsymbol{\nabla}^2 u(\boldsymbol{r})}{2k} + k u(\boldsymbol{r}) = 0 \tag{3.33}$$

where $\boldsymbol{r}$ represents (x,y) (transverse plane to the propagation direction z), $k = \frac{2\pi}{\lambda}$ is assumed to be constant (monochromatic light), and $\boldsymbol{\nabla}^2 u(\boldsymbol{r}) = \frac{\partial^2}{\partial x^2} u(\boldsymbol{r}) + \frac{\partial^2}{\partial y^2} u(\boldsymbol{r})$ is the Laplacian of the complex field along the transverse plane.

The complex field $u(\boldsymbol{r})$ travels along the z direction, and can be broken down into the amplitude (intensity) and phase or wavefront parts,

$$\begin{aligned} u(\boldsymbol{r}) &= \sqrt{I(\boldsymbol{r})} e^{i\phi(\boldsymbol{r})} \\ I(\boldsymbol{r}) &= |u(\boldsymbol{r})|^2 \\ W(\boldsymbol{r}) &= \frac{\lambda\phi}{2\pi} = \frac{\lambda \arg(u(\boldsymbol{r}))}{2\pi} \end{aligned} \tag{3.34}$$

The parabolic equation may be thought of as a transport equation, (a Field Transport Equation (FTE)), that describes the evolution of the complex field $u(\boldsymbol{r})$ along the z-axis.

$$\frac{\partial u}{\partial z} = i\frac{\nabla^2 u}{2k} + iku \tag{3.35}$$

(dropping the $(\boldsymbol{r})$ for succinctness)

In Teague's analysis [95,97], the Field Transport Equation is broken down into the Intensity and Wavefront Transport Equations.

The ITE is

$$\frac{\partial I}{\partial z} = -I\nabla^2 W - \nabla I \cdot \nabla W \tag{3.36}$$

and the WTE is

$$\frac{\partial W}{\partial z} = 1 - \frac{|\nabla W|^2}{2} + \frac{\lambda^2}{16\pi^2}\frac{\nabla^2 I}{I} - \frac{\lambda^2}{32\pi^2}\frac{|\nabla I|^2}{I^2} \tag{3.37}$$

Here, similar to the previously defined Laplacian $\nabla^2 W = W_{xx} + W_{yy}$, the gradient is taken in the plane transverse to the optical axis $\nabla W = W_x\hat{x} + W_y\hat{y}$. $|\nabla W|^2$ stands for $W_x^2 + W_y^2$.

This has the advantage of separating the intensity distribution (image) of a complex field, a measurable quantity, from the wavefront distribution, which is not directly measurable. Solutions to the ITE for image propagation over short distances, in setups similar to phase diversity [95,98], have been proposed as a method for phase retrieval [39,40,45] and wavefront sensing [87,114].

In the wavefront sensor known as the curvature sensor, scintillation at the telescope aperture is ignored. The Intensity Transport Equation can be applied with the approximation $\frac{\partial I}{\partial z} = -I\nabla^2 W$, ignoring the second term, $-\nabla I \cdot \nabla W$, which represents the intensity gradient. Any changes in intensity during propagation is approximated by the wavefront curvature at the telescope aperture. The second term of the Intensity Transport Equation describes the displacement or directionality of light propagation due to the wavefront slope. An alternative interpretation of the ITE is previously described in Equation 3.9 and Equation 3.10.

The Intensity Transport Equation has been studied in great detail in the literature cited previously. Understandably, less attention has been given to the Wavefront Transport Equation, since the wavefront is not directly measurable. However, certain properties of the WTE are useful in describing geometric optics as a subset of diffractive optics.

The geometric optics approximation of the WTE is given by its first two terms

$$\frac{\partial W}{\partial z} = 1 - \frac{|\nabla W|^2}{2} \tag{3.38}$$

This simplified WTE was first introduced in Equation 3.8 and ignores the diffractive wave nature of light by letting $\lambda = 0$. Equation 3.38 describes the direction of propagation of a wavefront in terms of the wavefront slope, affirming the principle of ray tracing at a direction normal to the wavefront, previously described in Equation 3.9 and Equation 3.10.

Chapter 4

Adaptive optics

This chapter examines the effects of atmospheric turbulence on optical systems. Atmospheric turbulence is a random process that follows Kolmogorov statistics [53], and is modelled within optical systems as optical aberrations. It is typically characterised by a few parameters that are introduced in Section 4.1.1.

Wavefront sensors are used to measure the aberrations caused by atmospheric turbulence. Section 4.3 introduces the problem of slope estimation. Section 4.4 generalises slope estimation to full wavefront sensing, introducing the four major classes of wavefront sensors studied in this thesis, the Shack-Hartmann, pyramid, geometrical and curvature sensors. The four wavefront sensors will be further developed and uniformly compared in subsequent chapters.

4.1 Kolmogorov turbulence

Big whorls have little whorls,
Which feed on their velocity;
Little whorls have smaller whorls,
And so on unto viscosity.
L. F. Richardson (1881-1953)

The atmosphere of the Earth is in a constant state of change, driven by heat from the sun, pressure differences across the globe, and the rotation of the Earth itself. The dissipation

of heat energy creates vortices of turbulence in the atmosphere, gradually shrinking in size until the energy is lost to the friction from the viscosity of air. The largest and smallest vortex sizes in this energy transfer correspond to the outer and inner scale of the turbulence. Based on dimensional analysis of the energy transfer from the outer and inner scales, the statistics of the turbulence spectrum can be shown to obey a $-\frac{11}{3}$ power law[1], known as the Kolmogorov power law [54].

$$\Phi_n(\boldsymbol{f}) = 0.033 C_N^2 \boldsymbol{f}^{-\frac{11}{3}} \tag{4.1}$$

where $\boldsymbol{f}$ is the frequency, and C_N^2 is known as the index structure coefficient.

It is the irregular temperature and pressure changes in the atmosphere that causes fluctuations in the refractive index of air which ultimately degrades the image quality. The structure function of the refractive index fluctuations is given by the index structure coefficient, which varies according to

$$D_n(\boldsymbol{\rho}) = \left\langle (n(\boldsymbol{r}) - n(\boldsymbol{r}+\boldsymbol{\rho}))^2 \right\rangle = C_N^2 \, |\boldsymbol{\rho}|^{\frac{2}{3}} \tag{4.2}$$

where $\boldsymbol{\rho}$ and $\boldsymbol{r}$ are 3-dimensional position vectors, and $n(\boldsymbol{r})$ is the refractive index at position $\boldsymbol{r}$.

The fluctuations in the refractive index are assumed to be symmetrical in all directions for small distances. Under this isotropic behaviour, the scalar quantity $|\boldsymbol{\rho}|$ is sufficient to describe $D_n(\boldsymbol{\rho})$. Changes in the refractive index of the atmosphere causes deformations in the wavefront of the light passing through. The total phase fluctuation at any point on the ground is found by integrating the deformations over the whole path of the light ray through the atmosphere.

$$\phi(x,y) = \frac{2\pi}{\lambda} \int_0^\infty n(x,y,z) \, dz \tag{4.3}$$

where λ is the wavelength of the light.

Equation 4.2 and Equation 4.3 allows us to derive the phase structure function, which determines the statistics of the wavefront aberration at ground level due to turbulence.

[1] As such, many important quantities in this section have characteristic power laws in fractions of $\frac{1}{3}$.

$$D_\phi(\boldsymbol{x'}) = \left\langle (\phi(\boldsymbol{x}) - \phi(\boldsymbol{x}+\boldsymbol{x'}))^2 \right\rangle = 2.91k^2 \sec\gamma \, \left|\boldsymbol{x'}\right|^{\frac{5}{3}} \int_0^\infty C_N^2(z)\, dz \tag{4.4}$$

where $\boldsymbol{x}$ and $\boldsymbol{x'}$ are 2-dimensional position vectors, and γ is an angular distance from the zenith. The air mass, $\sec\gamma$, a measure of the thickness of the atmosphere (and turbulence) that light needs to travel through, is minimised by timing astronomical observations to take place near the zenith (overhead).

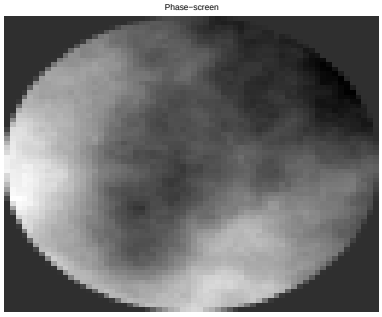


Figure 4.1 Simulation of a phase-screen obeying Kolmogorov statistics.

The profile of $C_n^2(z)$ over height is specific to the observatory site and conditions. For analysis purposes, a few models are frequently used for comparison. Each model relates the average C_N^2 to the height z. For example, the most common Hufnagel-Valley Boundary model is

$$C_n^2(z) = 5.94 \times 10^{-23} z^{10} e^{-z} \left(\frac{W}{27}\right) + 2.7 \times 10^{-16} e^{-\frac{2z}{3}} + A e^{-10z} \tag{4.5}$$

where W is related to the wind speed, and A the ground boundary layer.

Other wind/turbulence profile models like the SLC Day and Night models are also used. To a first approximation, most of the atmospheric turbulence can be assumed to be confined to a few strong layers, and in many cases, a single dominant layer close to the ground. In this thesis, simulations of turbulence use only a single layer represented using a phase-screen. As such, the precise $C_N^2(z)$ profile is not considered.

4.1.1 Optical effect of atmospheric turbulence

The optical effect of atmospheric turbulence can be summarised with a few commonly cited parameters.

Fried's parameter

The degradation in resolving power of a telescope is measured by Fried's parameter, r_0. Also known as the seeing cell size, it is the effective telescope diameter caused by atmospheric turbulence.

$$r_0 = \left(0.423k^2 \sec\gamma \int_0^\infty C_n^2(z)\, dz \right)^{-\frac{3}{5}} \tag{4.6}$$

Hence, in the presence of uncompensated atmospheric turbulence, the maximum achievable resolution is equivalent to that from a telescope of diameter r_0 without the atmosphere.

The most instructive trends from Equation 4.6 for r_0 are $r_0 \propto \lambda^{\frac{6}{5}}$ and $r_0 \propto \sec\gamma^{-\frac{3}{5}}$. The increase in r_0 with increasing wavelength increases the effective telescope diameter. This can also be seen in Equation 4.3, where the larger r_0 results in reduced phase errors. Many imaging telescopes work in the infra-red spectrum to reduce the effects of atmospheric turbulence. Adaptive optics compensation in the infra-red is also more effective than at shorter wavelengths. Wavefront sensor systems often piggy-back on the same path, sensing in the unused ultra-violet region.

Using r_0, we can also rewrite Equation 4.4 into a more convenient form.

$$D_\phi(\boldsymbol{x}') = \left\langle \left| \phi(\boldsymbol{x}) - \phi(\boldsymbol{x}+\boldsymbol{x}') \right|^2 \right\rangle = 6.88 \left(\frac{|\boldsymbol{x}'|}{r_0} \right)^{\frac{5}{3}} \tag{4.7}$$

Isoplanatic angle

The blurring caused by the atmosphere is not uniform across the whole sky. However, for a limited area, it is relatively constant. The isoplanatic angle is a rough measure of the angular distance over which no appreciable changes can be observed.

$$\theta_0 = \left(2.91k^2 (\sec\gamma)^{\frac{8}{3}} \int_0^\infty C_n^2(z) z^{\frac{5}{3}}\, dz \right)^{-\frac{3}{5}} \tag{4.8}$$

The isoplanatic angle is strongly determined by the turbulence higher in the atmosphere, since there is a $z^{\frac{5}{3}}$ height dependency in Equation 4.8. The $\sec^{\frac{8}{5}}\gamma$ factor of the air mass also

makes off-zenith imaging problematic.

Greenwood frequency

The evolution of turbulence over time is typically related to its spatial statistics using Taylor's frozen flow hypothesis. Under this hypothesis, the turbulence itself is assumed to be static, but is blown across the field of view of the telescope. The temporal statistics of atmospheric turbulence can thus be determined by the spatial structure function and the wind speed [44].

When wind velocity profiles are available, the atmospheric rate of change can be described using the Greenwood frequency.

$$f_G = 2.31\lambda^{-\frac{6}{5}} \left(\sec\gamma \int_0^\infty C_n^2(z) V(z)^{\frac{5}{3}}\, dz \right)^{\frac{3}{5}} \tag{4.9}$$

where $V(z)$ is the wind velocity at height z.

Zernike modes of atmospheric turbulence

When circular apertures are used for imaging, as shown in Figure 4.1, the phase function can be described in terms of its component Zernike modes (Equation 2.81, reproduced here in rectangular coordinates), as in classical optics.

$$\phi(x,y) = \sum_{i=1}^{\infty} \alpha_i Z_i(x,y) \tag{4.10}$$

An example is shown in Figure 4.2, where a static phase-screen is decomposed into its individual Zernike modes, and arranged in increasing order using Noll's numbering scheme.

The most significant contribution to the atmospheric phase-screen tend to come from the lower order Zernike modes, corresponding to image displacement and defocus, followed by higher order aberrations like astigmatism and coma. The expected magnitude of each Zernike component is 0, but the expected power (squared magnitude) of each component can be found analytically and expressed as the phase covariance matrix [65].

Although theoretically, any set of orthogonal bases functions is acceptable, when we are limited to a truncated representation of phase function due to the use of a finite num-

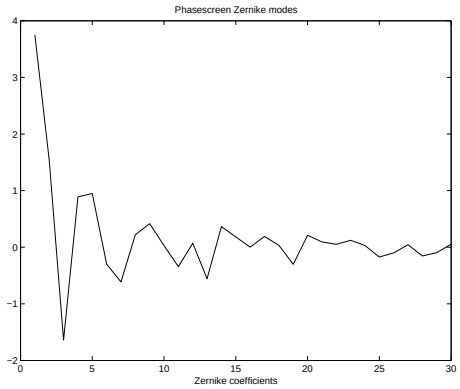


Figure 4.2 The magnitudes of the Zernike modes of one simulated atmospheric turbulence instance. Note that the piston term, corresponding to mode 1 (equivalent to the mean value of the simulated phase representation), does not affect the image.

ber of modes, the choice of functions should be selected to contain as much information as possible (on average). The optimal choice would be the Karhunen-Loeve[2] functions, whose coefficients are statistically uncorrelated (diagonal covariance matrix). Although the Karhunen-Loeve transform for atmospheric turbulence cannot be expressed analytically, it can in practice be expressed in terms of Zernike polynomial expansions.

As can be seen in Table 4.1, the covariance matrix for the Zernike terms is almost diagonal, showing low correlation between terms and decreasing power with higher orders. When truncating the Zernike coefficient representations, the lowest order modes should be retained to represent the most amount of energy. The almost diagonal covariance matrix means the Zernike polynomial representation is a good approximations to the optimal Karhunen-Loeve transform for atmospheric turbulence. For a finite number of terms, the covariance matrix can be de-correlated (diagonalised into principal components) to give the Karhunen-Loeve functions.

The phase variance over a circular region scales as $(D/r_0)^{\frac{5}{3}}$, where D is the diameter of that region. Most of the power is present in the lower order modes (diagonal matrix terms), showing that near fit of the Zernike polynomials to the Karhunen-Loeve functions of turbulence. For higher order modes, the residual estimation error after removing the first N modes from Kolmogorov turbulence is given by

[2]This is equivalent to the use of the Karhunen-Loeve transform as the most efficient compression scheme in signal processing. The Karhunen-Loeve functions are in effect the eigenfunctions of Kolmogorov turbulence.

$\left(\frac{D}{r_0}\right)^{\frac{5}{3}}$	α_2	α_3	α_4	α_5	α_6	α_7	α_8	α_9	α_{10}
α_2	0.448	0	0	0	0	0	-0.0141	0	0
α_3	0	0.448	0	0	0	-0.0141	0	0	0
α_4	0	0	0.0232	0	0	0	0	0	0
α_5	0	0	0	0.0232	0	0	0	0	0
α_6	0	0	0	0	0.0232	0	0	0	0
α_7	0	-0.0141	0	0	0	0.0062	0	0	0
α_8	-0.0141	0	0	0	0	0	0.0062	0	0
α_9	0	0	0	0	0	0	0	0.0062	0
α_{10}	0	0	0	0	0	0	0	0	0.0062

Table 4.1 Covariances of the first 10 Zernike coefficients, scaled by $\left(\frac{D}{r_0}\right)^{\frac{5}{3}}$.

$$E_N \approx 0.2944 N^{-\frac{\sqrt{3}}{2}} \left(\frac{D}{r_0}\right)^{\frac{5}{3}} \tag{4.11}$$

4.2 Laser guide stars

In adaptive optics, the wavefront distortion for the object of interest is usually estimated from a nearby reference star. This avoids using light from the object itself, a method which reduces throughput to the observation path. It also allows adaptive optics correction to be used for objects that may be too dim for wavefront sensing. The reference star has to be bright and close enough to the object to provide a good wavefront estimate, ideally well within the isoplanatic angle θ_0. There are not enough natural guide stars to allow observations of all interesting astronomical objects.

Observatories today use laser beacons as artificial guide stars to provide a bright reference source, extending the sky coverage [29, 69, 80, 110]. Laser guide stars are formed using one or more laser beams pointed near the object of interest. The artificial star can be formed by either Rayleigh scattering or by sodium resonance fluorescence.

Rayleigh scattering is based on scattering by air molecules in the lower atmosphere. The height of laser guide stars based on Rayleigh scattering is limited to between 5 to 20 km, which is approximately the same height as atmospheric turbulence. The thinner atmosphere at higher altitude also limits the brightness of Rayleigh laser guide stars. However, although

the turbulence layers here are weak, they are still significant because of the $z^{\frac{5}{3}}$ height dependence of the atmospheric isoplanatic angle (Equation 4.8) and temporal rate of change (Equation 4.9).

An alternative scattering method uses the sodium layer present at 90 km in the mesospheric layer [63]. Sodium atoms are resonant at 589.2 (D_2) and 589.6 (D_1) nm, scattering laser beams of that wavelength. Sodium laser guide stars can create light sources that are not only brighter than Rayleigh laser guide stars of equivalent power, but also higher, allowing the turbulence at higher altitudes to be measured. Sodium beacons are thus generally preferred to Rayleigh beacons.

4.2.1 Cone effect and anisoplanatism

A guide star and the object of interest are not affected by the same patch of atmospheric turbulence. The angular separation between the object and the guide star is bounded by the isoplanatic angle θ_0 [101], the angle over which turbulence effects may be considered to be constant [28]. The mean squared phase error for a separation of θ is on average

$$\sigma_\phi^2 = \left(\frac{\theta}{\theta_0}\right)^{\frac{5}{3}} \text{rad}^2 \tag{4.12}$$

Additionally, in laser guide stars, the limited height of the beacon (as opposed to the very distant natural star), also gives rise to what is known as the cone effect.

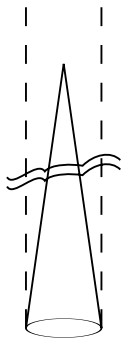


Figure 4.3 The limited height of laser guide stars compared to distant stars restricts its measurement of atmospheric turbulence.

The use of laser guide stars is often associated with multi-conjugate adaptive optics [50,58]. Multiple guide stars are used to cover a larger patch of the sky [78]. Additionally, within the

adaptive optics correction system, multiple mirrors are conjugated (hence the name) to various heights in the atmosphere to provide optimal correction. Since propagated light suffers from both phase and amplitude fluctuations, and deformable mirrors can only provide phase compensation, the most effective compensation is obtainable only by conjugating the mirror compensation to the height of the turbulence, effectively compensating the turbulence before the propagation that causes intensity fluctuations [25].

Although laser guide stars seem to provide the ultimate solution to the sky coverage and guide-star brightness problem, they have a major drawback. The laser beam displacements in the outgoing and returning beam cancel because they pass through the same turbulence, causing no apparent displacement in the guide star image. Because of this, laser guide stars cannot be used to improve the wavefront slope estimate. This is a great disadvantage as wavefront slope comprises 87% of the wavefront errors caused by the atmosphere (as seen in the first 2 terms of the covariance matrix in Table 4.1). Optimal slope detection under limited light thus remains a very important step for image improvement, and motivates the discussion in Chapter 5.

4.3 Wavefront slope estimation

The phase of the complex field at optical wavelengths cannot be measured directly. Its effect on images can however be seen when light is allowed to propagate. Wavefront sensing is basically a means of relating intensity measurements to phase aberrations. We begin by observing the propagation of light through an aperture under Fresnel diffraction, as shown in Figure 4.4. The simulation is carried out with a discrete approximation to Equation 3.27.

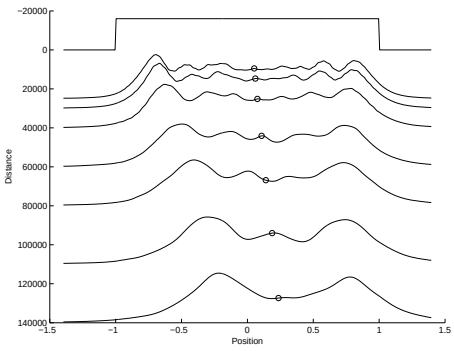


Figure 4.4 The effects of wavefront errors on the propagation of light through free space.

The circles represent the centroid of the image after propagating through free space. In

Teague's moment analyses [96, 97], the first moment, the centroid, travels in a straight line at a direction perpendicular to the mean wavefront slope at the aperture. This also agrees with the geometric optics model of light (Equation 3.9 and Equation 3.10), which predicts that the image is displaced in proportion to the global wavefront slope at the telescope aperture.

Over longer distances, the image displacement is larger, and an increasing variation in the intensity is observed. We expect the image at infinity to show the most amplitude variation in response to phase fluctuations at the aperture. In practice, a focusing lens can be placed at the aperture to reduce the equivalent propagation distance to the focal length of the lens. The lens also concentrates light, intensifying the image signal. Intuitively, the focal plane is thus the optimal position for slope detection.

4.3.1 Focal plane image displacement and the wavefront slope

The relationship between the mean slope and the displacement of the image centroid lies at the heart of most wavefront sensors. This relationship can be shown mathematically in the case of a uniformly illuminated and symmetric aperture as

$$\begin{aligned}
& \int_{-\infty}^{\infty} x\,|i(x)|^2\ dx \\
= \ & -i\int_{-\infty}^{\infty} ix\mathscr{F}\left\{\left(A(u)e^{i\phi(u)}\right)\star\left(A(u)e^{i\phi(u)}\right)\right\}\ dx \\
= \ & -i\,\frac{d}{du}\left[\left(A(u)e^{i\phi(u)}\right)\star\left(A(u)e^{i\phi(u)}\right)\right]\Big|_{u=0} \\
= \ & -i\int_{-\infty}^{\infty} A(u)e^{i\phi(u)}\overline{\frac{d}{du}A(u+0)e^{i\phi(u+0)}}\ du \\
= \ & -i\int_{-\infty}^{\infty} A(u)\frac{d}{du}A(u)\ du - \int_{-\infty}^{\infty} A(u)^2\phi_u(u)\ du \\
= \ & -A(u)^2\int_{-\infty}^{\infty}\phi_u(u)\ du
\end{aligned} \tag{4.13}$$

where $x = \frac{2\pi u}{\lambda z}$.

Least-squares slope estimate

The mean slope is the averaged wavefront slope over the whole aperture. However, due to the averaging operation of the centroid estimator, under certain conditions, an image

may “look” displaced, but still have a centroid of zero. An example of this is shown in Figure 4.5.

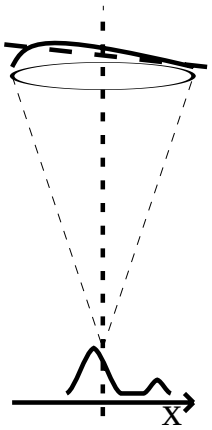


Figure 4.5 A wavefront with 0 mean slope at the aperture, resulting in an image with a centroid of 0. Although the bulk of the image is slightly displaced to the left, the centroid is weighted by distance over the whole image, and is sensitive to the position of distant speckles.

To maximise the Strehl ratio of an image, an adaptive optics system should compensate for image displacements, not by centroiding, but by centering on the brightest point of a speckle (the shift-and-add algorithm). Since the Strehl ratio is approximately related to the squared phase error, brightest spot centering corresponds to estimating the plane of best fit to wavefront aberrations in the least-square sense, as pointed out by Glindemann [33]. This is shown as a line of best fit to the wavefront in Figure 4.5.

The optimal estimates of the second and third Zernike terms are defined to be the least-squares fit of a plane to the wavefront. The position of the brightest point, an alternative to the centroid as a displacement estimator, is thus useful for measuring the wavefront slope in the least-squares sense.

In practice, the wavefront sensors examined in this thesis make use of quad-cells (examined later) for estimating displacement or slope. The undersampling in the quad-cell renders the position estimation of the brightest spot impractical. Subsequently, since the displacement estimate provided by the quad-cell is in fact an estimate of the image centroid, more emphasis is placed on the centroid estimator.

Estimation of higher order wavefront modes

If we are only interested in the global slope estimation from image displacements, then the optimal position for measuring the image is at the focal plane, as explained by van Dam and Lane [104]. While the global slope measurement provided by the quad-cell accounts for most of the wavefront error in Kolmogorov turbulence, it is still necessary to estimate the shape of the wavefront aberration function within the whole aperture. Wavefront sensors are used to detect the higher order modes in the turbulence.

Global slope estimation with quad-cells can be extended to estimate higher order modes by subdividing the wavefront aberrations at the telescope aperture into smaller regions. Within the smaller area, the effects of higher order aberrations are less severe, and the effects of the local wavefront function dominates. The local wavefront within each sub-region can then be measured independently. The sensor signal is typically linear with respect to some function of the wavefront (for example, the wavefront slope or curvature). The sensor signal $\boldsymbol{d}$ is thus given by a matrix operation ($\boldsymbol{H}$) with the wavefront Zernike coefficients $\boldsymbol{\alpha}$.

$$\boldsymbol{d} = \boldsymbol{H}\boldsymbol{\alpha} + \boldsymbol{n} \tag{4.14}$$

where $\boldsymbol{n}$ represents measurement noise in the wavefront sensor.

The higher order wavefront aberrations at the telescope aperture can be found from a linear combination of the local slope signals. Here, the sensor measurements $\boldsymbol{d}$ is a finite vector, while the wavefront coefficients $\boldsymbol{\alpha}$ is in fact infinite. In practice, a finite number of coefficients are estimated, since (as shown in Equation 4.11), the error in the subsequent higher order modes decreases, so the energy in Kolmogorov turbulence is mostly present in the lower order modes.

Depending on the statistics of the noise present in the sensor signal $\boldsymbol{d}$, various solutions for $\boldsymbol{\alpha}$ are obtained. As introduced previously in Section 2.5.2, two Bayesian methods for inverting Equation 4.14, the Maximum Likelihood and Maximum A Posteriori solutions, are commonly used here.

The Maximum Likelihood solution, assuming white noise $\boldsymbol{n}$, is given by

$$\hat{\boldsymbol{\alpha}} = (\boldsymbol{H}^T\boldsymbol{H})^{-1}\boldsymbol{H}^T\boldsymbol{d} \tag{4.15}$$

Given a finite number of measurements in $\boldsymbol{d}$, under low light conditions, with higher noise levels in the measurement, or when more coefficients ($\hat{\boldsymbol{\alpha}}$) are required from the wavefront sensors, additional constraints are required to condition the problem. In such situations, prior knowledge of the statistical distributions of coefficients can be useful, resulting in a Maximum A Posteriori solution of Equation 2.117. Here, the covariance matrix of the Zernike coefficients provide a convenient way to specify prior knowledge of the turbulence statistics. The MAP estimate is optimal as long as the model of the prior is accurate, and is independent of the basis functions chosen to represent the prior.

In many adaptive optics systems, the compensating mirror (Section 1.1.5) is built from actuators that correct the wavefront using local mechanical perturbations. Thus, an alternative problem that is also linear, but consisting of zonal wavefront estimates, can be formulated. In zonal estimation systems, prior information, in the form of the measurement covariance matrix, can be obtained from the covariance analysis in Wallner [107].

4.4 Wavefront sensors

In this thesis, four different wavefront sensors: the Shack-Hartmann, pyramid, curvature and geometric wavefront sensors, are examined in detail. For comparison purposes, a unified framework is developed to place the wavefront sensors in context.

4.4.1 Shack-Hartmann sensor

The Shack-Hartmann sensor [72, 85] consists of an array of lenslets placed at a plane conjugated to the telescope aperture. Each lenslet subdivides the aperture plane into smaller subapertures, and forms a low resolution image of the object at its focal plane.

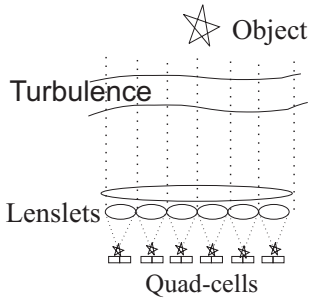


Figure 4.6 Simplified layout of a Shack-Hartmann sensor.

When the lenslets are approximately the same size as the coherence length r_0 of the atmospheric turbulence, the images formed by the lenslets are approximately the same size as the equivalent diffraction limited images formed by the lenslet (for example, with perfect telescope optics, and no atmospheric turbulence). At this size, the major effect of turbulence is in the local wavefront slope over each lenslet, giving rise to random image displacements, as shown in Figure 4.7.

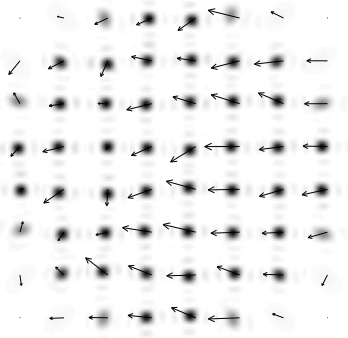


Figure 4.7 Simulated image from a Shack-Hartmann sensor, with sensor signals superimposed (not drawn to scale).

The signal from a Shack-Hartmann sensor is formed from the image displacements, which are linearly related to the wavefront slopes. Since differentiation (slope of wavefront) is a linear process, the slopes are in related to the coefficients of the Zernike polynomials in Equation 2.81.

The intensity of the image under each lenslet is measured with CCD detector arrays. The displacement of the image can be measured using the centroid estimator of Equation 5.1. In practice, to reduce the effect of read-noise in the CCD detectors, quad-cells are used to determine the displacement of the image.

The lenslet size is typically unchangeable for a fixed optical configuration, and needs to be tailored to the local turbulence conditions. There is a trade-off between the more precise estimate available from larger lenslets, with the better spatial resolution or sampling available from having more (and smaller) lenslets. A simulation for choosing the optimal lenslet size is presented in Section 6.3.2.

Image displacement is almost[3] independent of wavelength. This allows the Shack-Hartmann

[3]Some wavelength dependent refractive effect exists, and is used in polychromatic guide stars for tip/tilt estimation [27].

sensor to be used with broadband or white light, which maximises the amount of light used. Additionally, smaller extended objects are not resolvable through the small lenslets, and are suitable for use as guide stars. The simplicity and robustness of the Shack-Hartmann sensor has led to its widespread adoption in adaptive optics systems.

4.4.2 Pyramid wavefront sensor

The pyramid sensor consists of a pyramid-shaped prism at the focal plane of the telescope, and some re-imaging optics behind it. It was first suggested in various forms by Babcock and Ragazzoni [9, 76], and improves upon the qualitative Foucault knife edge test [26, 111] by allowing quantitative measurements of wavefronts to be made.

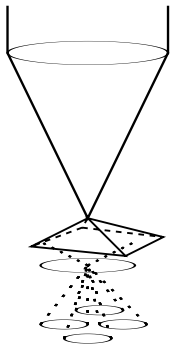


Figure 4.8 The pyramid wavefront sensor consists of a pyramidal prism at the focal plane. The subdivided field in each quadrant is re-imaged into 4 separate sub-images.

The pyramid sensor subdivides the complex field at the telescope focal plane into quadrants, and re-images each quadrant into 4 images of the telescope aperture. The pyramidal prism is there simply to spread out the sub-images to avoid overlaps. For analysis purposes, the prism may be ignored, as only the subdivision operation is important here.

The sensitivity and linearity of the pyramid sensor are a function of the image size, and artificially enlarging the image size on the pyramid can be beneficial. The most common method is to achieve this by a repetitive motion to increase its apparent size. This is examined further in Section 6.4.

As a first approximation, the light distributions within each individual sub-images may be ignored, by considering only their total intensities. This results in four intensity measurements, one for each quadrant in the focal plane. This is of course equivalent to a focal plane quad-cell for estimating image displacement, which corresponds to the global wavefront slope at the aperture.

By extension, due to linearity in the image intensities, as shown in Section 6.4, the intensity distribution in the sub-images is proportional to local wavefront slopes in the aperture plane. Figure 4.9 shows how a local wavefront slope in the aperture translates to localised intensity changes in each of the re-imaged apertures beyond the telescope focal plane.

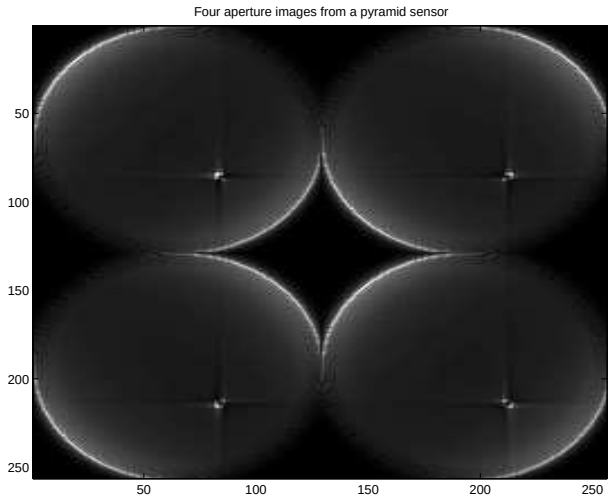


Figure 4.9 The re-imaged telescope aperture in the pyramid sensor, showing the signal arising from a flat wavefront with a small local perturbation.

The re-imaged copies of the aperture are blurred by the prism subdivision at the telescope focal plane. Roughly speaking, each subdivision, or facet of the pyramid, retains only $\frac{1}{4}$ of the illumination at the focal plane. This loss of information from the subdivision process determines the ultimate limit to the resolution of the wavefront estimate of a pyramid wavefront sensor.

Chapter 6 further demonstrates that the pyramid wavefront sensor is in fact a dual of the Shack-Hartmann sensor, with many equivalent functions performed in the dual Fourier space.

4.4.3 Curvature sensor

The curvature sensor is an image based wavefront sensor that measures wavefront curvature instead of slope. It was proposed by Roddier [86] as a simple low-order wavefront sensor especially tailored to astronomical imaging applications as opposed to earlier systems oriented towards military uses. It has found widespread use in infra-red applications where the effect of turbulence is less severe. The curvature sensor consists of two imaging planes placed before and after the nominal focal plane of a telescope.

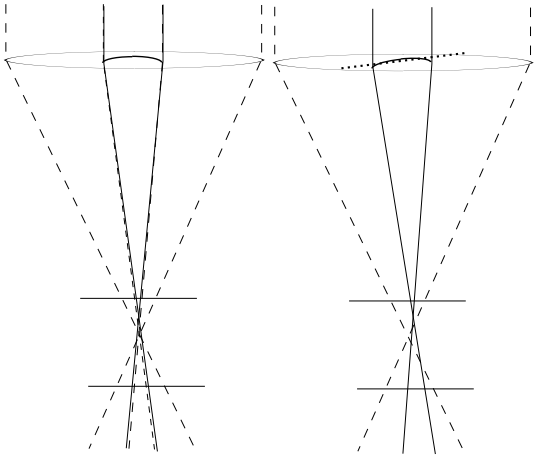


Figure 4.10 Layout of a curvature sensor showing the in-focus (above) and outside-focus imaging planes (below). The dashed lines represent the paths of light rays when there are no wavefront errors. On the right, the same aberration with a local slope causes an opposing displacement in the intensity signals in each image plane.

Using this layout, any wavefront errors at the aperture plane shows up as opposing intensity changes in the two out-of-focus imaging planes. In the example shown in Figure 4.10, a small negative curvature is added to the wavefront in the centre of the aperture. This causes the focal point for that sub-region in the aperture to move forward, so the corresponding region becomes brighter (and smaller) in the in-focus image, and darker (and larger) in the outside-focus image. The intensity within that sub-region in the two image planes is approximately proportional to the wavefront curvature, as shown in Equation 3.11.

The sensor output is taken to be the intensity difference between the two imaging planes, and is proportional to the wavefront curvature. For a small change in curvature $\Delta H(x,y,0)$ at the aperture, the corresponding change in intensity after propagating a distance of z is

$$\Delta I(x,y,z) \approx -zI(x,y,0)\Delta H(x,y,0) \tag{4.16}$$

In the original curvature sensor, it was proposed that the sensor outputs be sent directly to a bimorph deformable mirror, which will respond with a proportional curvature on its surface. In practice, additional processing of the sensor signal is required to match the characteristics of each component.

There are questions as to the accuracy of a sensor signal formed from the intensity differences in the two imaging planes. As exaggerated in Figure 4.10, curvature errors in the

wavefront not only result in local intensity changes, but also changes in the size of the image (sub-region). The most significant outcome of this effect can be seen at the edges of images. In practice, wavefront sensors treat the boundary of images as differential signals proportional to the wavefront slope. This raises the significant question of how the boundary slope signal is to be separated from the internal curvature signal.

The misalignment error also arises when the mean local slope within a sub-region results in a displacement of the intensity signal, so that the bright and dark spots in the two defocused imaging planes are no longer aligned, as shown in Figure 4.10 (right). A better solution to the curvature sensor equation has been proposed by van Dam and Lane [102] to take into account the full geometric optics behaviour of light. This new method is known as the geometric wavefront sensor.

4.4.4 Geometric wavefront sensor

The geometric wavefront sensor is a slope sensor. The physical layout of the geometric wavefront sensor is identical to the curvature sensor. However, it uses an improved interpretation of the intensity distribution in the out-of-focus images, using an exact geometric optics solution to recover wavefront aberrations by a ray tracing process.

To illustrate the underlying philosophy of the geometric wavefront sensor, we simplify the wavefront propagation problem to 1D, as shown in Figure 4.11.

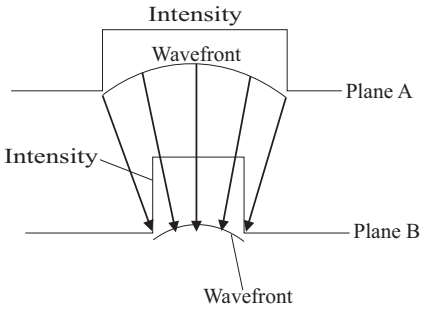


Figure 4.11 A simple defocus in the wavefront causes the image of the aperture to be smaller but brighter.

Light propagates in a direction perpendicular to the wavefront slope (Equation 3.9). At the same time, the intensity changes as it is concentrated or dispersed, as described by Equation 3.11. With a 1D aperture, these equations are reduced to

$$x_B = x_A + \Delta z W_x(x_A, z_A) \tag{4.17}$$

where $\Delta z = z_B - z_A$ and W_x is the wavefront slope along the x-axis.

for a light ray at x_A in the aperture, travelling to x_B in the image plane. The intensity changes are given by

$$I(x_B, z_B) = \frac{I(x_A, z_A)}{1 + \Delta z W_{xx}(x_A, z_A)} \tag{4.18}$$

The wavefront slope at the aperture can be recovered by tracing the light ray path between the aperture and image planes. Figure 4.12 shows the same wavefront from Figure 4.11, with the light rays found from comparing the intensities between the two planes. Intuitively, due to the conservation of light, the total intensity between any two light rays (shaded region of Figure 4.12) must be constant.

$$\int_{-\infty}^{x_A} I(x, z_A) dx = \int_{-\infty}^{x_B} I(x, z_B) dx \tag{4.19}$$

The wavefront slope at x_A is given by $\frac{x_B - x_A}{\Delta z}$.

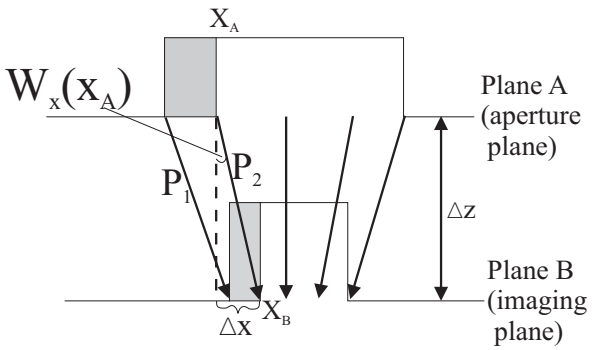


Figure 4.12 Geometric optics model for the propagation of light.

Equation 4.19 allows the positions of the light rays to be recovered from the intensity distribution at planes A and B. Ray tracing provides an exact solution to the problem, as can be seen from equating the intensity between the two light rays P_1 and P_2.

$$\begin{aligned}\int_{-\infty}^{x_A} I(x,z_A)\,dx &= \int_{-\infty}^{x_B} I(x',z_B)\,dx' \\ &= \int_{-\infty}^{x_B} \frac{I(x',z_A)}{1+\Delta z W_{xx}(x',z_A)}\,dx' \\ &= \int_{-\infty}^{x_B-\Delta z W_x} \frac{I_A(x)}{1+\Delta z W_{xx}}(1+\Delta z W_{xx})\,dx \end{aligned} \tag{4.20}$$

substituting $x' = x + \Delta z W_x(x,z_A)$ for $dx' = (1+\Delta z W_{xx}(x,z_A))dx$.

The exact wavefront slope can be estimated by equating the limits to the integrals in Figure 4.20.

$$x_B - x_A = \Delta z W_x \tag{4.21}$$

so the wavefront slope is exactly $W_x(x,z_A) = \frac{\Delta x(x)}{\Delta z}$.

Chapter 7 expands on the application of this method to the wavefront sensing, and provides a comparison of the performance of the geometric sensor to the curvature sensor.

4.4.5 Unifying theme

All the wavefront sensors examined share the same principles of operation. To estimate the complex field at the telescope, scintillation is assumed to be insignificant, and the amplitude can be assumed to be constant, leaving only the phase to be estimated. The wavefront or phase is not directly measurable, but wavefront slope or curvature can be inferred through intensity measurements.

In wavefront sensors, a wavefront is propagated through an aperture, producing intensity fluctuations in the propagating field. The effect of wavefront aberrations on the intensity of a propagating field is most pronounced when the propagation distance is large. In all four wavefront sensors, the most appropriate models for the diffraction effects are the Fresnel or Fraunhofer approximations.

In the presence of strong wavefront aberrations, diffraction effects are small by comparison to geometric effects, so geometric optics [38] provides a good approximation of the intensity propagation model, and so is a good description of how wavefront sensors work. However, as the image is compensated and approaches its diffraction limit, the geometri-

cal optics assumptions begin to fail. Under such conditions, Fourier optics is required and diffraction effects determine the ultimate performance limits of the wavefront sensors.

Using a geometric optics approximation allows wavefront sensing with extended objects or under broadband light. This flexibility extends the range of application of wavefront sensors. Additionally, the linearity and "localisation" property of geometric optics allows the wavefront to be subdivided directly into smaller sub-problems. Through such subdivision, higher order wavefront components can be estimated.

Depending on the wavefront sensor, the wavefront at the aperture can be subdivided explicitly at the aperture plane, implicitly at the focal plane, or somewhere in between. Within a sub-region in the divided aperture, a linear relation exists between the intensity and the local wavefront slope or curvature. Once such a model or forward problem of a wavefront sensor is obtained, the wavefront estimation problem is solved by inverting the forward problem. In this thesis, the inversion if framed in terms of the maximum-likelihood or maximum-a-posteriori methods.

Resolution-precision trade-off

The subdivision operation is equivalent to a wavefront sampling operation. The resolution of the wavefront estimate is thus dependent on the size of the subdivision; the smaller the sub-apertures, the finer the sampling. However, using smaller sub-apertures results in a lower precision in the individual local wavefront estimates. The trade-off between resolution and precision is an example of the space-bandwidth constant in the dual-space description of signals, and is examined in Section 6.1.

Most wavefront sensors have a tunable gain or sensitivity that affects the precision of the wavefront estimate. The sensitivity may be directly adjusted as in an optical modulation scheme or an electronic gain. Alternative, it may only be present implicitly in the image Strehl, and is not directly adjustable. Where possible, by reducing the sensitivity of a wavefront sensor, it may be possible to reduce the non-linearities in the sensor. Hence, the precision of the wavefront estimate can also be balanced against non-linear errors in wavefront sensors.

4.5 Conclusion

By focusing on how the designs of the wavefront sensors are connected, it is our hope that ultimately, all four wavefront sensors considered here can be shown to be equivalent. Interestingly, two related techniques in scintillation estimation, the scidar and slodar devices, resemble the pyramid sensor physically, so a greater unified theory for understanding wavefront and scintillation detection could be a good extension to the current framework.

In Chapter 6, the Fourier equivalence between the Shack-Hartmann and pyramid wavefront sensors is developed. By the equivalence between individual components of the wavefront sensors, the performance of the wavefront sensors can be compared.

Chapter 7 compares the curvature sensor against the geometric wavefront sensor. Since the wavefront sensors are physically identical, a comparison of the noise propagation through their algorithms is made.

Finally, although this thesis will focus mostly on the fundamental performance of wavefront sensors, in practice, we also need to consider instrument noise. Practical considerations usually result in design configurations that do not allow the full use of the wavefront sensors as described in the following sections.

Chapter 5

Quad-cells

In this chapter, we examine the problem of wavefront slope estimation in greater depth. We have shown in Section 4.3.1 that the global mean wavefront slope at the telescope aperture is proportional to the image centroid at the focal plane.

The fundamental limit to the estimation of image centroid arises from photon noise in the image intensity measurements. Photon noise refers to the fluctuations in the photon count in each image detector element due to the Poisson arrival process of photons. For an expected mean value of N, the photon count fluctuates about its expected mean value with a variance also equal to N.

The intensity distribution in an image is proportional to the density of photon arrivals. Consequently, an image can alternatively be seen to represent the probability density function for photons. Estimating some of the properties of an image, like its displacement, is then equivalent to parametric estimation of a known probability distribution. Using the Cramer-Rao bound [51], the ideal theoretical performance for any displacement estimator is shown to be related to the image shape.

Starting with the measurement of images with CCD arrays [112], we show how the presence of instrument read noise and photon noise lead to trade-offs which lead to the quad-cell. Although the quad-cell is the most commonly used image displacement estimator, because of under-sampling, it does not (strictly speaking) measure the image centroid. In this chapter, we compare the performance of quad-cells to the theoretical Cramer-Rao lower bound for displacement estimators.

Most conventional analyses have concentrated on the image width as the factor that determines performance [66, 69, 110]. In this chapter, the image peak is shown to be a more appropriate measure of quad-cell performance. This also simplifies the analysis of the closed-loop performance of the quad-cell. The performance of the quad-cell derived here is then extended to general slope estimation in wavefront sensors in subsequent chapters.

5.1 Displacement estimation

Images are typically measured using an array of detectors at the image plane, providing a sample of the intensity distributions at discrete points in the image plane.

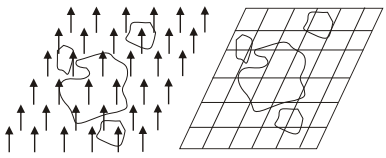


Figure 5.1 The sampling of an image using a finite array of detectors.

The maximum frequency component in an image is limited by the extent of the aperture correlation function (Equation 2.56). For a circularly symmetric aperture of diameter D, the radius of the aperture correlation function, which has two times the extent of the aperture function, is D. The sampling interval that satisfies the Nyquist sampling frequency (Section 2.4.3 and Shannon [90]) (reciprocal of $2D$, or two times the highest frequency in the signal) is given by $\frac{\Delta x}{\lambda f} = \frac{1}{2D}$, that is $\Delta x = \frac{\lambda f}{2D}$, which is approximately a quarter the size of the diffraction limited image. Assuming the Nyquist sampling criterion is satisfied, from a square array of image samples, the image centroid calculated by

$$\hat{x} = \frac{\sum_x \sum_y xI(x,y)}{\sum_x \sum_y I(x,y)},\ \hat{y} = \frac{\sum_x \sum_y yI(x,y)}{\sum_x \sum_y I(x,y)} \tag{5.1}$$

is related to the mean wavefront slope at the aperture by

$$\hat{W}_x = \frac{\hat{x}}{f} \tag{5.2}$$

where f is the focal length of the telescope, and $\hat{W}_x$ is the mean wavefront slope at the telescope aperture in the x-axis.

In the presence of photon noise, the centroid is a random variable. The variance of the denominator in the centroid term can be ignored when the noise level is low. Photon noise has Poisson statistics, so the variance in any intensity measurement in a photo-detector is the same as the expected intensity in the detector. Within a scale factor, the mean and variance of the centroid estimator is given by

$$\begin{aligned} \left\langle \sum_x \sum_y xI(x,y) \right\rangle &= \sum_x \sum_y x \langle I(x,y) \rangle \\ &= \sum_x \sum_y xI(x,y) \end{aligned} \tag{5.3}$$

$$\begin{aligned} \mathrm{var}\left(\sum_x \sum_y xI(x,y) \right) &= \sum_x \sum_y x^2 \mathrm{var}(I(x,y)) \\ &= \sum_x \sum_y x^2 I(x,y) \end{aligned} \tag{5.4}$$

The expected mean and variance of the centroid estimator corresponds to the mean and variance of the image when it is interpreted to be a probability distribution.

5.1.1 Centroid estimator variance

For a finite aperture with a discontinuity at the edges, the asymptotic decay in the focal plane image intensity over the image width, x, is x^{-2}. Unfortunately, because the intensity decay is slower than x^{-1}, this means that the variance in the image centroid estimate in Equation 5.4 is infinite when computed over an infinite plane [47].

In practice, the image measurement area is finite as shown in Figure 5.1. The centroid variance for a x^{-2} intensity decay is proportional to the area over which the image is measured, so the upper bound to the centroid variance is limited by the truncated measurement region. The image truncation produces a bias in the centroid estimator towards zero and also removes any intensity beyond the outer boundaries of the detector. Any intensity here

is spread out by the high frequency phase noise at the aperture, so image truncation at the focal plane effectively acts as a low-pass filter for the phase signal [73].

However, image truncation also causes some information in the image to be lost, so any CCD centroid estimators is no longer statistically optimal. Section 5.3.1 examines the statistical optimality property of a displacement estimator.

In modal wavefront estimation, an aberration function is often expressed as a combination of a finite number of Zernike terms. The relationship between the individual Zernike terms and the resulting image displacement is dependent on how the image displacement is measured. Usually, the image displacement refers to either the image centroid displacement, or the displacement of the brightest spot in the image. The two displacement measures actually correspond to different slope estimates.

The centroid displacement corresponds to the mean wavefront slope at the aperture. Most Zernike terms have a mean slope component, so their presence in a wavefront can result in a displacement in the image centroid. Given a centroid estimate, the corresponding Zernike coefficients cannot be determined unambiguously, since the slope component may be attributed to any Zernike term with a non-zero mean slope.

Alternatively, the displacement of the brightest spot in the image corresponds to the least-squares wavefront slope at the aperture, as previously explained in Section 4.3.1. Since the Zernike slope term is orthogonal to all the other higher order Zernike modes, the presence of higher order Zernike modes do not contribute to any image displacement. Therefore, a single displacement estimate can unambiguously determine the magnitudes of the tip and tilt Zernike terms. Additionally, the least-squares estimate, being independent of the higher order Zernike terms, are not affected by image truncation. That is, in the spatial domain, the position of the bright spot is a local measurement, and cannot be affected by the boundaries of the image.

Suppose the intensity distribution is modelled as a probability distribution, the centroid then corresponds to the mean of the distribution, while the position of the brightest spot corresponds to the mode of the distribution. As will be shown here, the image displacement estimate provided by a quad-cell corresponds to the median of the distribution. From this point of view, the quad-cell represents a compromise between finding the mean slope and the least-squares slope, taking into account the limitations imposed by read noise.

5.2 Slope detection with Quad-cells

To maximise the signal strength and minimise read noise in the centroid estimator, the image is frequently under-sampled. At the extreme, this leads to the quad-cell detector, which consists of 2x2 detector elements as shown in Figure 5.2. The quad-cell detector is a common feature in many wavefront sensors, and is examined in detail by Tyler and Fried [100].

The quad-cell image is nominally centered on the corner adjacent to all four cells. In this position, all four cells will measure the same amount of light. Any image displacements in the *x*-direction can then be measured by comparing intensity changes in the two halves of the plane made up of $A_1 + A_3$ on one side, and $A_2 + A_4$ on the other. Similarly, any displacements in the *y*-direction is given by comparing $A_1 + A_2$ with $A_3 + A_4$ [69, 100].

For small displacements Δx, the intensity in the two halves $A_1 + A_3$ and $A_2 + A_4$ of the plane will show opposing changes. The quad-cell formula is commonly taken to be the differential signal $(A_2 + A_4) - (A_1 + A_3)$, which is monotonically related to the displacement of the image. This differential signal is in fact the centroid formula. However, because of the loss of information in the image truncation and now, also from sub-Nyquist sampling, the signal no longer corresponds to the mean wavefront slope at the aperture.

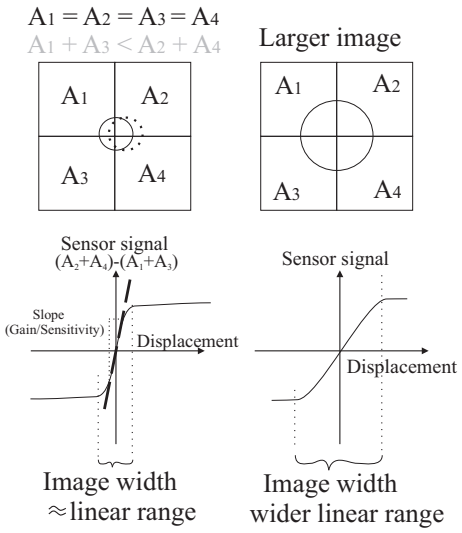


Figure 5.2 Detection of image spot displacement with a quad-cell. When the image is shifted as outlined by the dots, the intensity measurements (no longer equal, as printed in grey) on both halves of the quad-cell plane provide a displacement estimate. This signal is approximately linear for small displacements, and saturates for larger displacements. A larger image size (right) results in a wider range for which the signal is linear, at the expense of the signal gain or sensitivity.

5.2.1 Quad-cell formula

The precise variation in intensity with image displacement depends on the intensity distribution within the image itself. Consider the normalised angular spectrum $h(u,v)$, where

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} h(u,v)\,du\,dv = 1 \tag{5.5}$$

Changing from angular to spatial coordinates $(\frac{x}{f},\frac{y}{f}) = (u,v)$, the normalised image at the focal plane is $h'(x,y) = \frac{1}{f^2}h(\frac{x}{f},\frac{y}{f})$ where f is the telescope focal length. For total mean intensity of N photons, the image itself is $Nh'(x,y)$. Figure 5.3 shows the intensity signal in each half of the image plane for a displacement Δx.

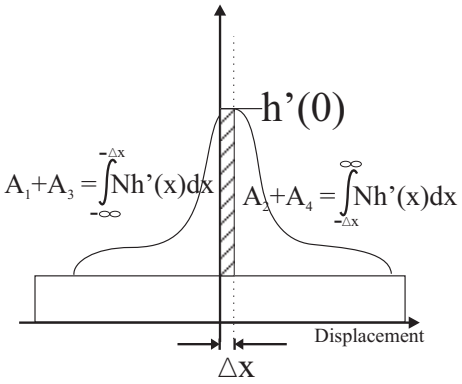


Figure 5.3 Small displacements in the 1D PSF results in opposing intensity changes in each half of the quad-cell. The mean wavefront slope at the aperture is then given by $W_x = \frac{\Delta x}{f}$.

The differential quad-cell signal is given by

$$\begin{aligned}
& (A_2+A_4)-(A_1+A_3) \\
= & \int_0^{\infty}\int_{-\infty}^{\infty} Nh'(x-\Delta x,y-\Delta y)dy\,dx - \int_{-\infty}^{0}\int_{-\infty}^{\infty} Nh'(x-\Delta x,y-\Delta y)dy\,dx \\
= & N\left(\int_{-\Delta x}^{\infty} h''(x')dx' - \int_{-\infty}^{-\Delta x} h''(x')dx'\right) \\
= & N\left(\int_0^{\infty} h''(x')dx' - \int_{-\infty}^{0} h''(x')dx'\right) - 2N\left(\int_0^{-\Delta x} h''(x')dx'\right)
\end{aligned} \tag{5.6}$$

where $h''(x)$ is a 1D projection of the normalised image distribution, $h''(x) = \int_{-\infty}^{\infty} h'(x,y)\,dy$, and $x' = x-\Delta x$. The integral over y also eliminates the displacement Δy in the orthogonal

axis, allowing us to ignore it, simplifying the equation.

The first term of Equation 5.6 is independent of the image displacement. It is usual here [66, 67, 80, 110] to make the assumption that $h'(x,y)$ is modelled by a Gaussian profile. However, the much weaker assumption that the image intensity is equal on both sides of the image peak, is enough to allow us to ignore the first term. This leads to

$$\begin{aligned}(A_2+A_4)-(A_1+A_3) &= -2N\left(\int_0^{-\Delta x} h''(x)dx\right)\\ &\approx 2N\Delta x h''(0)\end{aligned} \tag{5.7}$$

The quad-cell signal is a non-linear function of the image displacement Δx. The non-linearity, shown in Figure 5.2, is of the form of a saturation curve. For small displacements, it is approximately linear, with the signal gain (slope) being determined by the shape of the speckle image. For larger displacements, the signal saturates, and it is no longer possible to estimate the magnitude of the image displacement.

Using the linear approximation, the wavefront slope, obtained by dividing the image displacement over the focal length f (from Equation 5.2), is

$$\hat{W}_x = \frac{1}{2Nh''(0)f}(A_2+A_4-A_1-A_3) = \frac{1}{2Nh(0)}(A_2+A_4-A_1-A_3) \tag{5.8}$$

with the corresponding formula $(A_3+A_4-A_1-A_2)$ for the slope in the y-direction (refer to Figure 5.2). $h(0)$ is the 1D angular spectrum $\int_{-\infty}^{\infty} h(0,v)\,dv$.

The sensitivity of the quad-cell, $2Nh(0)$, measures the ratio between changes in the wavefront slope and the corresponding changes to the quad-cell intensity measurements. This is the reciprocal of the gain, which is a scale factor tuned during operation to estimate the slope from the quad-cell intensity measurements.

5.2.2 Slope estimation errors

In the presence of photon noise, with an expected photon count of N, the signal in each quad-cell is independent of the other quad-cells with a variance of $\frac{N}{4}$. The variance of $(A_2+A_4)-(A_1+A_3)$ is the sum of their individual variances, giving N. Thus combined,

the quad-cell slope variance (error due to photon noise) is [66, 100]

$$E_p = \left\langle (\hat{W}_x - W_x)^2 \right\rangle_p = \left(\frac{1}{2Nh(0)} \right)^2 N = \frac{1}{4Nh(0)^2} \tag{5.9}$$

In practice, the quad-cell signal is also corrupted by detector read noise. Assuming independent and uniform read noise in each quad-cell detector of σ_r^2, the slope variance due to read noise is

$$E_r = \left\langle (\hat{W}_x - W_x)^2 \right\rangle_r = \left(\frac{1}{2Nh(0)} \right)^2 4\sigma_r^2 = \frac{\sigma_r^2}{N^2 h(0)^2} \tag{5.10}$$

In both error expressions, $h(0)$ can be expressed as $\Gamma h_o(0)$, where $h_o(0)$ is the diffraction-limited image peak, while Γ is the 1D analogue of the Strehl ratio.

The performance of the quad-cell is derived by Tyler and Fried [100], assuming diffraction-limited imaging ($\Gamma = 1$) when the image is an Airy disc (this has the form $\mathrm{Jinc}(x)^2$, as explained in Equation 2.64). Additionally, analytical solutions to diffraction-limited images of extended round objects were also given.

The results from [100] may be summarised more simply by using Equation 5.9 and some identities. In a $\mathrm{Jinc}(r)^2$ circularly symmetric image, the volume under the surface is $\frac{\pi}{4}$ (Equation 2.22), and the image height is $\frac{\pi}{4}$ (Equation 2.19). The maximum height of its 1D projection, $\int_{-\infty}^{\infty} \mathrm{Jinc}(\sqrt{x^2+y^2})^2 \, dy$ at $x = 0$ is given by Equation 2.20 as $\frac{2}{3}$. The first zero of the image is at $r = 1.22$. By appropriately scaling the dimensions of the $\mathrm{Jinc}(r)^2$ function to match the image at the telescope focal plane, we can derive the focal plane image peak due to a circularly symmetric aperture.

For a circular telescope aperture of diameter D, with a diffraction-limited image (of an Airy disc) (Equation 2.17) with its first zero crossing at $1.22\frac{\lambda}{D}$ radians, and photon count of 1 (corresponding to the volume), the peak of the 1D projection is

$$\begin{aligned} h(0) &= \left(\frac{\frac{2}{3}}{\frac{\pi}{4}}\right)\left(\frac{1.22}{1.22\frac{\lambda}{D}}\right) \\ &= \frac{\left(\frac{8}{3\pi}\right)}{\frac{\lambda}{D}} \end{aligned} \tag{5.11}$$

Substituting $h(0)$ into the quad-cell displacement estimator variance (Equation 5.9) results in

$$\begin{aligned} E_p &= \frac{\left(\frac{3\pi}{16}\right)^2}{N}\left(\frac{\lambda}{D}\right)^2 \\ &\approx \frac{0.35}{N}\left(\frac{\lambda}{D}\right)^2 \end{aligned} \tag{5.12}$$

For comparison, the commonly used Gaussian approximation of the image on the quad-cell, as used in Welsh and Gardner [110] or Parenti and Sasiela [69] (which had actually started from the image Strehl), is shown here.

$$\begin{aligned} h(u,v) &= \frac{1}{2\pi\sigma^2}e^{-\frac{u^2+v^2}{2\sigma^2}} \\ h(u) &= \int_{-\infty}^{\infty} h(u,v)dv = \frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{u^2}{2\sigma^2}} \end{aligned} \tag{5.13}$$

where σ is the width of the image, and $h(0)$ is $\frac{1}{\sqrt{2\pi}\sigma}$.

For a circular telescope aperture of diameter D and diffraction-limited imaging, the best Gaussian approximation is for $\sigma = 0.43\frac{\lambda}{D}$ radians. Using the Gaussian approximation to the Airy disc, the error contributions to the slope estimate due to photon noise is given by

$$E_p = \frac{1}{4Nh(0)^2} = \frac{0.29}{N}\left(\frac{\lambda}{D}\right)^2 \tag{5.14}$$

The read noise contribution is

$$E_r = \frac{\sigma_r^2}{N^2 h(0)^2} = \frac{1.16\sigma_r^2}{N^2}\left(\frac{\lambda}{D}\right)^2 \tag{5.15}$$

Unlike the photon noise variance which, obeying Poisson noise statistics, is inversely proportional to the total intensity illuminating the quad-cell, the read noise is inversely proportional to the squared intensity.

5.3 Fundamental bound on quad-cell performance

The main properties of a quad-cell are the extent of its linear region, and the sensor gain or sensitivity, which affects the signal-to-noise ratio. They are dependent on the shape of the image and the operating light level. The operating performance of a quad-cell is determined by the light level. However, image shape, which is also critical, has a less clear impact on performance.

The sensitivity of a quad-cell is $2Nh(0)$, where the value of $h(0)$ can be approximated by the maximum value or peak of the image. For a Gaussian image, the image peak varies in inverse proportion to the image width. Hence, it is common for the image width to be used as an indication of the sensor sensitivity. However, as shown above, the image peak is the more direct performance measure, and for irregularly shaped images (more common in closed loop adaptive optics systems after partial correction), is the correct and more accurate quantity to use.

The amplitude of the image is more conveniently expressed as a fraction of the peak amplitude of the diffraction-limited image. It is in fact the 1D analogy of the conventional 2D Strehl ratio. From simulations, it was found empirically that $\Gamma_{1D} \approx \Gamma_{2D}^k$ in Kolmogorov turbulence, where k is around 0.6 to 0.7.

5.3.1 Cramer-Rao Lower Bound

Given a set of observations (photon locations) derived randomly from the image, the estimation of image displacement can be formulated as a statistical estimation problem. Using the Fisher information (leading to the Cramer-Rao lower bound) of a probability density function, we can quantify the fundamental limit on the performance of a displacement estimator. With a lower Cramer-Rao bound, the slope estimate can potentially be more precise. Equivalently, fewer photons are required to achieve the desired level of precision.

The Cramer-Rao bound provides a useful comparison with the estimator variance provided by the quad-cell, quantifying the loss of information in the quad-cell arising from its coarse sampling. In contrast, as pointed out in Section 5.1.1, the centroid estimator has an infinite variance, and so has limited use as a benchmark for comparison.

The quad-cell image is a probability density function parametrised by its position $\theta = x'$, which is the displacement in 1D. The Fisher information of the image $f(x,y|\theta)$ is given by

$$J = \left\langle \left[\frac{\partial}{\partial \theta} \ln f(\boldsymbol{x}|\theta) \right]^2 \right\rangle = - \left\langle \frac{\partial^2}{\partial \theta^2} \ln f(\boldsymbol{x}|\theta) \right\rangle \tag{5.16}$$

Here, the expression for the Fisher information may be simplified further, since the parameter θ simply describes a translation of the density function $f(x,y|\theta = x') = f(x-x',y)$,

$$\begin{aligned} \frac{\partial}{\partial \theta} \ln f(x,y|\theta = x') &= \frac{\partial}{\partial x'} \ln f(x-x',y) \\ &= -\frac{\partial}{\partial x} \ln f(x-x',y) \end{aligned} \tag{5.17}$$

The expectation is taken over all points (x,y). The shape of the image remains unchanged when shifted, so the expectation is independent of the position x', which can be ignored. The CRLB when observing 1 photon is the inverse of the Fisher information.

$$\sigma^2_{\hat{x}'} \geq \frac{1}{\left\langle \left[\frac{\partial}{\partial x} \ln f(x,y) \right]^2 \right\rangle} = \frac{1}{\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \left[\frac{\partial}{\partial x} \ln f(x,y) \right]^2 f(x,y)\, dx\, dy} \tag{5.18}$$

A more realistic comparison with the quad-cell would restrict the image intensity distribution to 1D, since the quad-cell measurement $(A_2 + A_4) - (A_1 + A_3)$ is fully integrated over 1 axis, and is unable to measure the full 2D shape of the image. The 1D CRLB is given by

$$\sigma^2_{\hat{x}'-1D} \geq \frac{1}{\left\langle \left[\frac{\partial}{\partial x} \ln f(x) \right]^2 \right\rangle} = \frac{1}{\int_{-\infty}^{\infty} \left[\frac{\partial}{\partial x} \ln f(x) \right]^2 f(x)\, dx} \tag{5.19}$$

where $f(x) = \int_{-\infty}^{\infty} f(x,y)\, dy$.

Equation 5.19 shows that the Cramer-Rao bound for any image displacement estimator depends only on the shape of the image. The best images for displacement estimation has low Cramer-Rao bounds. The denominator in the CRLB, $\left[\frac{\partial}{\partial x}\ln f(x)\right]^2$, is maximised by images with strongly varying profiles or slopes. Similarly, the displacement of smooth images (which are highly blurred) is harder to estimate. In simulations, Equation 5.18 and Equation 5.19 are computed numerically from random speckle images because of the lack of an analytical formula for a random speckle.

5.4 Signal modulation and extended objects

The sensitivity and linear range of a quad-cell is dependent on image shape, which is dependent on atmospheric turbulence and the effects of adaptive optics compensation. Sometimes, the sensitivity of the quad-cell may be too high, and the image will be difficult to position on the centre of the quad-cell. Here, we show how the image shape can effectively be modified using modulation to provide more control over the operating range of the quad-cell. Modulation in a quad-cell signal reduces its sensitivity and increases the linear range [22, 76].

The signal from the quad-cell can be modulated by oscillating the image over the quad-cell in a periodic motion using an oscillating tip/tilt mirror. The ideal modulation path is a diamond shaped traverse that spends the same amount of time over each quadrant of the quad-cell. Practical modulation schemes approximate this with a circular path, as shown in Figure 5.4.

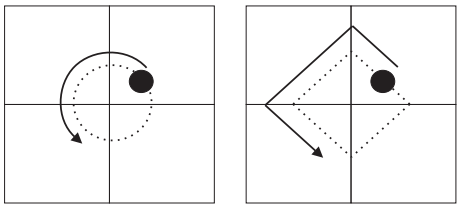


Figure 5.4 Modulation by displacing the image at the focal plane along a path. In practical implementations, the circular path on the left approximates the diamond shaped path typically used for analysis (right).

During modulation, the image on the quad-cell is displaced depending on its position along the modulation path l. The normalised image is now

$$h'(x,y,l) = h'(x-x',y-y') \tag{5.20}$$

where x' and y' are the displacement along the x and y axes for each position l along the modulation path.

To analyse the properties of a modulated quad-cell, the problem can be reduced to the estimation of a 1D displacement Δx. At each modulation position, the signal from the quad-cell, extending Equation 5.7, is

$$(A_2 + A_4) - (A_1 + A_3) = -2N \int_0^{-\Delta x} h''(x - x')\, dx \tag{5.21}$$

where $h''(x) = \int_{-\infty}^{\infty} h'(x,y)dy$.

The full modulated signal is obtained by integrating the quad-cell signal over the whole modulation path.

$$\oint m(l)\, ((A_2 + A_4) - (A_1 + A_3))\, dl \tag{5.22}$$

where the quad-cell signal (as given in Equation 5.21) is dependent on the modulation path position l.

$m(l)$ is the modulation function, representing the weighting for the time spent in each modulation position along the x-axis. Here, we see the advantage of using a diamond modulation path, since the "projected" modulation sweep speed is constant, so $l = x'$.

$$m(x') = \begin{cases} \frac{1}{l_m} & \text{for } -\frac{l}{2} < x' < \frac{l}{2} \\ 0 & \text{elsewhere} \end{cases} \tag{5.23}$$

where l_m represents the modulation width.

Equation 5.22 effectively blurs the image over a larger area on the quad-cell, as shown in Figure 5.5.

Using a diamond modulation path, the modulated signal is

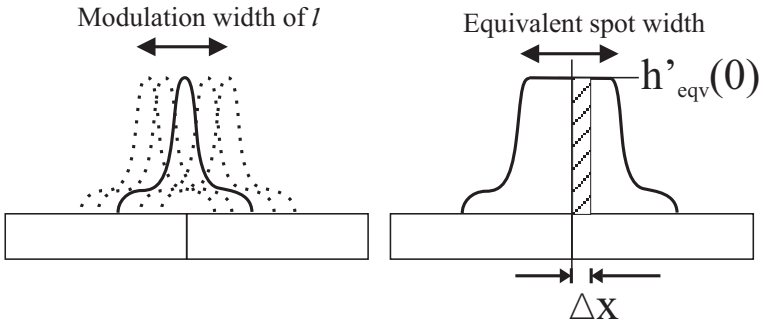


Figure 5.5 Shown in 1D, the modulation function blurs the image (left) in a controlled manner, producing the equivalent image with a rectangular shape on the right. For illustrative purposes, the equivalent image function has not been normalised, so it has a larger area under its curve - for analysis, the area under each image should be held constant.

$$\begin{aligned}
&\int_{-\frac{l}{2}}^{\frac{l}{2}} m(x')\,((A_2+A_4)-(A_1+A_3))\,dx' \\
=\;& \int_{-\infty}^{\infty} m(x')\,\left(-2N\int_0^{-\Delta x} h''(x-x')\right)dx\,dx' \\
=\;& -2N\int_0^{-\Delta x}\int_{-\infty}^{\infty} m(x')\,h''(x-x')\,dx'\,dx \\
=\;& -2N\int_0^{-\Delta x}\left[m(x)\odot h''(x)\right]\,dx
\end{aligned} \tag{5.24}$$

Indeed, as pointed out by many authors [46, 100], the blurring caused by the modulation in Equation 5.24 is equivalent to imaging with an extended object $o(x,y)$. In that case, instead of dealing with the point-spread-function $h'(x,y)$, the image at the focal plane is expressed as a convolution of the object with the point-spread-function. By substituting with $o(x,y)\odot h'(x,y)$, the modulated signal is (by Equation 2.40)

$$-2N\int_0^{-\Delta x}\left[o'(x)\odot h''(x)\odot m(x)\right]dx \tag{5.25}$$

with $o'(x)=\int_{-\infty}^{\infty} o(x,y)\,dy$ being the 1D distribution of the extended object, and $m(x)$ the modulation function defined above.

Comparing Equation 5.7 to Equation 5.24 and Equation 5.25, we see that the effect of extended objects, modulation, or both, can be simplified by assuming an equivalent image at the quad-cell.

$$h''_{eqv}(x) = o'(x) \odot h''(x) \odot m(x) \tag{5.26}$$

The effect of a modulation is to linearise the response of a quad-cell and reduce its sensitivity in a controlled fashion. Typically, the modulation width is selected to be larger than the image width itself, so the exact shape of the image no longer matters. For a modulation width l, the height of the equivalent image $h''_{eqv}(0)$ is then $\frac{1}{l}$. The slope estimate from a modulated quad-cell is given by

$$\begin{aligned} \hat{S}_x &\approx \frac{(A_2+A_4)-(A_1+A_3)}{2Nh''_{eqv}(0)f} \\ &\approx \frac{l}{2Nf}(A_2+A_4-A_1-A_3) \end{aligned} \tag{5.27}$$

The spatial modulation width l is scaled by the telescope focal length f to give the equivalent angular modulation width of $\frac{l}{f}$ radians. The slope estimate is no longer dependent on the image shape, and is now linear over the wider range of $\frac{l}{f}$ radians.

The trade-off under modulation is the decreased sensitivity of the quad-cell, so the slope variance increases with the modulation width. The slope variance under modulation is

$$E_p = \frac{l^2}{4N^2f^2}\text{var}\{A_2+A_4-A_1-A_3\} = \frac{l^2}{4Nf^2} \tag{5.28}$$

5.4.1 Circular modulation paths

The diamond shaped path used for analysis above is not smooth enough for use in physical systems, where circular paths are used instead. For a circular modulation, the weighting function $m(l)$ is equally weighted over a circular path. The modulated signal is

$$\int_0^{2\pi} \frac{1}{2\pi}\left((A_2+A_4)-(A_1+A_3)\right)d\theta \tag{5.29}$$

where the quad-cell signal is dependent on the modulation position.

Expressing Equation 5.29 in transformed rectangular coordinates, we arrive at the equiva-

lent modulation weighting function in 1D by integrating along the circular path parametrised by the angle θ ($l' = \frac{l}{2}\cos\theta$).

$$\int_{-\frac{l}{2}}^{\frac{l}{2}} \frac{1}{\pi}\left((A_2+A_4)-(A_1+A_3)\right)\frac{1}{\sqrt{(\frac{l}{2})^2-l'^2}}\,dl' \tag{5.30}$$

compared with Equation 5.22, this gives $m(l') = \frac{1}{\pi\sqrt{(\frac{l}{2})^2-l'^2}}$.

Equivalent alternatives to image modulation have been suggested, and include using diffuser plates [77], imaging of extended objects [46], or using the blurring caused by atmospheric turbulence itself [21].

5.5 Closed-loop operation

The above analysis, in common with most published analyses, assumes a constant image at the focal plane on the quad-cell. For a more complete treatment, we are also interested in the behaviour of the quad-cell when the image shape is a function of random turbulence. In practice, the time averaged performance of the quad-cell is not only a function of turbulence, but also the characteristics of the closed-loop adaptive optics system used.

5.5.1 Statistical analysis of quad-cell performance

The performance of the quad-cell in open loop is found from the ensemble average of the slope variance (error) over the turbulence process and also the photon arrival process.

In a closed-loop system, the performance of the quad-cell is linked to other components in the system, although it is often attributed only to the quad-cell sensitivity [22, 75]. The control system of a closed-loop wavefront compensation system is shown in Figure 5.6. Successive wavefront estimates are added to the current wavefront estimate through a correcting deformable mirror, allowing the system to track the ever changing turbulence. The integrating function of the correcting mirror results in a control system with an internal state. A complete analysis of such a closed-loop system using control theory is presented by Parenti and Sasiela [69, 116]. Here, we have simplified the analysis by considering only the steady state response of a closed-loop system. This approximates slowly varying turbulence or equivalently a fast loop response, and is sufficient to illustrate the performance

improvement compared to open loop conditions.

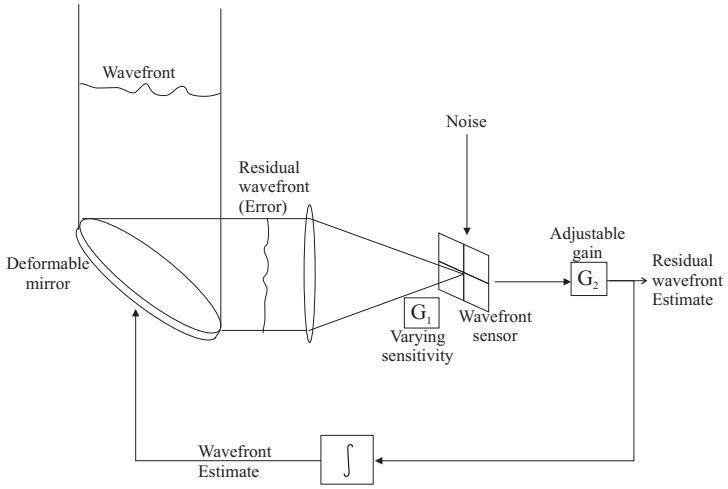


Figure 5.6 Control path of a wavefront sensor in a closed-loop adaptive optics system. Here, $G_1 = \frac{1}{2Nh(0)}$ (refer to Appendix) is the changing sensitivity of the quad-cell caused by the changing image, and G_2 is the adjustable feedback loop gain.

During normal operation, the temporal slope signal is estimated from the quad-cell signal through an adjustable gain G_2 (analogous to G_2 in Figure 5.6 without loop closure), which also determines the slope variance. After adjustment, the optimal value for this gain, which now remains constant over the course of operation, minimises the slope error,

$$\begin{aligned} \left\langle\left\langle(\hat{W}_x - W_x)^2\right\rangle\right\rangle &= \left\langle\left\langle(G_2(A_2 + A_4 - A_1 - A_3) - W_x)^2\right\rangle\right\rangle \\ &= \left\langle\left\langle(G_2(2Nh(0)W_x + n_p) - W_x)^2\right\rangle\right\rangle \\ &= \left\langle\left\langle(G_2(G_1 W_x + n_p) - W_x)^2\right\rangle\right\rangle \end{aligned} \tag{5.31}$$

where n_p represents the photon noise term in the quad-cell signals ($A_4 + A_2 - A_1 - A_4$), and $G_1 = 2Nh(0)$ the sensitivity or gain of the quad-cell. The expectations are taken over the random wavefront and photon arrival processes.

The wavefront slope and photon noise distributions both have zero mean. Assuming no correlation between the wavefront slope, photon noise, and image ($\langle W_x h(0)\rangle = \langle W_x n_p\rangle = \langle n_p h(0)\rangle = 0$), the optimal value of G_2 is

$$G_2 = \frac{\langle G_1\rangle \langle W_x^2\rangle}{\langle G_1^2\rangle \langle W_x^2\rangle + \langle n_p^2\rangle} \tag{5.32}$$

A simpler solution, using $\langle G_2(A_2+A_4-A_1-A_3)-W_x\rangle = 0$ or

$$G_2 = \frac{1}{\langle G_1\rangle} = \frac{1}{2N\langle h(0)\rangle} \tag{5.33}$$

can be used instead. Keeping the combined feedback loop gain constant, this solution represents an approximation to Equation 5.32 when $\langle n_p^2\rangle$ is very small compared to the other quantities, and the variance of $h(0)$ is small compared to $\langle h(0)\rangle^2$.

The slope error expression of Equation 5.9 remains valid as a special case of Equation 5.33 when the image speckle is unchanging. Note that the short term exposure equivalent of $G_2 = \left\langle \frac{1}{G_1} \right\rangle = \left\langle \frac{1}{2Nh(0)} \right\rangle$ for Equation 5.33 cannot be realised in practice, since over the time scale involved, the system gain G_2 is static.

5.6 Non-linear errors in the quad-cell

The derivation of the quad-cell error in Equation 5.9 assumes that the image displacement is small. The linear approximation in Equation 5.7 is more exactly

$$\begin{aligned} \hat{W}_x &= \frac{(A_2+A_4)-(A_1+A_3)}{2Nh(0)} \\ &= \frac{-2N\int_0^{-\Delta x} h(x)\,dx}{2Nh(0)} \\ &= \frac{-\int_0^{-\Delta x} h(x)\,dx}{h(0)} \end{aligned} \tag{5.34}$$

When the image displacement is large, the non-linearity in Equation 5.34 becomes significant. In the extreme case where the quad-cell signal is frequently over-saturated, the quad-cell signal can be simplified to a piecewise-linear approximation, where the signal is either saturated (constant) or linear with respect to image displacement.

$$(A_2+A_4)-(A_1+A_3) = -2N\int_0^{-\Delta x} h(x)\,dx = \begin{cases} N & \text{when } W_x > \frac{1}{2h(0)} \\ 2Nh(0)W_x & \text{when } -\frac{1}{2h(0)} < W_x < \frac{1}{2h(0)} \\ -N & \text{when } W_x < -\frac{1}{2h(0)} \end{cases} \tag{5.35}$$

In that case, the error is

$$e^2_{\hat{W}_x} = (\hat{W}_x - W_x)^2 = \begin{cases} \frac{1}{2h(0)} - W_x & \text{when } W_x > \frac{1}{2Nh(0)} \\ 0 & \text{when } -\frac{1}{2Nh(0)} < W_x < \frac{1}{2Nh(0)} \\ -\frac{1}{2h(0)} - W_x & \text{when } W_x < -\frac{1}{2Nh(0)} \end{cases} \tag{5.36}$$

and the expected error is

$$\left\langle e^2_{\hat{W}_x} \right\rangle = \int_{-\infty}^{\infty} e^2_{\hat{W}_x} P(W_x) dW_x \tag{5.37}$$

5.7 Quad-cell performance comparisons

In this section, the behaviour of Equation 5.9 in turbulence is estimated using a simulation. From the simulation, the slope variance from the quad-cell is compared to the theoretical Cramer-Rao lower bound for slope estimators. Additionally, we also confirm the simulation results by comparison with Yura's ([115]) approximation for tip/tilt corrected image profiles.

In the simulation, we model the effects of atmospheric turbulence as wavefront aberrations with Kolmogorov statistics. A sample of Kolmogorov phase-screens at various $\frac{D}{r_0}$ is generated using the fractal method of Harding and Johnston [42]. We assume a single layer of turbulence at the telescope aperture plane, which is focused onto a quad-cell. The peak of the turbulence degraded image on the quad-cell then determines the variance of the slope estimator as given by Equation 5.9. The average image peak at each turbulence level describes the performance of the quad-cell. As shown in Equation 5.33, the averaged slope variance is $\frac{1}{2N\langle h(0)\rangle}$.

At the same time, the shape of the image distribution at the focal plane also determines the Cramer-Rao bound for the slope estimate, as given by Equation 5.18 and Equation 5.19. In this simulation, the CRLB is calculated by discrete numerical differentiation. Again, the CRLB is averaged over all simulated phase-screens at each level ($\frac{D}{r_0}$) of turbulence.

5.7.1 Tip/tilt compensated approximation

Using the rule-of-thumb that atmospheric turbulence degrades and reduces the resolution of a large telescope to be equivalent to a smaller telescope with diameter r_0, the resolution of the image at the focal plane is approximated by $\frac{1.22\lambda}{r_0}$. In fact, the long-term exposure image is Gaussian shaped, with the best fit to the image when the width (standard deviation) of the Gaussian is $\frac{1}{2\sqrt{2}}\frac{1.22\lambda}{r_0} \approx 0.43\frac{1.22\lambda}{r_0}$.

In Yura's work [115], the effect of tip/tilt compensation on the image size is accounted for by the enlargement of r_0 to $(1+0.37\left(\frac{r_0}{D}\right)^{\frac{1}{3}})$. For this centroid[1] based slope estimator, the variance, as given by the best fitting Gaussian, is

$$\text{Std.dev}\{\hat{W}_x\} = \begin{cases} \frac{1.22\lambda}{2\sqrt{2}r_0(1+0.37\left(\frac{r_0}{D}\right)^{\frac{1}{3}})} & \text{for } \frac{D}{r_0} > 1 \\ \frac{1.22\lambda}{2\sqrt{2}D} & \text{for } \frac{D}{r_0} < 1 \end{cases} \tag{5.38}$$

When $\frac{D}{r_0} < 1$, for low levels of turbulence, the effects of turbulence are negligible, so the variance is limited by the size of the telescope aperture D instead of r_0.

Simulation results

The measured quad-cell errors, CRLB and Yura's theoretical approximations are shown in Figure 5.7. Not surprisingly, with higher turbulence when the focal plane image is highly blurred, the estimation error increases. The measured errors agree very closely with Yura's approximation, confirming the validity of our approach. Compared to the CRLB, the errors are not more than a few times larger than the theoretically achievable minimum, so using a quad-cell for slope sensing represents an acceptable trade-off, given its simplicity.

Based on the previous assumption of the peak of each image being centered on the quad-

[1] Given a Gaussian profile, the centroid estimator no longer has infinite variance, and is in fact the optimal displacement estimator. Note that in this model, the estimator performance is derived from the *image width*, but is equivalent to the formulation based on image height since the image shape is fixed.

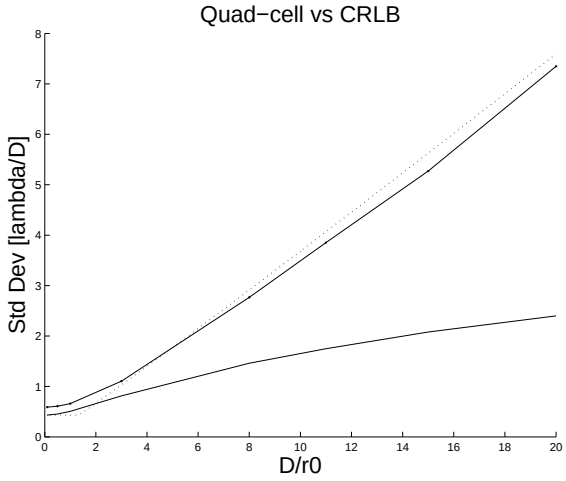


Figure 5.7 Errors in the slope estimate of a quad-cell due to photon noise (solid-dotted line) as compared to the CRLB in 1D (solid line) and Yura's approximation (dotted). The slope standard deviation is expressed in multiples of $\frac{\lambda}{D}$ [rad].

cell, the expressions developed here are only valid for small image displacements. Under open loop conditions, when the randomly displaced images are grossly misaligned with the centre of the quad-cell, an additional non-linear error is introduced. The exact value of $h(0)$ is also subject to the randomness of each image, so its difference from the average quad-cell sensitivity will give rise to further errors. These errors are collectively grouped into the non-linear error term, and will be included in simulations of the wavefront sensors in the following sections.

5.8 Conclusion

After examining the direct centroiding approach for calculating image displacement, and encountering problems with photon and read noise, we reduce the displacement or slope estimator to a quad-cell. Assuming small image displacements in the quad-cell, a linear approximation is use to examine the estimation errors in the quad-cell. The critical factor affecting the performance of the quad-cell is the image shape on the quad-cell. The closed-loop behaviour of the quad-cell is abstracted to a model of the image shape on the quad-cell. The quad-cell modulation process can also be described as a shape-manipulation operation for adjusting the performance of the quad-cell.

In this chapter, a simulation of the errors in the quad-cell is compared with the fundamental performance limit for any image displacement estimator, the Cramer-Rao lower bound. The work in Yura [115] is also extended and modified slightly to provide a second data-point for

comparison and validation. This leaves us satisfied that the quad-cell is most appropriate as a practical image displacement estimator.

The slope measurement process in the quad-cell forms the basis for wavefront sensing in the Shack-Hartmann and the pyramid wavefront sensors. Having studied the behaviour of the quad-cell, the extension to wavefront sensing of higher order modes is straight-forward, and we can begin to examine these wavefront sensors in the next chapter.

Chapter 6

Comparison of the Shack-Hartmann and pyramid wavefront sensor

The previous chapter has shown that the precision of the quad-cell slope estimate is determined by the image intensity and Strehl ratio. This chapter looks at the subdivision operation used to split the basic quad-cell arrangement (Section 4.3.1) into smaller problems. Global slope estimation with quad-cells can be extended to estimate higher order modes in wavefront aberrations by subdividing the wavefront aberrations at the telescope aperture into smaller regions, or subdividing the imaging plane and re-imaging, as in the pyramid sensor. All the wavefront sensors introduced in Chapter 4 subdivide the complex field, but this is performed along the optical path at different positions.

6.0.1 Resolution and precision

The subdivision of a complex field forms the common basis of both wavefront sensors examined in this chapter. In estimating the overall wavefront function, two important factors to consider are the resolution and precision of the wavefront estimate. In a quad-cell, the precision of the wavefront slope estimate at the telescope aperture plane is determined by the image Strehl and intensity at the telescope focal plane. Large telescope apertures (more light), or small levels of turbulence (higher Strehl), result in more precise wavefront estimates.

In the original quad-cell arrangement, only a plane of best fit to the wavefront, derived from the global wavefront slope estimate, is available. Through a subdivision process,

more slope measurements within the same area can be obtained. The spatial resolution of the wavefront refers to the spatial sampling of the wavefront estimate within the aperture. At higher resolutions, the sensor estimate can approximate the wavefront more closely, allowing more types of aberrations to be corrected.

Given a finite amount of light, the subdivision size and position is crucial to achieving optimal performance. Precision and resolution are in fact determined by a space-bandwidth trade-off. We examine here the implications of the space-bandwidth trade-off in wavefront sensors.

6.1 The Fourier Transform in wavefront sensors

In an optical system aimed at a point-source object, the complex fields between the aperture and focal planes are related by the Fourier transform. Propagating a complex field $A(u,v)e^{i\phi(u,v)}$ from the telescope aperture plane, where $A(u,v)$ is the aperture magnitude function, and $\phi(u,v)$ is the phase function, results in a complex field $u(x,y)$ at the focal plane given by (Equation 3.28)

$$u(x,y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} A(u,v)e^{-i\phi(u,v)}e^{-i\frac{2\pi}{\lambda f}(ux+vy)}\,du\,dv = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} p(u,v)e^{-i\frac{2\pi}{\lambda f}(ux+vy)}\,du\,dv \tag{6.1}$$

where (u,v) and (x,y) represent the spatial coordinates in the aperture and focal planes respectively, λ is the wavelength of the monochromatic light source (the wavefront is given by $W(u,v) = \frac{2\pi}{\lambda}\phi(u,v)$), and f is the focal length of the optical system. The complex field $u(x,y)$ should not be confused with the coordinate u in the aperture plane.

The image at the focal plane is given by the squared-magnitude of $u(x,y)$. When normalised to sum to 1, it is the point-spread-function of the optical system.

$$h(x,y) = \frac{|u(x,y)|^2}{\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}|u(x,y)|^2\,dx\,dy} \tag{6.2}$$

Using the Fourier equivalence of functions in corresponding Fourier domains, we can better understand the operations of both wavefront sensors and explicitly compare their functions. The Fourier relationship between the aperture and focal planes enables us to derive dual

operators in each domain. As shown in Equation 4.13, the displacement at the spatial plane is linearly related to the wavefront slope in the Fourier plane. This duality between image displacement and wavefront slope forms the fundamental limitation of wavefront sensing.

The resolution and precision constraint is located in opposing Fourier spaces, and is subject to the space-bandwidth limitation. In the spatial domain, the spatial width of a signal is

$$\Delta x = \sqrt{\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x^2 \left|u(x,y)\right|^2 \, dx \, dy} \tag{6.3}$$

In the Fourier domain, the width of the corresponding spectrum is

$$\Delta u = \sqrt{\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} u^2 \left|p(u,v)\right|^2 \, du \, dv} \tag{6.4}$$

According to the uncertainty principle, the space-bandwidth product is a constant, and represents a more general limit to the precision that is achievable in any physical system. In wavefront sensors, this represents the sampling between the spatial and the Fourier frequency domain.

$$\Delta x \Delta u = \frac{1}{4\pi} \tag{6.5}$$

Depending on other system constraints that need to be satisfied, one can choose between having many high noise measurements, or fewer low noise measurements, while still satisfying the space-bandwidth limit.

6.2 Wavefront subdivision

The quad-cell at the focal plane of a telescope provides only a global slope measurement. The quad-cell can be replaced with a transmissive pyramidal prism to re-image the aperture, as seen in Figure 6.1(a). Integrating the total intensity in each aperture image will reproduce what is still equivalent to the quad-cell, providing a way to measure the global wavefront slope. However, since images of the aperture are now available, more local slope variations within the aperture can be measured [9, 76].

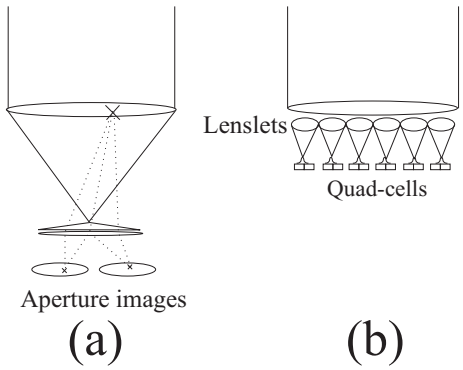


Figure 6.1 Extension of the quad-cell to estimate higher order wavefront slopes.

Alternatively, the complex field at the telescope can be directly divided into separate regions using a lenslet array in the aperture, and refocused, as shown in Figure 6.1(b). The image displacement at the focal plane of each sub-region corresponds to the local wavefront slope over that region [72,85]. The subdivision operation may also be implicit, as in the curvature and geometric sensors, where the equivalent subdivision occurs at an intermediate position between the aperture and focal planes. Chapter 7 examines the curvature and geometric wavefront sensors.

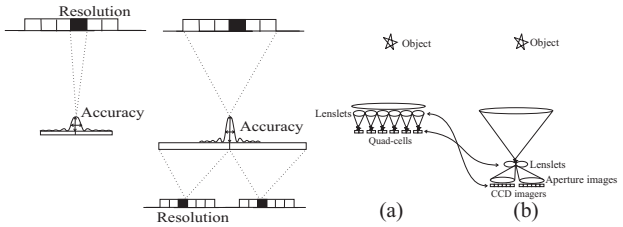
This chapter examines two extreme positions for subdividing the complex field, as represented by the Shack-Hartmann (aperture subdivision, [12, 16]) and pyramid wavefront (focal plane subdivision) sensors. In the following sections, we consider the effects of photon and read noise on the slope estimation errors in both types of wavefront sensors. By extending Equation 5.9 to the measurement of local slopes (for higher order aberrations), we may directly compare the slope estimation performance of two wavefront sensors.

The Fourier duality between the two wavefront sensors also provides additional insight into their similarities and differences. Many operations in both wavefront sensors can be shown to be equivalent. However, there are also critical differences that confer advantages to the pyramid wavefront sensor.

6.2.1 Resolution and precision of wavefront sensors

Figure 6.2 shows a simplified layout of the Shack-Hartmann sensor and a focal plane lenslet array or the pyramid sensor. In the Shack-Hartmann sensor, a lenslet array produces multiple images of an object through the telescope aperture. In contrast, the focal plane lenslets

re-image the aperture plane through the equivalent of a 2x2 lenslet array at the focal plane, creating multiple images of the aperture.



(a) The relationship between resolution and precision in wavefront sensors.

(b) The duality and equivalence between the subdivision and slope detection in wavefront sensors.

Figure 6.2 A comparison of the Shack-Hartmann and pyramid wavefront sensors.

In wavefront sensors, the resolution of the slope estimate is inversely proportional to the spatial extent over which the local slopes are estimated, while the precision, or variance of the estimates, is determined by the measurement fluctuations caused by photon noise. The aperture subdivision operation in the Shack-Hartmann is a rectangular windowing operation. This is equivalent to convolution with the sinc function in the lenslet focal plane (Equation 2.53), or a blurring operation.

In both wavefront sensors, a displacement measurement at the focal plane corresponds to a wavefront slope measurement in the aperture plane. To be more precise, the displacement measurement is initially performed by sampling the total intensity within the rectangular CCD arrays used to subdivide the measurement plane. This convolution and sampling operation corresponds to a multiplication (with a sinc) and sampling operation at the aperture plane. The intensity summing operation results in a loss in the higher frequency components in the recovered wavefront.

The Shack-Hartmann sensor subdivides the field at the aperture plane, and forms arrays of images of the object (assumed here to be an unresolved point-source) at the focal plane of the lenslets. The lenslet images are blurred by the subdivision operation at the dual aperture plane and thus enlarged, are focused onto arrays of quad-cells. Each quad-cell consists of 2x2 intensity detector sections that subdivide the image plane.

The pyramid sensor subdivides the complex field at the telescope focal plane, and re-images the telescope aperture onto CCD detectors. The image sampling process by the CCD array

implicitly subdivides the image. The aperture image subdivision has a direct analogy to the aperture subdivision with lenslets in the Shack-Hartmann sensor. Similarly, the focal plane subdivision operation is analogous to the quad-cell slope measurement operation in the Shack-Hartmann sensor. In contrast to the Shack-Hartmann sensor, slope sensing in the pyramid sensor occurs before aperture subdivision, so no blurring of the focal plane image occurs.

Although the similarity may not be obvious at first, the slope measurement operation is in fact identical in both wavefront sensors. In the Shack-Hartmann, this is performed by comparing the intensity within the quad-cells, while in the pyramid sensor, the displacement of the single focal plane image is derived from comparisons of the intensity measurements in each facet of the pyramidal prism.

6.3 Shack-Hartmann wavefront sensor

As shown in Figure 6.2(a), the spatial resolution of the wavefront at the aperture is given by the size of the lenslets of the Shack-Hartmann sensor —with more lenslets, more slope measurements are obtained, providing finer sampling of the wavefront.

On the other hand, with fewer larger lenslets, the images at the focal plane of the lenslets are smaller and brighter, providing better slope estimates (lower variance). Although larger lenslets have higher illumination and higher image peaks (a function of image shape), the higher illumination alone within a lenslet does not lead to any improvement in the precision of the global averaged slope. As will be shown in Section 6.3.1 (and summarised in Table 6.1), since the total illumination remains constant, the only improvement in overall precision arises from the higher image peaks in each lenslet.

A zonal wavefront estimate can be reconstructed by interpolating between the local sensor slope signals. Southwell [93] explored the different slope reconstruction geometry for interpolating between measurements and derived their respective error performances.

Alternatively, the wavefront estimate can be expressed in terms of the Zernike coefficients, giving rise to a modal estimate. Section 4.3.1 reconstructed the full wavefront estimate from sensor (in the Shack-Hartmann, local slope) measurements. Although the Zernike polynomials are orthogonal, their slopes are not, so a full matrix inversion is required to solve for wavefront coefficients. The optimal MAP solution for the Shack-Hartmann sensor is derived in Bakut et. al [10].

The following section derives a performance measure for the Shack-Hartmann based on the image Strehl, as an extension of the quad-cell, to facilitate a comparison with the pyramid sensor. The same statistical estimation framework is also used to examine the optimal subdivision size for the lenslets in the Shack-Hartmann sensor. This completes the discussion on the trade-off between the precision and resolution of wavefront estimates in the Shack-Hartmann sensor. The analysis here also provides the background for understanding the way the pyramid sensor "side-steps" the resolution-precision limit.

6.3.1 Shack-Hartmann slope errors

In the Shack-Hartmann sensor, the variance in the slope estimate for each lenslet is caused by two components, photon noise and read noise in the CCD detectors. Using the photon noise error expression of Equation 5.9 and Equation 5.10 (which assumes a read noise of σ_r^2 in each detector element of the quad-cells), the variance in $\hat{w}_{s_i}$, the slope estimate for a lenslet (averaging over the random photon and read-noise induced variations), is

$$\begin{aligned} e_s &= \left\langle (\hat{w}_{s_i} - w_{s_i})^2 \right\rangle = e_{s_p} + e_{s_r} \\ &= \frac{1}{4N_i h_i(0)^2} + \sigma_r^2 \frac{1}{N_i^2 h_i(0)^2} \end{aligned} \tag{6.6}$$

where e_{s_p} is the slope error due to photon noise and e_{s_r} is the slope error due to read noise. $h_i(0)$ is the peak of the projected angular spectrum of the lenslet ($h_i(x) = \int_{-\infty}^{\infty} h_i(x,y)dx$).

The global slope estimate at the aperture is formed by a weighted sum of the local slope signal in all lenslets. The weighting assigned to each lenslet signal is given by the proportion of the lenslet area to the total aperture area. Assuming there are M lenslets in the Shack-Hartmann sensor, the global slope estimate is

$$\hat{W}_s = \frac{\sum_{i=1}^{M} R_i \hat{w}_{s_i}}{\sum_{i=1}^{M} R_i} \tag{6.7}$$

where $\hat{W}_s$ is the global wavefront tilt estimate in the Shack-Hartmann sensor, and $\hat{w}_{s_i}$ is the local wavefront tilt estimate in the i^{th} lenslet. R_i represents the area of the i^{th} lenslet.

The photon count within a lenslet can be assumed to be proportional to the lenslet area,

giving

$$\hat{W}_s = \sum_{i=1}^{M} \frac{N_i}{N_{tot}} \hat{w}_{s_i} \tag{6.8}$$

with N_i being the photon count in the i^{th} lenslet, and $N_{tot} = \sum_i N_i$ being the total photon count over the whole aperture.

The total variance in the global mean slope given by the Shack-Hartmann sensor (from averaging the fluctuations due to photon noise) is

$$\begin{aligned} E_s &= \left\langle (\hat{W}_s - W_s)^2 \right\rangle = E_{s_p} + E_{s_r} = \sum_{i}^{M} \left(\frac{N_i}{N_{tot}} \right)^2 e_{s_i} \\ &= \frac{1}{N_{tot}^2} \sum_{i=1}^{M} \left(\frac{N_i}{4h_i(0)^2} + \frac{\sigma_r^2}{h_i(0)^2} \right) \end{aligned} \tag{6.9}$$

with the i in e_{s_i} to differentiate the local slope error between each lenslet.

Assuming a local slope estimator that is optimal in the statistical sense (the minimum variance unbiased estimator [51] that achieves the Cramer-Rao bound, as examined in Section 5.3) is available for each measurement, and the measurements in each sub-region are statistically independent, then the global slope estimated with the weighting proposed in Equation 6.7 is optimal, and forms a minimum variance unbiased estimator.

The quad-cell, due to under-sampling and measurement truncation, is not a minimum variance estimator (as demonstrated in Figure 5.7). Also, there is usually some correlation between the measurements in neighbouring Shack-Hartmann lenslets. So although the global slope estimate in Equation 6.7 is not a minimum variance estimate, it is a good estimate that compares well to the theoretical limit (Section 5.7.1).

6.3.2 Lenslet size

In this section, we examine the performance trade-off involved in varying the subdivision size in the Shack-Hartmann wavefront sensor, and show how one derives the optimal lenslet size. Frequently in closed-loop systems, the image displacements in the Shack-Hartmann quad-cell detectors are small, allowing non-linearities in the Shack-Hartmann quad-cell

detectors to be ignored, giving a linear wavefront modal estimator. From the quad-cell signals in each lenslet, a maximum-likelihood solution to the wavefront is obtained. In spite of its slightly lower performance, the maximum-likelihood solution is chosen in favour of a *maximum a posteriori* solution because it is sufficient for the analysis here and is simpler.

The performance of the Shack-Hartmann sensor is determined by the size of the lenslets used. A trade-off exists between larger lenslet sizes which provide more precise wavefront estimates, and smaller lenslet sizes, which increase the resolution of the wavefront estimate. To examine this trade-off, we compare the error terms in the Shack-Hartmann sensor over different lenslet sizes. The wavefront sensor is modelled with (from Section 2.5.2)

$$\boldsymbol{d} = \boldsymbol{H}\boldsymbol{\alpha} + \boldsymbol{n} \tag{6.10}$$

where $\boldsymbol{H}$ is the model of the wavefront sensor that includes the effect of subdivision size, and the noise in the slope measurements, $\boldsymbol{n}$, are modelled by zero-mean Gaussian noise (Equation 6.6). The Zernike decomposition of Kolmogorov turbulence $\boldsymbol{\alpha}$ are also zero-mean and take on Gaussian statistics.

The maximum-likelihood inverse of $\boldsymbol{H}$ is (Equation 2.119)

$$\boldsymbol{H}^{+} = (\boldsymbol{H}^{T}\boldsymbol{H})^{-1}\boldsymbol{H}^{T} \tag{6.11}$$

The forward and inverse matrices are generally not of full rank, so $\boldsymbol{H}^{+}\boldsymbol{H} = \boldsymbol{P}$ is not the identity matrix, but a projection matrix describing the detectable modes in the coefficient vector $\boldsymbol{\alpha}$. Terms in $\boldsymbol{\alpha}$ that are not detectable corresponds to zeroes in the $\boldsymbol{P}$ matrix.

The wavefront estimation error, taking the expectation over the turbulence $\boldsymbol{\alpha}$ and photon noise $\boldsymbol{n}$ is

$$\begin{aligned} & \left\langle trace\left((\boldsymbol{H}^{+}\boldsymbol{d} - \boldsymbol{\alpha})(\boldsymbol{H}^{+}\boldsymbol{d} - \boldsymbol{\alpha})^{T}\right)\right\rangle_{\boldsymbol{\alpha},\boldsymbol{n}} \\ = \;& trace\left((\boldsymbol{I} - \boldsymbol{P})\left\langle \boldsymbol{\alpha}\boldsymbol{\alpha}^{T}\right\rangle_{\boldsymbol{\alpha}}(\boldsymbol{I} - \boldsymbol{P})\right) + trace\left(\boldsymbol{H}^{+}\left\langle \boldsymbol{n}\boldsymbol{n}^{T}\right\rangle_{\boldsymbol{n}}\boldsymbol{H}^{+T}\right) \end{aligned} \tag{6.12}$$

We used the property that the turbulence and photon noise are zero-mean Gaussians and uncorrelated to each other $\langle \boldsymbol{\alpha} \boldsymbol{n}^T \rangle = \mathbf{0}$.

Simulation

Simulations of a Shack-Hartmann sensor is performed by generating 200 random Kolmogorov phase-screens ($\frac{D}{r_0} = 8$) as the turbulence, then measuring the sensor performance when estimating 8 Zernike modes in the turbulence. The wavefront estimates across different configurations of the Shack-Hartmann with different number of lenslets (ranging from 1, 2, 4, 8, 16, 32, to 64 lenslets across the telescope aperture of 256 pixels), adding Poisson noise with a mean of 800 photons (averaged over 50 photon noise frames for each turbulence instance), are then compared.

Figure 6.3 illustrates the trade-off between sensor precision and resolution, and shows each term of Equation 6.12 separately.

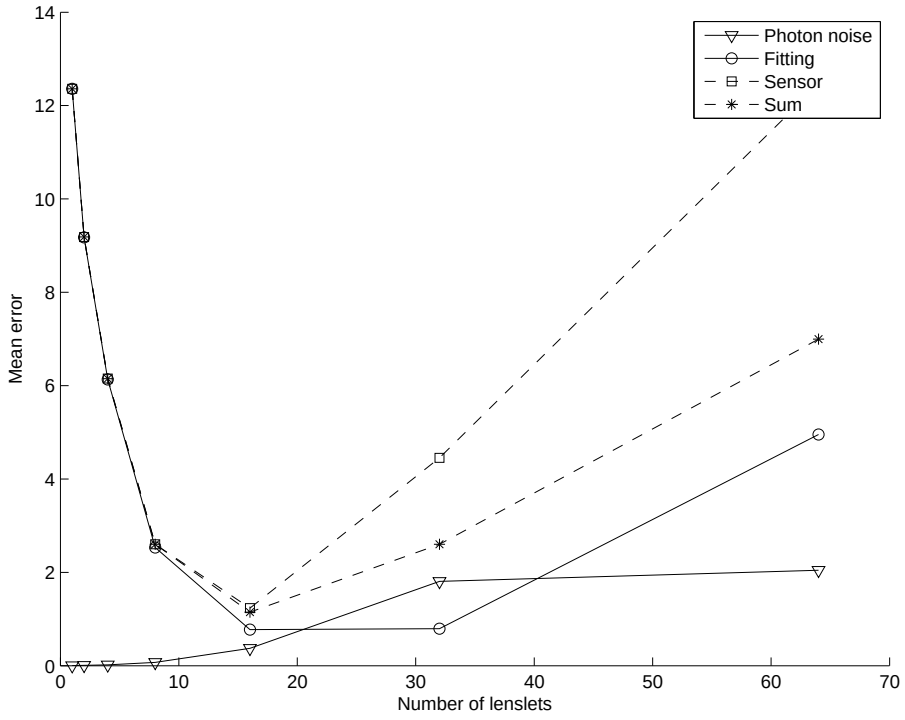


Figure 6.3 The effect of increasing the number of lenslet (and reducing their size correspondingly) in the Shack-Hartmann sensor.

Due to the limited resolution of the wavefront sensor, only a limited and finite number of Zernike modes can be estimated. The first term of Equation 6.12 quantifies this error, which is effectively a wavefront fitting error. This error is dependent on the Zernike coefficient

covariance matrix $C_{\boldsymbol{\alpha}} = \left\langle \boldsymbol{\alpha}\boldsymbol{\alpha}^T \right\rangle$, which is derived from the statistics of Kolmogorov turbulence. When more (smaller) lenslets (plotted over the x-axis of Equation 6.12) are used to subdivide the telescope aperture, more slope measurements can be made, increasing the sensor resolution and reducing the fitting error (shown as the dashed "fitting" error curve). The fitting error increases again when there are more than 32x32 lenslets used because of the increasing inaccuracy in modelling Zernike wavefront functions as discrete pixel grid elements[1].

The second term of Equation 6.12 is the photon noise error propagation term, and describes the precision of the wavefront estimate produced by the wavefront sensor (shown as the dot-dashed "Photon noise" error curve). More (smaller) lenslets produce larger images at their focal planes, reducing the precision of their slope estimate. Here, in contrast to the fitting error, having more lenslets lead to less precise slope estimates, corresponding to higher errors in the wavefront estimate. Again, modelling inaccuracies lead to a break in the error trends beyond 32x32 lenslets.

The sum of the fitting error and photon noise errors (the dotted "Sum" error curve) do not correspond to the actual measured sensor error (the solid "Sensor" error curve). The discrepancy is small and can be ignored as it arises from modelling inaccuracies due to the discretisation from pixelisation and increased non-linear errors at smaller lenslets sizes.

In summary, smaller lenslet sizes lead to more wavefront modes being detected, but with lower precision. Conversely, with larger lenslets, fewer wavefront modes are detectable, but with higher precision. The combined total error in the Shack-Hartmann sensor is minimised by matching the size of the lenslets to atmospheric turbulence. From this analysis of the trade-off between sensor precision and resolution, the optimal lenslet size is found to be related to the turbulence coherence length, r_0, confirming the rule-of-thumb used for sizing lenslets to match r_0 in the Shack-Hartmann sensor.

6.4 Pyramid wavefront sensor

For analysis purposes, the analogy between displacement estimation at the focal plane of the pyramid sensor and displacement estimation in quad-cells can be generalised to NxN

[1]The sampled, discretised Zernike polynomials (on a rectangular-array) are no longer mutually orthogonal. The residual errors due to the sampling process depend on the number of pixels used to represent the polynomials or the frequency content of the Zernike polynomial. Since the higher Zernike modes have a higher frequency content, the discretisation error is especially prominent at higher modes, so the number of modes simulated should be kept low.

centroid estimators, just as in the quad-cell [18]. In this equivalent arrangement, the pyramid sensor consists of an array of lenslets, subdividing the complex field at the focal plane. Each lenslet delineates a square section of the focal plane, through which the complex field is focused to form low-resolution images of the telescope aperture.

Mathematically, the propagation of the complex field at the telescope aperture to the telescope focal plane is described by an optical Fourier transform. At the focal plane, it is windowed or subdivided by the lenslet array, and each sub-region is then propagated again with a Fourier transform to the lenslet focal plane. Because the focal plane represents the frequency domain of the complex field at the aperture plane, the subdivision operation at the focal plane can be described by a filtering operation.

The lenslet windows act as two dimensional "brick-wall" filters in the frequency domain, so the equivalent operation in the spatial domain (after re-imaging the aperture) from the Fourier convolution-multiplication relationship is a blurring with the sinc kernel. This blurring or low-pass filtering is determined by the size and position of each lenslet.

The lenslet transmittance is a rectangular window with dimensions Δx by Δy, and centred on (x',y'), through which the complex field $u(x,y)$ is transmitted and re-imaged.

$$s(x,y) = \begin{cases} 1 & \text{for } (x'-\frac{\Delta x}{2}) < x < (x'+\frac{\Delta x}{2}),\ (y'-\frac{\Delta y}{2}) < y < (y'+\frac{\Delta y}{2}) \\ 0 & \text{otherwise.} \end{cases} \tag{6.13}$$

Each windowed complex field is propagated with another Fourier transform to the lenslet focal plane, where a blurred and inverted image of the telescope aperture is formed.

$$\begin{aligned} \alpha(\xi,\eta)e^{i\theta(\xi,\eta)} &= \mathscr{F}\{s(x,y)u(x,y)\} \\ &= \mathscr{F}\{s(x,y)\} \odot \mathscr{F}\{u(x,y)\} \\ &= \Delta x e^{-i\frac{2\pi}{\lambda f}x'\xi} sinc(\frac{\Delta x}{\lambda f}\xi)\Delta y e^{-i\frac{2\pi}{\lambda f}y'\eta} sinc(\frac{\Delta y}{\lambda y}\eta) \\ &\quad \odot A(-\xi,-\eta)e^{i\phi(-\xi,-\eta)} \end{aligned} \tag{6.14}$$

The extent of the blur is determined by the convolution kernel, a two-dimensional sinc function (first term of Equation 6.14). The wider the lenslets, the narrower the sinc function,

and the less blurring is present in their aperture images[2].

Returning to the problem of global mean slope estimation, we can ignore local slope distributions, and sum the image intensity over the whole aperture. The total intensity in each aperture image is (by Parseval's theorem) the same as the total intensity that passes through the corresponding lenslet.

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\left|\alpha(\xi,\eta)e^{i\theta(\xi,\eta)}\right|^2 d\xi d\eta = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}|s(x,y)u(x,y)|^2\ dxdy \tag{6.15}$$

This problem reduces to the familiar image displacement (centroid) estimation problem at the telescope focal plane.

A practical advantage to imaging with a CCD array in the pyramid sensor is the pixel binning function. Whether implemented in hardware or software, pixel binning easily allows for effectively variable pixel sizes. In the Shack-Hartmann sensor, this is equivalent to varying the size of the lenslets, a function that is not possible in practice.

6.4.1 Pyramid sensor slope errors

In this section, we restrict our attention to the pyramid wavefront sensor, which is a 2x2 quad-cell arrangement at the focal plane. The intensity over each quadrant in the focal plane is found from the total intensity of its corresponding aperture image.

In most analyses of the pyramid sensor performance, the image width is used as a measure of the sensitivity of the pyramid sensor [22,75]. In contrast, the analysis here uses the image height, as previously explained in Chapter 5. Using the quad-cell formula at the focal plane (Equation 5.9 and Equation 5.10), the error in the slope estimate is given by

$$\begin{aligned} E_p &= \left\langle(\hat{W}_p - W_p)^2\right\rangle = E_p^p + E_p^r \\ &= \frac{1}{4N_{tot}h_{tel}(0)^2} + P\sigma_r^2\frac{1}{N_{tot}^2 h_{tel}(0)^2} \end{aligned} \tag{6.16}$$

[2]Analytical approximations to Equation 6.14 have been derived by assuming a rectangular telescope aperture and a uniform wavefront slope within the telescope aperture [18]. Based on the assumptions outlined, the aperture images can be expressed as exponential integral functions defined to be $Ei(x) = -\int_{-x}^{\infty}\frac{e^{-t}}{t}\,dt$.

where the subscript p denotes the pyramid sensor, and $h_{tel}(0)$ is the 1D image peak (the projected angular spectrum from the telescope). In the read-noise calculations, we assume the use of P pixels to measure the aperture image, with independent read-noise of σ_r^2 in each detector element.

Given the global slope estimate, an average of the local slope measurements, the noise present in each local slope measurement is P times larger

$$\begin{aligned} e_p &= \left\langle (\hat{w}_p - w_p)^2 \right\rangle \\ &= PE_p \\ &= \frac{1}{4N_i h_{tel}(0)^2} + \sigma_r^2 \frac{1}{N_i^2 h_{tel}(0)^2} \end{aligned} \tag{6.17}$$

6.4.2 Duality with the Shack-Hartmann

Using Fourier optics we have seen that the telescope aperture and focal planes behave as dual spaces, where the subdivision operation in one plane results in a reduction in the resolution in the dual plane [17]. The Shack-Hartmann sensor may be seen as a complement to its dual, the pyramid sensor, which operates with the opposing planes in the telescope.

This duality reveals that the dual wavefront sensors are identical in all respects, except for the order of the subdivision and slope measurement operations. The performance limit of the wavefront sensors is closely tied to the subdivision and slope measurement operation, and the optical planes where these operations are performed.

In the Shack-Hartmann sensor, the slope is measured at the focal plane of lenslets which subdivide the aperture plane, so the sensor performance is limited by the size of the lenslets. In the pyramid sensor, the slope measurement is performed at the telescope focal plane, so its measurement precision is limited by the size of the telescope aperture. From comparisons of the correspondence between the wavefront sensors, we expect the performance of the pyramid wavefront sensor to be higher than the performance of the Shack-Hartmann sensor.

6.5 Comparisons of sensor performance

To compare the performance of the pyramid sensor to the Shack-Hartmann sensor, we assume that, relative to the aperture image size, the CCD detector pixel size in the pyramid sensor is matched to the relative size of the lenslets in the Shack-Hartmann sensor. This is done by setting M, the number of lenslets in the Shack-Hartmann sensor, equal to P, the number of pixels per aperture image in the pyramid sensor. For example, Figure 6.2(a) shows 6 lenslets across the telescope aperture in the Shack-Hartmann sensor, simplified to 1 dimension. Correspondingly in Figure 6.2(b), there are 6 imaging pixels across each aperture image in the pyramid sensor. This means that both sensors have the same number of slope measurements, and consequently can be expected to estimate the same number of modes in the turbulence.

To simplify the analysis, we assume a square telescope aperture. In the Shack-Hartmann sensor, we further assume that $h_i(0)$ no longer varies from lenslet to lenslet. Reducing the summation in Equation 6.9, and using $N_{tot} = MN_i$ (uniformly illuminated telescope aperture), the slope variance is

$$E_s = \frac{1}{4N_{tot}h_i(0)^2} + \frac{M\sigma_r^2}{N_{tot}^2 h_i(0)^2} \tag{6.18}$$

It should be noted that Equation 6.18 biases the slope variance marginally in favour of the Shack-Hartmann sensor, since in a circular telescope aperture, partially illuminated lenslets have a lower $h_i(0)$ and consequently contribute higher noise.

To compare the Shack-Hartmann sensor performance to the pyramid sensor, we divide Equation 6.18 by Equation 6.16. The mean sensor errors caused by photon noise, as derived from a single frame of turbulence, is

$$\frac{E_s}{E_p} = \frac{h_{tel}(0)^2}{h_i(0)^2} \tag{6.19}$$

Unlike conventional analyses which do not unify the wavefront sensors [106, 109], the advantage of using a common dual framework for describing the wavefront sensors has allowed us to "cancel" many similarities in two what initially looked very different sensors,

Slope variance[rad^2]	Shack-Hartmann		Pyramid	
	Photon noise	Read noise	Photon noise	Read noise
Per measurement	$\propto \frac{(\frac{\lambda}{d})^2}{N_i}$	$\sigma_r^2 \frac{\left(\frac{\lambda}{d}\right)^2}{N_i^2}$	$\propto \frac{(\frac{\lambda}{D})^2}{N_i}$	$\sigma_r^2 \frac{\left(\frac{\lambda}{D}\right)^2}{N_i^2}$
Averaged global tilt	$\propto \frac{(\frac{\lambda}{d})^2}{N_{tot}}$	$M\sigma_r^2 \frac{\left(\frac{\lambda}{d}\right)^2}{N_{tot}^2}$	$\propto \frac{(\frac{\lambda}{D})^2}{N_{tot}}$	$P\sigma_r^2 \frac{\left(\frac{\lambda}{D}\right)^2}{N_{tot}^2}$
Resolution [m]	d		d	

Table 6.1 Summary of the ideal wavefront sensor performance (photon noise). In the Shack-Hartmann sensor, $D = \sqrt{M}d$, and $N_{tot} = MN_i$, where M is the number of lenslets. The pyramid sensor configuration is matched to the Shack-Hartmann sensor by making $P = M$.

leaving a direct comparison of the differences between the two sensors.

6.5.1 Strehl as performance measure

As an ideal performance benchmark, the results for the diffraction-limited case is summarised in Table 6.1. Under ideal conditions, the performance of the pyramid sensor is $\left(\frac{D}{d}\right)^2$ times better than the Shack-Hartmann sensor[3].

In practice, the performance of wavefront sensors in operational adaptive optics systems is less than perfect. After averaging the "instantaneous" result of Equation 6.19 over time (or the turbulence process), the actual average performance of wavefront sensors can be related to the ideal situation using the Strehl ratio as shown in Equation 6.20.

$$\frac{\langle E_s \rangle}{\langle E_p \rangle} = \frac{\langle h_{tel}(0)^2 \rangle}{\langle h_i(0)^2 \rangle} = \frac{\langle \Gamma_{tel} \rangle^2 h_{tel_0}(0)^2}{\langle \Gamma_i \rangle^2 h_{i_0}(0)^2} \tag{6.20}$$

where h_{tel_0} and h_{i_0} represents the telescope (pyramid sensor) and lenslet (Shack-Hartmann) image peaks under diffraction limited conditions, and Γ the Strehl of their respective long-term exposure in closed loop ($\Gamma h_0(0) = h(0)$). Γ is in fact the 1D analogy of the conventional 2D Strehl ratio. From separate simulations, it is estimated that $\Gamma_{1D} \approx \Gamma_{2D}^k$ in Kolmogorov turbulence, where k is around 0.6 to 0.7.

The Strehl of the long-term exposure image provides the open loop [89] performance of

[3]The Shack-Hartmann sensor could be configured so that there is only one lenslet, $n = 1$, across the aperture (allowing only the global mean slope is to be estimated). Such a configuration results in equal performance between the two sensors.

the sensors. With tip/tilt correction, the sensor performance is given by the Strehl of the short-term exposure image. We are interested in the performance of the wavefront sensors when the higher order wavefronts are corrected, when the compensated image takes on a characteristic core and halo structure [85].

Under low turbulence levels (small $\frac{D}{r_0}$), the performance of the Shack-Hartmann sensor (the Strehl ratio of each lenslet image) does not change significantly. In contrast, the pyramid sensor image resolution is roughly equivalent to that from a telescope of diameter r_0, so the Strehl ratio and performance of the pyramid sensor image drops significantly. Under a closed-loop system, we expect the performance of both wavefront sensors to improve again. The pyramid sensor should now show a higher level of improvement in its performance.

6.6 Simulation of operating conditions

Kolmogorov phase-screens [42] are used to simulate the effects of atmospheric turbulence, and estimated using their Zernike modes. By keeping the number of Zernike modes under consideration low (20 modes) and using 64x64 pixels for the aperture size, discretisation errors are kept low, and are insignificant. The closed-loop wavefront is approximated by completely cancelling the 8 lowest modes in the wavefront.

In the Shack-Hartmann sensor, the complex field at the aperture is divided into 8x8 lenslets and propagated using a Discrete Fourier Transform onto quad-cells. The pyramid sensor divides the complex field at the focal plane into 2x2 quadrants, and re-images the aperture onto 8x8 pixels. This is equivalent to $M = P = 8^2$ in Equation 6.18 and Equation 6.16. Poisson noise with mean photon count of 800 is then added to the measured images. The final wavefront errors due to Poisson noise are normalised to be equivalent to a mean photon count of 1.

6.6.1 Photon noise

In the first simulation, the performance of the wavefront sensors is determined by the variance in their wavefront slope estimates only. This is measured from the difference between the slope estimates in the absence and presence of photon (Poisson) noise. The absolute slope errors (difference between estimated and true slopes) contain additional errors due to the non-linearity of quad-cells, and are considered separately.

We first confirm the accuracy of Equation 6.9 and Equation 6.16 by comparing them to the

simulated slope errors. The simulated sensor performance under closed-loop conditions, with the lowest 8 modes being fully compensated, is shown in Figure 6.4 as $\frac{D}{r_0}$ is varied from 0 to 25. In both sensors, the simulated and predicted slope errors are in close agreement for low turbulence levels of up to $\frac{D}{r_0} = 10$. This confirms the accuracy of the predictions given by Equation 6.9 and Equation 6.16 using the Strehl ratio. As expected, the performance of the pyramid sensor surpasses the performance of the Shack-Hartmann sensor. At low turbulence levels, the sensor performance approaches the ideal performance, with the error in the pyramid sensor being $\frac{D}{d} = 8$ (square-root of the quantities in Table 6.1) times lower than the Shack-Hartmann error.

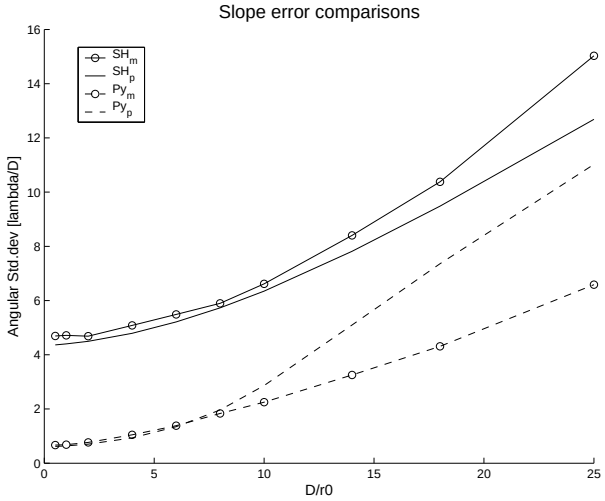


Figure 6.4 Comparison of the simulated and predicted slope errors due to photon noise as $\frac{D}{r_0}$ is varied, with closed-loop compensation in place. The curves represent the Shack-Hartmann sensor slope error measured directly in the simulations (SH_m) compared to the predicted slope error (SH_p, Equation 6.9), and the equivalent pyramid sensor slope error measured in the simulations (Py_m) compared to the predicted slope error (Py_p, Equation 6.16).

In Figure 6.5, we compare the open loop performance against the closed loop performance of both wavefront sensors, and confirm the improvement in closed loop. The performance of both sensors improve in closed loop because the long-term exposure images in both sensors now have higher Strehl ratios. In contrast to the Shack-Hartmann sensor, where the blurring of the long-term exposure image is dominated by random image displacements within each lenslet, the pyramid sensor image is blurred by the wavefront across the whole aperture, and consequently, improves much more in closed loop, particularly at higher turbulence ($\frac{D}{r_0}$) levels.

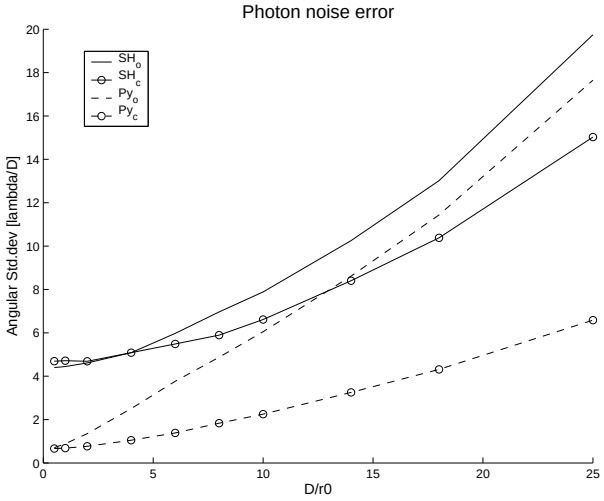


Figure 6.5 Simulations of the performance of the wavefront sensors with photon noise only. The curves represent the Shack-Hartmann sensor slope error in open loop (SH_o) and in closed loop (SH_c), along with the pyramid sensor slope error in open loop (Py_o) and in closed loop (Py_c).

6.6.2 Noise from non-linear errors

Although the sensor performance under photon noise as shown here clearly favours the pyramid sensor, in fact, the non-linearity of quad-cells also lead to errors in the slope estimate. The combined errors from photon noise and non-linearity in the wavefront sensors are shown in Figure 6.6. The estimation errors are now larger compared to Figure 6.5, with the performance in open loop of the pyramid sensor now being comparable to the Shack-Hartmann.

Under closed-loop operating conditions, the non-linear error in the both sensors is reduced to produce a better estimate of the wavefront. However, in the Shack-Hartmann, there is no increase in sensitivity of the measurement as a whole, since the size of the speckle image under each lenslet remains unchanged.

6.7 Conclusion

The Shack-Hartmann sensor subdivides the telescope aperture and measures the local slope within each subaperture using a quad-cell. The resolution of the wavefront estimate is inversely proportional to the size of subapertures, while the precision of the measurements is determined by the image height, which is roughly proportional to the size of the subapertures.

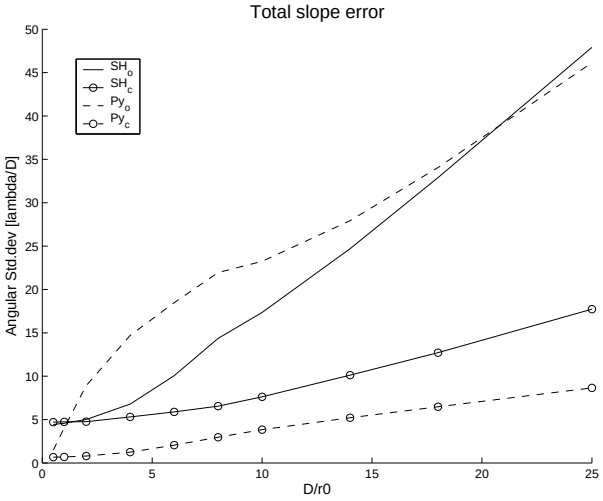


Figure 6.6 Simulations of the full performance of the sensors taking all other errors into account. The curves represent the Shack-Hartmann sensor in open (SH_o) and closed loop (SH_c) along with the pyramid sensor in open (Py_o) and closed loop (Py_c). In both cases, closed-loop operation (circled lines) show an improvement over open loop operation (uncircled lines).

Compensation of the wavefront results in a reduction in the mean slope across each lenslet, with no significant corresponding reduction in the lenslet spot size (sensitivity). On average, there is now a smaller signal, without a corresponding offset in increased sensitivity, which is limited by the lenslet size.

In the pyramid sensor, the wavefront slope is estimated by comparing the intensity changes in each facet of the pyramidal prism. The precision of the wavefront slope estimate is determined by the image height at the focal plane, which is in turn determined by the size of the telescope aperture.

The wavefront resolution of the pyramid sensor is given by the CCD sampling at the aperture image plane. Each detector element in the CCD array provides a measurement of the slope within the equivalent region bounded by the detector. More wavefront slope measurements can be obtained by increasing the sampling density of the CCD detector elements. This can be achieved by reducing the physical size of the CCD detectors[4], or equivalently, by optically magnifying the aperture image before sampling. In contrast, the Shack-Hartmann configuration cannot be re-sized dynamically. Thus freed from physical limitations to the subdivision size, the wavefront resolution in the pyramid sensor is only limited by blurring in the aperture images.

[4] Alternatively, the CCD pixel size may often be increased using the built-in on-chip binning function.

It is important to note that the resolution-precision constraints examined in Section 6.3.2 does not apply identically to pyramid sensor. The precision of the global wavefront slope is not constrained by the aperture image subdivision operation which occurs *after* the slope has been measured at the focal plane. Unlike the Shack-Hartmann sensor, the trade-off between resolution and precision is limited only by the size of the telescope aperture, not by the size of the aperture subdivisions.

In this chapter, we used the duality between the Shack-Hartmann sensor and the pyramid wavefront sensor to compare their performance, and have shown that the pyramid sensor is fundamentally better. We have shown, through simulations, that in practice, the pyramid sensor can provide significant advantages over the Shack-Hartmann sensor in closed-loop wavefront compensation systems. In open loop conditions, the performance of the pyramid sensor is roughly similar to the Shack-Hartmann sensor.

In our comparisons, we suggested the use of the Strehl ratio (defined on 1D-images), as opposed to the sometimes ambiguous image width, as a more precise and convenient measure of the sensitivity of the wavefront sensors, particularly in closed loop operation. The degradation in sensitivity of the sensors is thus characterised by the Strehl ratio of the adaptive optics system.

Chapter 7

Wavefront sensing from defocused images

This chapter examines the curvature sensor and the geometric wavefront sensors. In contrast to the explicit aperture subdivision process in the Shack-Hartmann and pyramid wavefront sensor, these sensors implicitly subdivide the telescope aperture. Under geometric optics, the propagation of light through a medium results in intensity fluctuations related to the wavefront. The changes in intensity can be used in the wavefront sensors to recover the wavefront aberrations.

Figure 7.1 shows the propagation of a 1D aberrated wavefront from plane A to plane B. As an intuitive analogy, wavefront aberrations are water ripples in a bathtub illuminated from the top. Ripples on the water surface change the direction of the light rays travelling downwards, resulting in corresponding light and dark fringes at the bottom of the bathtub. The direction and change in intensity as light propagates are described by Equation 3.9 and Equation 3.12.

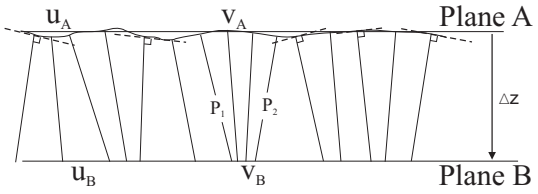


Figure 7.1 The effect of wavefront perturbations on the direction of light rays.

The actual wavefront sensing arrangement is shown in Figure 7.2, where the complex field at the telescope aperture is allowed to propagate, but not all the way to the focal plane. Instead, at two opposing out-of-focus planes, the defocused outline of the telescope aperture is imaged, and subdivided into local intensity measurements.

In Figure 7.2, a small aberrated wavefront section has been shown highlighted. The small positive wavefront curvature error causes the light rays within that region to be spread out, so they now focus at a point after the original prime focal plane. The corresponding changes in the out-of-focus intensity measurements allow this wavefront change to be measured and localised[1].

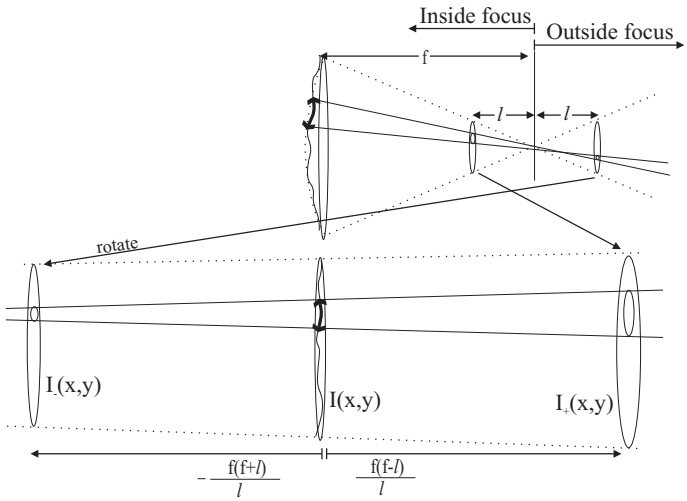


Figure 7.2 The physical layout (top) of a geometric wavefront sensor, with an optically equivalent arrangement, for ease of analysis, shown (bottom). This is equivalent to a wavefront (windowed by the aperture alone) propagating in free space. Note that the equivalent outside-focus image is rotated.

The intensity changes at the out-of-focus planes can be described by geometric optics. To simplify the analysis, Figure 7.2(bottom) also shows an equivalent optical arrangement [88] for wavefront sensing, where the focusing mirror or lens in the telescope is replaced by an equivalent free space propagation.

From Equation 3.30, a telescope with focal length f introduces a quadratic phase term $e^{-i\frac{k}{2f}(x^2+y^2)}$. Given a complex field $A(x,y)e^{i\phi(x,y)}$ at the telescope aperture, the complex field after propagating a distance of z is

[1]Here, the intensities at the out-of-focus measurement planes (at $f \pm l$) change in opposing ways - inside focus, it is dimmer, while outside focus, it is brighter.

$$
\begin{aligned}
& A(x,y)e^{i\phi(x,y)}e^{-i\frac{k}{2f}(x^2+y^2)} \odot e^{i\frac{k}{2z}(x^2+y^2)} \\
= & \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} A(x',y')e^{i\phi(x',y')}e^{-i\frac{k}{2f}(x'^2+y'^2)}e^{i\frac{k}{2z}((x-x')^2+(y-y')^2)}\,dx'dy' \\
= & e^{i\frac{k}{2f}\frac{z'}{z}x^2}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} A(x',y')e^{i\phi(x',y')}e^{i\frac{k}{2z'}((\frac{z'}{z}x-x')^2+(\frac{z'}{z}y-y')^2)}\,dx'dy' \\
= & e^{i\frac{k}{2f}\frac{z'}{z}x^2}\left(A(x,y)e^{i\phi(x,y)} \odot e^{i\frac{k}{2z'}(x^2+y^2)}\right)\Big|_{(\frac{z'}{z}x,\frac{z'}{z}y)}
\end{aligned}
\tag{7.1}
$$

where $\frac{1}{z'} = \frac{1}{z} - \frac{1}{f}$ or $z' = z\frac{f}{f-z}$ is the equivalent propagation distance without the quadratic phase term. This result is the same as the geometric optics based thin-lens equation of Equation 3.5.

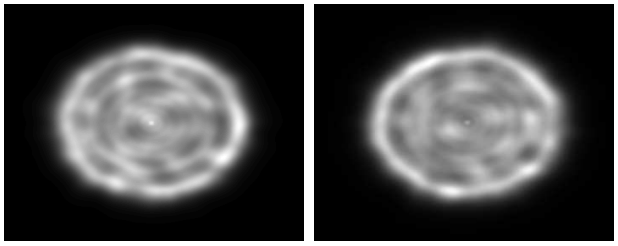
Under the equivalent optical arrangement, the image at the inside-focus plane $z = f - l$ is identical to (but smaller than) the image at $\frac{(f-l)f}{l}$ without the quadratic phase term. The outside focus, at $z = f + l$, is similarly equivalent to a virtual propagation distance of $-\frac{(f+l)f}{l}$, with an additional image inversion or rotation about the axis of propagation. The defocus l is usually small enough[2] that the equivalent virtual propagation distances are approximately $\pm\frac{f^2}{l}$.

The inputs to the wavefront sensors come from measurements of the out-of-focus images. A simulation of the propagated and defocused aperture images, with some turbulence, is shown in Figure 7.3. By design, l is adjusted so that the imaging plane is placed far enough from the focal plane to minimise the effects of diffraction on the defocused images. The diffraction effects are small enough that they are smoothed out by the image blurring and sampling operation carried out by CCD detectors, and are not visible in the sampled image. The blurred outlines of the telescope aperture remain visible, allowing the effect of any wavefront aberrations on the images to be described using geometric optics alone.

The displacement l trades off the sensitivity of the wavefront sensor against its resolution. As quantified in Equation 7.6, the intensity fluctuations in the defocused images are roughly proportional to l and I (the mean intensity). The resolution, corresponding roughly to the size of the dark and bright patches in the images, is determined by diffraction effects, and is inversely proportional to $\sqrt{l}$, as shown later in Section 7.4.3. With smaller l, or larger equivalent propagation distances z', the sensitivity is increased, at the cost of a lower

[2] As an example, on a 1m F/10 telescope (focal length 10m), a defocus of $l = 2cm$ (and corresponding image size of 2mm) is equivalent to a virtual propagation distance of $z \approx \frac{f^2}{l} = 5km$.

resolution [91].



(a) Inside-focus image from the wavefront sensor, $i_+(x,y)$.

(b) Outside-focus image from the wavefront sensor, $i_-(x,y)$.

Figure 7.3 Defocused images from two opposing planes.

In the following simulations, photon noise is simulated using a Poisson model, while read noise is ignored[3].

$$P(I(x,y)|i(x,y)) = \prod_{\forall(x,y)} \frac{e^{-i(x,y)} i(x,y)^{I(x,y)}}{I(x,y)!} \tag{7.2}$$

where $i(x,y)$ and $I(x,y)$ are the intensity measurements before and after the addition of noise, respectively.

The input measurements to the wavefront sensors are thus

$$I_+(x,y) = i_+(x,y) + n_+(x,y) \tag{7.3}$$

and

$$I_-(x,y) = i_-(x,y) + n_-(x,y) \tag{7.4}$$

where $i_{+/-}(x,y)$ and $I_{+/-}(x,y)$, represents the intensities before and after the addition of noise $n_{+/-}(x,y)$, respectively.

[3]Typical astronomical observations operate under low light levels, and with cooled equipment, to give low instrument noise. Avalanche photo diodes (which have no read noise) have also been used [85, 88] for curvature sensing. Since read noise is only an instrumentation limitation, it is ignored in subsequent analyses.

Figure 7.4 shows the large visible effects of photon noise, from a mean total photon count of 40000, on an image. Most simulations in this section assume even higher noise levels, with photon counts of 800, so a method for accumulating and averaging the signal is required. The fluctuations in the measured intensity distribution can be reduced by software averaging, or by adjusting the size of the CCD detector elements. The increased integration area of each detector element results in fewer detectors (fewer measurements) and lower read noise, but also a correspondingly reduced image spatial resolution.

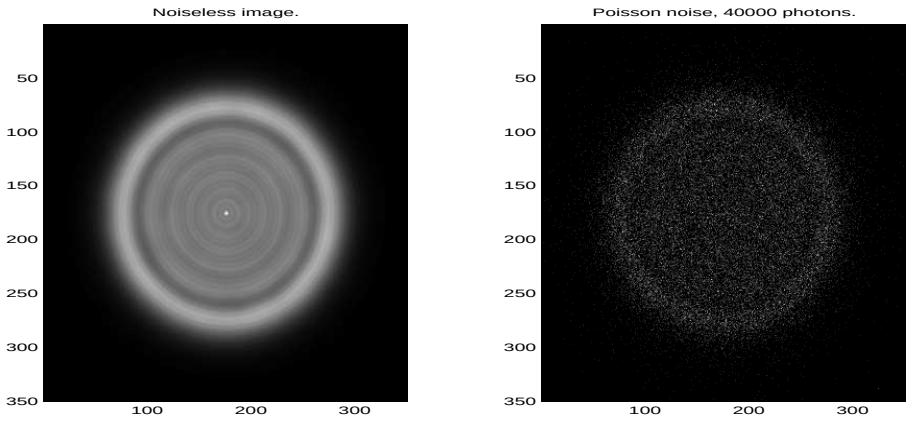


Figure 7.4 Effect of photon (Poisson) noise on input image with total flux of 30000 photons. The telescope aperture is 250 pixels in diameter, equivalent to 1m. The input image has been propagated 14km, assuming a light wavelength of 600nm. The wavefront aberrations are small enough $\frac{D}{r_0} = 0.1$ that no intensity fluctuations could be observed.

Section 7.1 and Section 7.2 examine the geometric optics formulation for recovering the wavefront. Section 7.3 then introduces the curvature sensor approximation. Finally, section 7.4 investigates the effects of photon noise on each wavefront sensor.

7.1 Geometric optics solution

In this section, we examine the free space propagation of wavefronts using the geometric optics model, and derive a solution for recovering the wavefront from its effects on light intensity, expanding on its original introduction by van Dam and Lane. [102]. Referring to Figure 7.1, the direction of travel of the light rays at a particular section of an aberration wavefront is perpendicular to the wavefront slope (slope of the water surface in the bathtub analogy) at that point. Mathematically, referring to Equation 3.9, the initial and final positions of a light ray are related to the wavefront slope by

$$x_B = x_A + \Delta z W_x(x_A) \tag{7.5}$$

where x_A and x_B are ray-intercepts of the ray with planes A and B respectively.

The irregular wavefront causes the propagating light rays to spread out and concentrate unevenly. The intensity at any point is proportional to the density of light rays passing through that point. For example, in Figure 7.1, the concentration of light rays around point v causes a relative brightening on the intensity at v_B compared to v_A, while the diffusion of light rays at point u causes a corresponding relative darkening of the intensity at u_B. In 1D, Equation 3.12 reduces to

$$I_B(x_B) = \frac{I_A(x_A)}{1 + \Delta z H(x_A) + \Delta z^2 K(x_A)} = \frac{I_A(x_A)}{1 + \Delta z H(x_A)} \tag{7.6}$$

where $I_A(x)$ and $I_B(x)$ represent the intensity distributions in the planes A and B respectively. In 1D, the mean wavefront curvature at plane A is $H(x) = W_{xx}(x)$, while the Gaussian curvature at plane A is $K(x) = 0$.

Figure 7.5 illustrates this for a wavefront at the originating plane A with a uniform negative curvature, $W(x) = -ax^2$, for $a > 0$. The illumination at A is assumed constant ($I_A(x) = I_A$) within a window representing a finite optical aperture. The wavefront slope at plane A, $W_x(x) = -2ax$, determines the direction in which the light rays leave plane A.

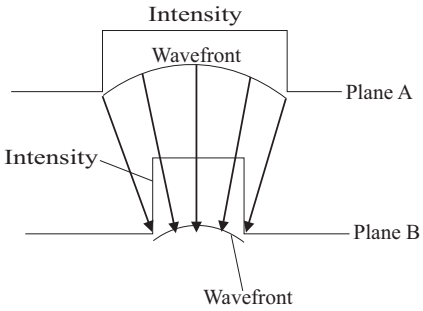


Figure 7.5 A simple defocus in the wavefront causes the image of the aperture to be smaller but brighter. All rays move at 90° from the wavefront slope.

$$x_B = x_A - 2ax_A\Delta z \tag{7.7}$$

The uniform parabolic wavefront gives rise to a uniform focusing action, which causes the intensity distribution at plane B to be the same shape as the intensity distribution at plane

A, but smaller and brighter.

$$I_B(x) = I_B = \frac{I_A}{1 - 2a\Delta z} \tag{7.8}$$

From the intensity distribution alone, we can recover the original wavefront aberrations using the positions of light rays. Figure 7.6 shows Figure 7.5 with some light ray positions inferred. The leftmost ray P_1 defines the edges of the aperture. P_2 is then reconstructed by making use of the fact that the total intensity between the two rays P_1 and P_2 is constant.

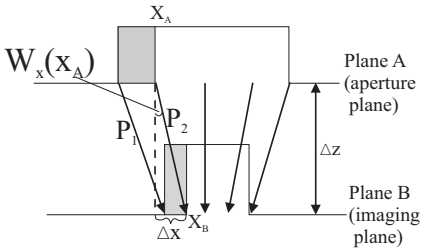


Figure 7.6 The shaded regions in each plane are equal in area (intensity), so the starting and ending points of the light ray P_2 lie along the boundary of the shaded regions. Thus given the direction of the light ray, the corresponding wavefront slope at plane A, $W_x(x_A)$, can be found.

Assuming that the light rays in the region from plane A to B never cross over each other (as when the wavefront distortions and propagation distances are small), the positions of P_1 and any ray P_2 can be recovered unambiguously[4]. Propagating from plane A to plane B, the intensity distribution is stretched and compressed but not lost by the changing light ray positions. This intuitive notion of the principle of the conservation of light can be expressed mathematically as

[4]If any light rays cross, it is no longer possible to unambiguously recover the positions and directions of the light rays. In regions where light rays intersect, also known as caustics, diffraction effects are especially prominent [94], and the geometric optics approximation breaks down. For example, in the extreme, the geometric optics model breaks down at the focal plane (where all light rays meet), and the intensity distribution has to be described using scalar diffraction theory instead.

$$\begin{aligned} C_{I_B}(x_B) &= \int_{-\infty}^{x_B} I_B(x')dx',\ x' = x + \Delta z W_x(x) \\ &= \int_{-\infty}^{x_B} \frac{I_A(x)}{1+\Delta z W_{xx}(x)} dx' \\ &= \int_{-\infty}^{x_A} \frac{I_A(x)}{1+\Delta z W_{xx}(x)} (1+\Delta z W_{xx}(x)) dx \\ &= \int_{-\infty}^{x_A} I_A(x)dx \\ &= C_{I_A}(x_A) \end{aligned} \tag{7.9}$$

where $I_A(x)$ and $I_B(x)$ are the intensity distributions across planes A and B respectively. The wavefront slope $W_x(x_A)$ is $\frac{x_B - x_A}{\Delta z} = \frac{\Delta x}{\Delta z}$.

The cumulative intensity matching process is also known as histogram specification [14, 37]. From this process, the displacement of each light ray is found. The wavefront slopes, in plane A at the base of each ray, are found from Equation 7.5. In the example given above, the original wavefront is $W(x_A) = -ax_A^2$). The cumulative intensity distributions are

$$\begin{aligned} C_{I_B}(x_B) &= I_B x_B + \tfrac{N}{2} \\ C_{I_A}(x_A) &= I_A x_A + \tfrac{N}{2} \end{aligned} \tag{7.10}$$

where $N = \int_{-\infty}^{\infty} I_A(x)\,dx = \int_{-\infty}^{\infty} I_B(x)\,dx$ is the total intensity.

By matching the ray positions using histogram specification, as shown in Figure 7.7,

$$\begin{aligned} C_{I_B}(x_B) &= C_{I_A}(x_A) \\ I_B x_B &= I_A x_A \end{aligned} \tag{7.11}$$

we recover the slope of the original wavefront. Consequently, the wavefront can be calculated exactly (to within the geometric optics approximation).

$$\begin{aligned} W_x(x_A)\Delta z &= \Delta x \\ W_x(x_A) &= \frac{x_B - x_A}{\Delta z} \\ &= \frac{x_A\left(\frac{I_A}{I_B} - 1\right)}{\Delta z} \\ &= -2ax_A \end{aligned} \tag{7.12}$$

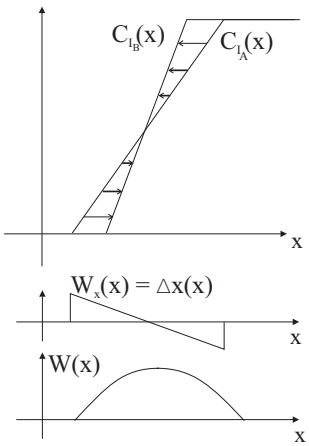


Figure 7.7 Solution to the wavefront slope using histogram specification. The ray positions are found by matching equal levels in the histograms. From the ray positions, the original wavefront slope and finally the wavefront itself, is recovered.

7.1.1 Minimising diffraction effects

To confirm the validity of the geometric optics model, and to show how diffraction effects can be ignored, the Fresnel diffraction formula is used. This allows simulations of free space propagation with full diffraction effects to be performed.

As an example of the effects diffraction can have during propagation, a random wavefront aberration at a square telescope aperture of length 1m is propagated to several distances ranging from $\pm$30km to $\pm$120km. Although the optical Fresnel propagation model is in 2D, the geometry of the problem is reduced to one dimension by letting the complex field at the aperture vary across 1 axis only. Figure 7.8 shows the intensity distribution at 30km, the closest propagation distance simulated.

The intensity distribution at this distance no longer has any sharp edges because of the

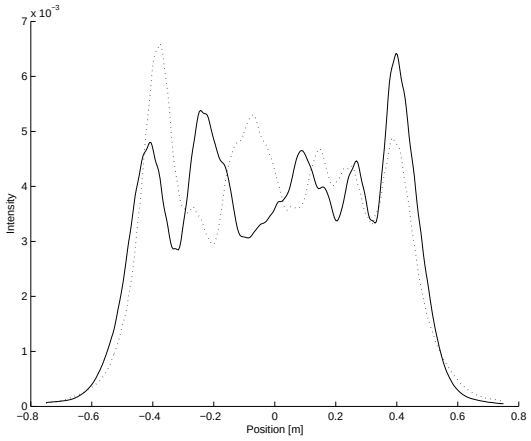


Figure 7.8 Geometric wavefront sensing with 1D images of an aperture 1m in diameter, propagated 30km in front of (solid line), and behind (dotted line), the aperture.

smoothing caused by diffraction[5]. The effects of diffraction are stronger at longer distances. By keeping the propagation distance suitably close, the geometric optics solution is kept accurate. At 30km, the effects of diffraction are still minimal, and the original outline of the aperture can still be seen. The intensity fluctuations in the propagated image form the input to the wavefront reconstruction process. The sensitivity of the sensor is proportional to the distance of propagation.

Using histogram specification, the wavefront is estimated and compared with the actual wavefront in Figure 7.9. At 30km, the wavefront estimate is a good approximation of the original simulated wavefront function. The propagation process has blurred the aperture image in a low-pass filtering operation. This causes the estimated wavefront to be smoother and lower in spatial resolution than the original wavefront. Over larger distances, the blurring increases, so our wavefront estimate becomes smoother and less accurate.

Short propagation distances ensure that the histogram specification process is accurate enough to obtain an accurate estimate of the wavefront at the telescope aperture. At long distances, when diffraction effects dominate, the relationship to the wavefront is non-linear, and falls into the class of phase retrieval problems. The presence of two images (previously shown to be optically equivalent to slightly defocused planes) correspond to two phase diverse measurements, and is commonly known as the phase diversity (with defocus) problem.

It is assumed in the following discussions that the imaging planes are sufficiently defocused

[5]This smoothing size is roughly on the order of $\sqrt{\lambda z}$, known as the Fresnel length [38]. The Fresnel blurring determines the resolution of the wavefront estimate.

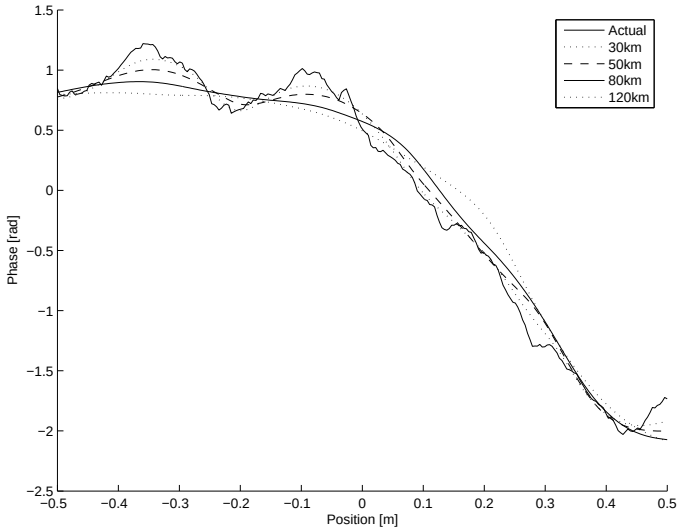


Figure 7.9 Comparison of the actual phase at the imaging aperture (solid, jagged line) with the phase estimate after propagating through various distances. The larger the propagation distance, the smoother the wavefront estimate, and the more the deviation from the actual wavefront.

from the focal plane that diffraction has minimal effects on our results.

7.2 Geometric wavefront sensor

In this section, the geometric wavefront sensor is generalised to estimate two dimensional wavefronts [105]. The geometric wavefront sensor is a slope based sensor. In two dimensions, light rays continue to travel perpendicular to wavefront slopes, and intensity, now determined by the density of light rays within an area, is still conserved. However, the endpoints of any light ray can no longer be inferred directly by ray tracing or histogram specification. For example, the left-edge of the 1D aperture in Figure 7.6 defines the two points in each plane A and B, corresponding to the initial and final positions of the leftmost light ray. However, in 2D, the outer edges of the aperture are now defined not by points, but by curves. The location and direction of light rays, now with an extra degree of freedom, can no longer be recovered.

In Section 7.1.1, a two dimensional wavefront was recovered by treating it as a one dimensional wavefront, since the wavefront function is constant in one axis. This provides a clue as to how a two dimensional wavefront can be estimated using geometric optics. The images are reduced by a series of projections to a number of one dimensional image slices, similar to the radon transform used in medical CT applications. Each slice, consisting of the

integrated intensity along an axis, as shown in Figure 7.10 for a single projection direction, is related to the wavefront projection along the same axis.

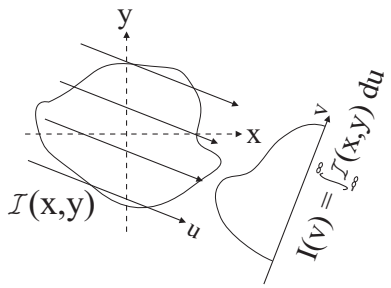


Figure 7.10 A single projection in the radon transform for 2D wavefront reconstruction.

To derive the relationship between the projections of images and the projections of wavefront functions, the ray-tracing histogram specification process is performed on the projected intensity distribution. For example, taking the y-axis as the projection direction, the 2D version of the histogram specification problem, the ray tracing process of Equation 7.9, is equivalent to

$$\int_{-\infty}^{x_A}\int_{-\infty}^{\infty} I_A(x,y)\, dy\, dx = \int_{-\infty}^{x_B}\int_{-\infty}^{\infty} I_B(x,y)\, dy\, dx \tag{7.13}$$

with $I_A(x,y)$ and $I_B(x,y)$ being the 2D intensity distributions at planes A and B.

The wavefront slope at the aperture is approximately

$$\frac{\int_{-\infty}^{\infty} I_A(x,y)W_x(x,y)dy}{\int_{-\infty}^{\infty} I_A(x,y)dy} = \frac{\Delta x(x)}{\Delta z} = \frac{x_B - x_A}{\Delta z} \tag{7.14}$$

To see how Equation 7.14 works, consider the example of a constant wavefront slope and intensity across the aperture. The intensity distributions across the two out-of-focus planes, $I_A(x,y)$ and $I_B(x,y)$, have the exact same shape except for a displacement. When projected into 1D, through histogram specification, the constant displacement (Δx) across the aperture can be found, allowing the magnitude of the slope in the wavefront to be recovered. Similarly, extending this to higher orders of aberrations requires the "slice displacements", Δx, to be measured from more projections over different directions[6].

[6]The number of projection angles used determines the resolution and fitting errors in the wavefront estimate. For example, simulations in this thesis use up to 10 projection angles to recover 20 Zernike modes.

In practice, using the linearity of the problem, we may associate each Zernike mode directly to their effects on the image intensity. Given a decomposition of some wavefront $W(x,y)$ into its Zernike coefficients, with each coefficient given by $\alpha_i = \frac{1}{\pi}\int W(x,y)Z_i(x,y)dx\,dy$, Equation 7.14 is linear function of the coefficients $\boldsymbol{\alpha}$.

$$\boldsymbol{d} = \boldsymbol{H}\boldsymbol{\alpha} \tag{7.15}$$

where $\boldsymbol{d}$ is the signal vector formed from the displacements $\Delta x(x)$[7], as found through histogram specification.

Given the signals obtained from histogram specification[8], Equation 7.15 can be inverted to recover the wavefront function. Although the Maximum A Posteriori solution is theoretically the most optimal, in practice, at high photon counts (low photon noise levels), a least-squares solution (equivalent to the Maximum-Likelihood solution, with uniform white noise assumptions) is found to be adequate.

$$\boldsymbol{\alpha} = (\boldsymbol{H}^T\boldsymbol{H})^+\boldsymbol{H}^T\boldsymbol{d} \tag{7.16}$$

The solution to the geometric wavefront sensor is a system of linear equations. The sensor output, although derived using a non-linear algorithm, can be linearly related to the input wavefront coefficients. Additionally, prior information on the wavefront coefficients can also be included in Equation 7.16, resulting in an MAP solution. Geometric optics represent a practical wavefront sensing solution that is physically simpler than the Shack-Hartmann and Pyramid wavefront sensors.

7.3 Curvature sensor

Due to the novelty of the method, the algorithm for geometric wavefront sensing has not been applied on working adaptive optics systems. Today, the algorithm used in wavefront estimation with defocused images is largely based on curvature estimation, first proposed by Roddier in 1988 [13, 83, 86].

[7] The vectors of displacement signals in each projection direction are combined by stacking the vectors together to form a single vector.

[8] Note that although histogram specification is a non-linear process, the remaining parts of the geometric wavefront sensor is linear.

The curvature sensor was initially proposed as a simple and effective method for low-order adaptive optics in infra-red applications. Requiring only two defocused image measurements, the physical simplicity of the curvature sensor has lead to its widespread use. The initial design for the sensor provides for curvature signals sent to adaptive optics systems with membrane or bimorph mirrors as wavefront correctors[9]. This is particularly convenient as the bimorph mirrors respond to a curvature signal because of their mechanical properties. In practice, instrumental limitations necessitate the use of more complex designs to match the signal between the wavefront sensor and the mirror actuators [85].

The curvature sensor uses the same defocused image data used by the geometric wavefront sensor. However, the curvature sensor makes some simplifying assumptions, resulting in the estimation of wavefront curvature instead of slopes. Equation 3.12 is reproduced in Equation 7.17 with the approximation $(x_B, y_B) = (x_A, y_A)$, essentially ignoring any displacements in the local intensity signals during image propagation. Furthermore, the wavefront shape is implicitly assumed to be locally spherical, so the eccentricity or Gaussian curvature $K(x,y) = 0$.

$$\begin{aligned} I_B(x_B,y_B) &= \frac{I_A(x_A,y_A)}{1+\Delta z H(x_A,y_A)+\Delta z^2 K(x_A,y_A)} \\ I_B(x,y) &\approx \frac{I_A(x,y)}{1+\Delta z H(x,y)} \\ &\approx I_A(x,y) - I_A(x,y)\Delta z H(x,y) \end{aligned} \tag{7.17}$$

The approximate intensity difference from propagating a wavefront a distance Δz is then

$$\Delta I(x,y) = I_B(x,y) - I_A(x,y) = -\Delta z I_A(x,y) H(x,y) \tag{7.18}$$

With two images defocused in opposing directions, symmetrically displaced about the focal point of the telescope, the intensity in each plane provides a differential signal[10] approximating the wavefront curvature. Where a region is brighter in one image, it is darker in the other. The curvature sensor signal, formed from the difference between these two out-of-

[9] This is essentially a zonal correction scheme, since the curvatures are computed within separate hexagonal regions, and corrected by mirrors driven by signals proportional to the local curvature.

[10] The differential signal allows scintillation or intensity fluctuations in the telescope aperture to be cancelled.

focus images, is

$$\begin{aligned} S(x,y) &= I_+(x,y) - I_-(x,y) \\ &\approx -2\Delta z I(x,y) H(x,y) \end{aligned} \tag{7.19}$$

where $I_+(x,y)$ and $I_-(x,y)$ represent the two defocused images measured by the curvature sensor[11]. $I(x,y)$ and $H(x,y)$ are the intensity and the wavefront curvature at the aperture plane.

The wavefront curvature is thus given by

$$\begin{aligned} H(x,y) &\approx -\frac{S(x,y)}{2I(x,y)\Delta z} \\ &\approx -\frac{S(x,y)}{(I_+(x,y) + I_-(x,y))\Delta z} \end{aligned} \tag{7.20}$$

The 1D example in Figure 7.5 is useful to illustrate the curvature sensing algorithm. The same wavefront is recovered by integrating the curvature signal twice, as shown in Figure 7.11.

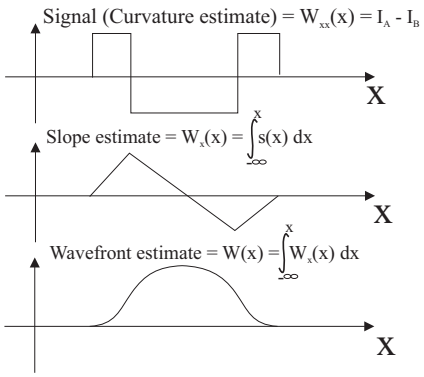


Figure 7.11 Estimation of wavefront from Figure 7.5 with the sensor signal $s(x)$ on top, and the recovered wavefront (with edge effect errors) at the bottom.

The wavefront estimate near the edges of the telescope aperture is no longer accurate. By propagating a flat wavefront with an overall tilt, Figure 7.12 reveals the presence of edge

[11] Recall from Figure 7.2 that the defocused images are equivalent to free space propagation, but $I_-(x,y)$ needs to be rotated 180 degrees.

effects [88] in the differential image signal. Since the wavefront has no curvature, a clear boundary between the edge signal and the zero curvature region can be seen.

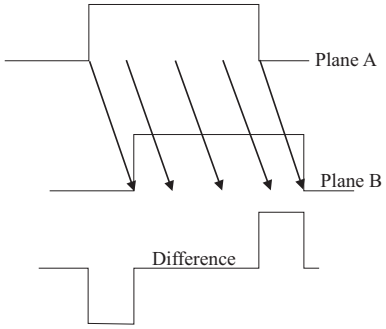


Figure 7.12 A wavefront that is only tilted has zero curvature and produces no curvature signal. An edge signal is still produced.

By ignoring displacements in the signal due to the wavefront slope, the curvature sensor has introduced estimation errors in the curvature signal. More significantly, as shown by the example in Figure 7.11 (compare Figure 7.7), an additional source of error in the edge signal is also present.

Arising from image subtraction over mis-matched aperture edges, the edge signal is proportional to the radial wavefront slope at the edges of the telescope aperture. Practical curvature sensors must therefore model the edge signal separately from the central curvature region [11, 31, 41]. The output from a curvature sensor thus has two components, a curvature signal, and an edge signal. In general, the exact extent of the edge signal cannot be determined, so the boundary to separate the two types of signals remains ambiguous.

7.3.1 Error approximation estimation

The error in the curvature sensor approximation, compared to the geometric wavefront sensor, is given by (continuing from Equation 7.6)

$$\begin{aligned} \Delta I_B(x_B,y_B) &= I_{B_{geo}}(x_B,y_B) - I_{B_{curv}}(x_B,y_B) \\ &= \frac{I_A(x_A,y_A)}{1+\Delta z H(x_A,y_A)+\Delta z^2 K(x_A,y_A)} - \frac{I_A(x_B,y_B)}{1+\Delta z H(x_B,y_B)} \end{aligned} \tag{7.21}$$

Due to the division operation, the error is a non-linear function of distance and wavefront curvature. Linear approximations [49, 61] to Equation 7.21, up to the first order, has been

derived from the equivalent Intensity Transport Equation (Equation 3.36) representation.

The error in the curvature sensor, extended to the second order by van Dam and Lane [103], is (all functions are evaluated at (x_B, y_B))

$$I(z+\Delta z) = \frac{I(z)}{1+H\Delta z+(K-T)\Delta z^2} \tag{7.22}$$

where $T = W_x W_{xxx} + W_x W_{xyy} + W_y W_{xxy} + W_y W_{yyy}$ = $W_x H_x + W_y H_y$ is the displacement error of H. The Laplacian curvature H represents a first order change in the intensity, while K and T are both second order errors.

However, even a second order error approximation is insufficient for extended analyses of the curvature sensor. The in-focus and outside-focus image planes, given by a Taylor series expansion about $I(z)$, are

$$\begin{aligned} I(z+\Delta z) - I(z) &= \sum_n \left(\frac{\partial^n}{\partial z^n} I(z) \right) \frac{(\Delta z)^n}{n!} \\ I(z-\Delta z) - I(z) &= \sum_n \left(\frac{\partial^n}{\partial z^n} I(z) \right) \frac{(-\Delta z)^n}{n!} \end{aligned} \tag{7.23}$$

The curvature sensor signal is the differential signal between the out-of-focus planes. The second order terms in the sensor signal cancel,

$$I(z+\Delta z) - I(z-\Delta z) = 2\sum_n \left(\frac{\partial^n}{\partial z^n} I(z) \right) \frac{\Delta z^n}{n!}, \ \forall\ n \text{ odd} \tag{7.24}$$

The third order error term thus needs to be retained for further analysis of the curvature sensor. As an example, the first few terms in the Taylor series expansion of Equation 7.22 are

$$\begin{aligned} I(z+\Delta z)-I(z) &= I_z\Delta z + I_{zz}\frac{\Delta z^2}{2} + I_{zzz}\frac{\Delta z^3}{6} + \ldots \\ &\approx -I(z)H\Delta z - I(z)(K-T+H^2)\Delta z^2 \\ &\quad +I(z)H(2(K-T)-H^2)\Delta z^3 \end{aligned} \tag{7.25}$$

where $I_z \approx -IH$ and $I_{zz} \approx -2I(K-T+H^2)$, and $I_{zzz} = 6IH(2(K-T)-H^2)$.

Therefore, even a slightly extended analysis of signal displacement (T) in the curvature sensor must incorporate at least the third order in the error expansion. In contrast, the geometric sensor has the advantage of an exact geometric model. This effectively accounts for both the displacement (T) and curvature uniformity (K) transparently.

7.3.2 Direct comparison with the geometric wavefront sensor

To compare the curvature sensor to the geometric wavefront sensor, we re-formulate the image difference as the difference between two integrated images. This is similar to, and allows comparison with, the histogram specification step, as shown in Figure 7.13. In the histogram specification step, the geometric wavefront sensor makes use of the displacement signal between two defocused images. In contrast, the curvature sensor uses the direct difference signal between the two images[12].

Both sensors then integrate the resultant difference signal to arrive at the wavefront. From this comparison, we can see that the key difference between the two wavefront sensors comes from the geometric wavefront sensor taking the horizontal difference in the histograms, which corresponds to the light ray displacement, giving the actual wavefront slope.

Here, the more accurate geometric sensor model eliminates any signal mismatch between the two image planes, removing the distinction between the edge and curvature regions. The errors introduced by the curvature sensor approximation are quantified by the difference between the "horizontal" and "vertical" histogram differences.

For small wavefront perturbations, as in a closed-loop adaptive optics system, C_{I_A} and C_{I_B} will be very similar, and the difference between the geometric sensor and the curvature sensor is small. With larger wavefront aberrations, as in open loop operating conditions,

[12]Hence, in its simplest form, a closed-loop control strategy simply tries to cancel the curvature sensor signal by matching the two out-of-focus images.

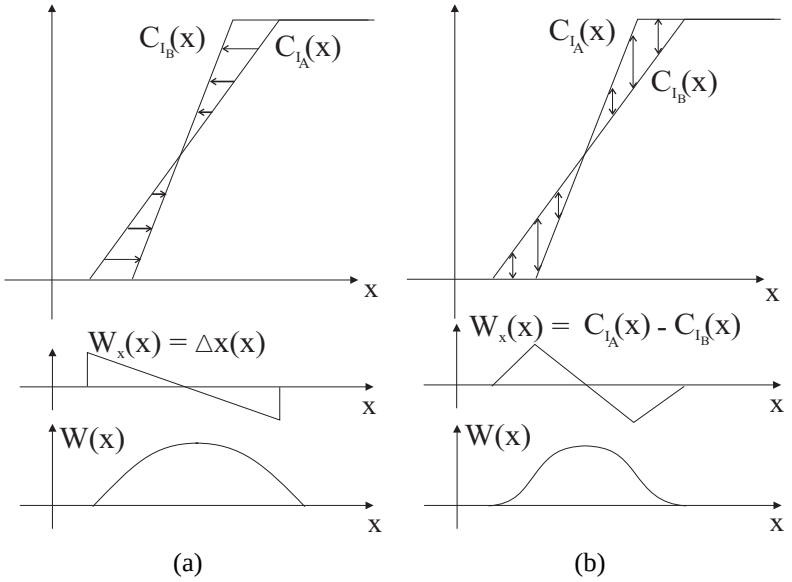


Figure 7.13 Comparison of the histogram specification process(a) with curvature sensing(b) in the estimation of slopes.

the edge signal errors in the curvature sensor become more significant, and the geometric wavefront sensor can provide more accurate wavefront estimates.

7.4 Theoretical performance

7.4.1 Photon noise analysis

Although the curvature and geometric wavefront sensors use the same inputs as data, the theoretical treatment of the wavefront sensing problem as presented by the geometric wavefront sensor is more precise. The principle of ray tracing to deduce the wavefront is also intuitively more consistent with geometric optics especially when applied to regions near the aperture edge, where the curvature sensor treatment is more messy.

This section examines the effect of photon noise on both wavefront sensors. A comparison of the two wavefront sensors is performed while ignoring read-noise to avoid detracting from the main analysis. The effects of photon noise, as continuing from Figure 7.4, are assumed to obey Poisson noise statistics, with independent noise in each imaging detector. The effect of various steps of each wavefront sensor algorithm on this noise is described, and the methods required to filter the noise are derived.

7.4.2 Intensity normalisation

The measured photon counts in the defocused images are determined by the Poisson statistics of photon noise, with variance equal to the mean or expected photon count. The photon noise in each pixel is independent and under bright illuminations, with high photon flux levels (above 50 photons), is approximately Gaussian. With the fluctuation caused by photon noise, the total photon count in each defocused image may no longer be equal. This difference in intensity results in mismatched histograms with unequal heights, as shown in Figure 7.14, so the histogram specification process is no longer defined. This problem is especially significant at extremely low light levels when individual photons are measurable.

Since an assumption of the ray tracing algorithm is that intensity is conserved, in order to apply histogram specification to wavefront estimation, the histograms must be matched. To satisfy this constraint, the intensities in the two images must be equalised, either by the addition or subtraction of a constant offset, or by scaling the intensity values of the two images.

The addition or subtraction of constant offsets may result in negative image intensity values. Furthermore, a constant offset maintains the mismatch in their histograms. Since it is the intensity *distribution* or shape that is used for estimating the wavefront, and not the absolute intensity levels, a more appropriate solution is to normalise the images by scaling the intensity values. This aligns the endpoints of the image histograms, as shown in Figure 7.14.

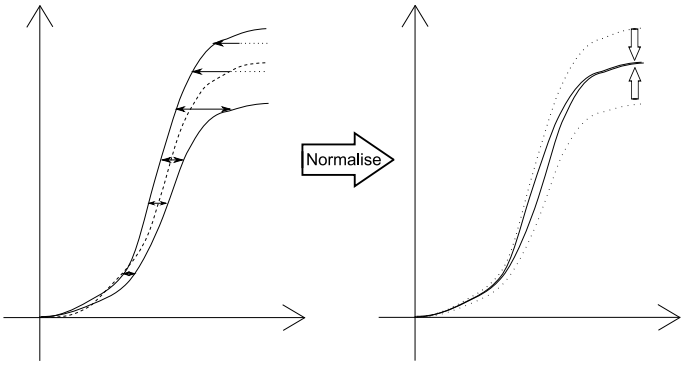


Figure 7.14 Due to fluctuations in the measured intensity, the image histograms are no longer matched (left), and have an undefined histogram specification. The images are equalised by the normalisation of the total intensity (right) to a nominal photon count of 1.

The normalisation step is dependent on the noise present in each image. Figure 7.15 shows the equivalent noise after intensity equalisation, obtained by subtracting the normalised

noisy image (or its histogram) from the original image[13]. The division by the total photon count (image plus noise) introduces some negative correlation into each pixel in the measurement plane.

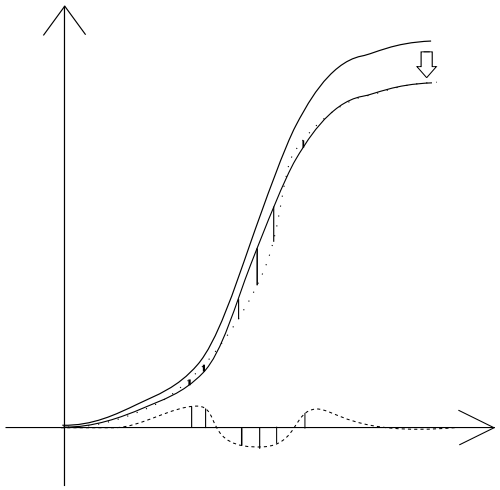


Figure 7.15 The noise after normalisation of the histogram is largest in the centre of the aperture, and zero at both endpoints, corresponding to the edges of the aperture.

The correlation can be derived by returning to the example in Section 7.1, where we start with the expected (noiseless) intensity measurement $I_B(x)$, and add noise to get $I_B(x) + n_B(x)$. The signal is then scaled to equalise the intensity level. The equivalent noise in the scaled signal is defined to be

$$\frac{I_B(x) + n'_B(x)}{\sum_x I_B(x)} = \frac{I_B(x) + n_B(x)}{\sum_x [I_B(x) + n_B(x)]} \tag{7.26}$$

where $n_B(x)$ is the photon noise at plane B, and the high intensity Gaussian approximation is assumed to hold true. $n'_B(x)$ is the equivalent normalised noise term after scaling the image intensity.

Re-arranging Equation 7.26, the normalised noise term now has an additional term dependent on the total noise level and the intensity in each pixel.

[13]Drawn to scale, in actual simulations, the histogram noise is too small to be seen against the scale of the histograms.

$$\begin{aligned} n'_B(x) &= \frac{n_B(x)\sum_x I_B(x) - I_B(x)\sum_x n_B(x)}{\sum_x (I_B(x) + n_B(x))} \\ &\approx \frac{n_B(x) I_{tot} - I_B(x)\sum_x n_B(x)}{I_{tot}} \\ &= n_B(x) - \frac{n_{tot}}{I_{tot}} I_B(x) \end{aligned} \tag{7.27}$$

with the approximation $\sum_x n_B(x) = n_{tot} = 0$ in the denominator.

The modified noise covariance matrix can be expressed in terms of the original raw Poisson noise covariance matrix. The original noise is assumed to be approximately Gaussian (due to a high photon count) and independent between pixels, with the noise variance equal to the intensity in that pixel,

$$\langle n_B(x) n_B(y) \rangle = \delta_{xy} I_B(x) = \delta_{xy} I_B(y) \tag{7.28}$$

where δ_{xy} is the Kronecker delta (being 1 for $x = y$, and 0 otherwise).

The noise in the pixels is independent from each other, (the noise correlation between different pixels is zero), allowing Equation 7.29 the be reduced using $\langle n_B(y) n_{tot} \rangle = \langle n_B(y) \sum_i n_B(i) \rangle = \langle n_B(y)^2 \rangle = I_B(y)$ and $\langle n_{tot}^2 \rangle = \langle (\sum_i n_B(i)) (\sum_j n_B(j)) \rangle = \sum_{ij} \langle n_B(i) n_B(j) \rangle = I_{tot}$, as follows

$$\begin{aligned} C_{n'_B}(x,y) &= \langle n'_B(x) n'_B(y) \rangle \\ &\approx \left\langle \left(n_B(x) - \frac{n_{tot}}{I_{tot}} I_B(x) \right) \left(n_B(y) - \frac{n_{tot}}{I_{tot}} I_B(y) \right) \right\rangle \\ &= \langle n_B(x) n_B(y) \rangle - \frac{I_B(x)}{I_{tot}} \langle n_B(y) n_{tot} \rangle - \frac{I_B(y)}{I_{tot}} \langle n_B(x) n_{tot} \rangle + \langle n_{tot}^2 \rangle \frac{I_B(x) I_B(y)}{I_{tot}^2} \\ &= \langle n_B(x) n_B(y) \rangle - \frac{I_B(x) I_B(y)}{I_{tot}} \end{aligned} \tag{7.29}$$

As a special case, let the noise variance in each pixel with a uniform intensity distribution across N pixels be σ^2 (equal to the intensity $I_B(x)$ in each pixel). Being uniform uncorrelated Gaussian noise, $C(x,y) = \delta_{xy}\sigma^2$. The covariance of the normalised noise, with a slight negative correlation between each pixel, is then

$$\begin{aligned} C_{n'_B}(x,y) &= \left(\delta_{xy} - \frac{1}{N}\right)\sigma^2, \text{ or, arranged into matrices} \\ C_{n'_B} &= \left(\boldsymbol{I} - \frac{1}{N}\mathbf{1}\right)\sigma^2 \end{aligned} \tag{7.30}$$

In the subsequent histogram specification step, the image and noise are first integrated to form a histogram. Histogram formation, a cumulative summing operation, is linear and can be described using the matrix operation $\boldsymbol{C}_{sum}$. The integrated noise was originally a Brownian noise. With normalisation, the histogram noise is still similar to Brownian noise, but with the additional condition that the noise at the endpoints (edges of aperture) is constrained to be 0.

The covariance matrix for the normalised histogram or integrated noise is given by

$$\begin{aligned} \left\langle \boldsymbol{C}_{sum}\boldsymbol{n'_B}(\boldsymbol{C}_{sum}\boldsymbol{n'_B})^T \right\rangle &= \boldsymbol{C}_{sum}\left\langle \boldsymbol{n'_B}\boldsymbol{n'^T_B} \right\rangle \boldsymbol{C}^T_{sum} \\ &= \boldsymbol{C}_{sum}\boldsymbol{C}_{n'_B}\boldsymbol{C}^T_{sum} \\ &= \left(min(x,y) - \frac{xy}{N}\right)\sigma^2 \end{aligned} \tag{7.31}$$

The variance of the normalised histogram noise, given by the diagonal elements of the matrix in Equation 7.31, is $(x - \frac{x^2}{N})\sigma^2$. Such a noise distribution is also commonly encountered in Monte-Carlo analysis and is known as the Brownian bridge[14].

After the image histograms have been formed, the next step in the geometric wavefront sensing algorithm is histogram specification, a non-linear process, introducing higher order errors into the data. To simplify analysis, especially at lower noise levels, histogram specification can be approximated with histogram subtraction as shown in Figure 7.16.

At low noise levels, the change from histogram specification to subtraction has negligible effects on the noise statistics, as the output noise is not noticeably different from the input. This allows us to replace the non-linear step with a linear one for noise analysis purposes.

[14]The Brownian bridge is commonly defined to be the process $w(x) - xw(1)$ bound to 0 at the endpoints $x = 0$ and $x = 1$, with $w(x)$ being a Brownian process with variance $\text{var}\{w(x)\} = x$. The Brownian bridge has a variance that depends on position, $x - x^2$.

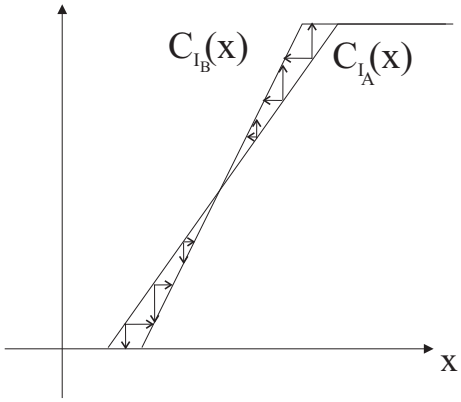


Figure 7.16 Approximating histogram specification with histogram subtraction.

The total effect on the noise is thus a combination of all linear steps —normalisation (resulting in $\boldsymbol{N'}$), image projection ($\boldsymbol{P}$), subtraction, and histogram formation or integration ($\boldsymbol{C}$)

$$\boldsymbol{N} = \boldsymbol{CPN'P}^T\boldsymbol{C}^T \tag{7.32}$$

7.4.3 Limits to resolution due to diffraction

Diffraction limits the resolution of the wavefront estimate in both the geometric and curvature wavefront sensors. The spatial resolution of the wavefront estimate is determined by the spatial blurring of the images at the out-of-focus images. The operation of the wavefront sensors put them in the Fresnel diffraction region, so Fresnel diffraction is the dominant operation. The extent of the smoothing during intensity propagation is known as the Fresnel length, or the Fresnel invariant or scale [38] (pg70).

Fresnel length

To illustrate the general behaviour of field propagation under Fresnel diffraction, we observe the effect of a small localised phase perturbation, $\Delta\phi(x,y)$, in a complex field, $A(x,y)e^{i\phi(x,y)}$. The propagated intensity in the Fresnel region is given by the Fresnel convolution equation (from Equation 3.27)

$$|u(x,y)|^2 = \left|A(x,y)e^{i\phi(x,y)} \odot F\right|^2 \tag{7.33}$$

where F is the Fresnel kernel $e^{i\frac{k}{2z}(x^2+y^2)}$.

Due to due to the small phase perturbation $\Delta\phi(x,y)$, the change in the propagated image intensity is

$$\begin{aligned}
& \left|(A(x,y)e^{i(\phi(x,y)+\Delta\phi(x,y))}) \odot F\right|^2 - \left|A(x,y)e^{i\phi(x,y)} \odot F\right|^2 \\
= & \left|A(x,y)e^{i\phi(x,y)} \odot F + p(x,y) \odot F\right|^2 - \left|A(x,y)e^{i\phi(x,y)} \odot F\right|^2 \\
= & |u(x,y) + p(x,y) \odot F|^2 - |u(x,y)|^2 \\
= & |u(x,y)|^2 + 2Re\{u(x,y)(\overline{p(x,y) \odot F})\} + |p(x,y) \odot F|^2 - |u(x,y)|^2 \\
= & 2Re\{u(x,y)\overline{p(x,y) \odot F}\} + |p(x,y) \odot F|^2
\end{aligned} \tag{7.34}$$

with the field perturbation $p(x,y)$ being related to the phase perturbation by

$$\begin{aligned}
p(x,y) &= A(x,y)e^{i\phi(x,y)+i\Delta\phi(x,y)} - A(x,y)e^{i\phi(x,y)} \\
&\approx A(x,y)e^{i\phi(x,y)}i\Delta\phi(x,y)
\end{aligned} \tag{7.35}$$

In Equation 7.34, the second order perturbation term $|p(x,y) \odot F|^2$ can be ignored, leaving the larger first order perturbation term $2Re\{u(x,y)\overline{p(x,y) \odot F}\}$. This change in intensity can also be represented as

$$\begin{aligned}
& 2Re\{u(x,y)\overline{p(x,y) \odot F}\} \\
= & 2Re\{u(x,y)\}Re\{p(x,y) \odot F\} + 2Im\{u(x,y)\}Im\{p(x,y) \odot F\}
\end{aligned} \tag{7.36}$$

For simplicity, the phase perturbation is assumed to be a small circular region with a constant phase offset.

$$\Delta\phi(x,y) = m\ \mathrm{circ}(kx,ky) \tag{7.37}$$

For a small enough circular diameter D (given by a large k), the Fresnel convolution of the field perturbation approximates Fraunhofer diffraction. The Fraunhofer diffraction pattern from a circular disc is given by the Jinc or Airy function with a quadratic phase term.

$$\begin{aligned} & 2Re\{u(x,y)\}Re\{e^{i\frac{k}{2z}r^2}Jinc(\frac{D}{2\lambda z}r)\} + 2Im\{u(x,y)\}Im\{e^{i\frac{k}{2z}r^2}Jinc(\frac{D}{2\lambda z}r)\} \\ = \ & 2Re\{u(x,y)\}Jinc(\frac{D}{2\lambda z}r)cos(\frac{k}{2z}r^2) + 2Im\{u(x,y)\}Jinc(\frac{D}{2\lambda z}r)sin(\frac{k}{2z}r^2) \end{aligned} \tag{7.38}$$

Equation 7.38 is effectively a modulation of the intensity by Jinc and sinusoidal functions. The perturbation modulation functions are shown in Figure 7.17, separately (top) and combined (bottom). The widths of the Jinc and sinusoidal terms are $\frac{\lambda z}{D}$ and $\sqrt{\lambda z}$ respectively.

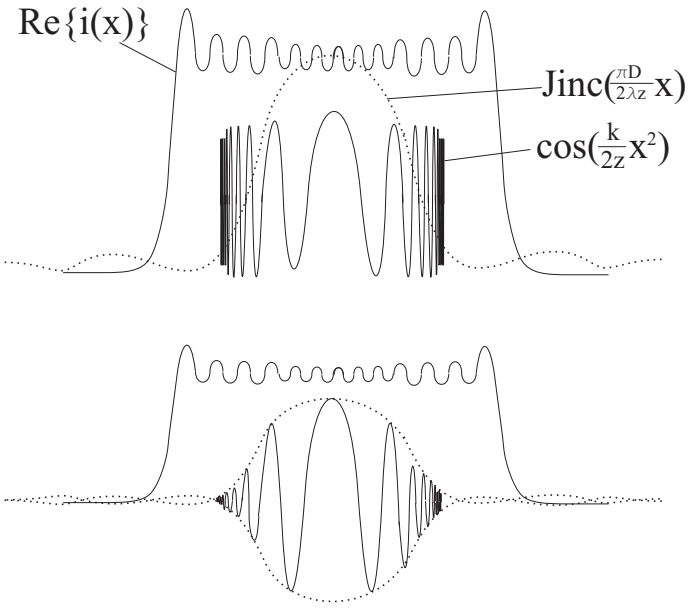


Figure 7.17 "Linearised" point-spread-function of Fresnel propagation for a sub-aperture.

Although not strictly accurate, the blurring function in Figure 7.17 may be considered to be the approximate extent of the point-spread-function of the Fresnel kernel. This describes the propagation of the aperture phase function to the defocused imaging plane, and consists of a central region about $\sqrt{\lambda z}$ in width, and side-lobes bound by an envelope that is about

$\frac{\lambda z}{D}$ in width. The fringes in the side-lobes oscillate so fast that they are smoothed out when averaged over the whole phase function, and in any case, are under-sampled in practice. The bound on the spatial resolution of the wavefront estimate is thus determined by the central lobe, $\sqrt{\lambda z}$.

This measure of blurring applies only for short distances, where the Fresnel approximation is valid. At larger distances, the effects of Fraunhofer diffraction (on the order of $\frac{\lambda z}{D_{tel}}$, with D_{tel} being the aperture of the optical system) supersedes Fresnel diffraction, so the Fresnel length is no longer the dominant blurring term. Shown in Figure 7.18, the nominal division between the Fresnel and Fraunhofer regions is normally consider to be the Rayleigh distance $z_R = \frac{D_{tel}^2}{\lambda}$, which is also where the Fresnel length is equal to the aperture diameter of the imaging system, $\sqrt{\lambda z} = \frac{\lambda z}{D_{tel}}$(leading to $\sqrt{\lambda z} = D$).

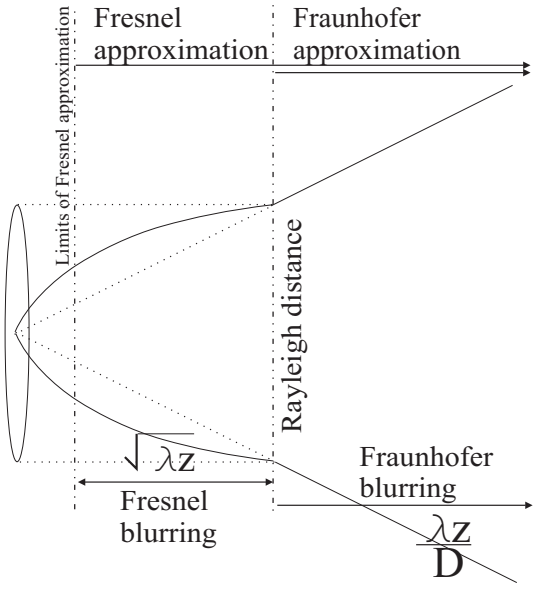


Figure 7.18 Approximate boundaries of the Fresnel and Fraunhofer regions for a planar wavefront.

The blurring due to Fresnel diffraction is independent of the telescope aperture size and the complex field at the aperture. In particular, the severity of atmospheric turbulence has no significant effect on the resolution of the geometric and curvature wavefront sensors when operated in the geometric optics region.

Although the Fresnel approximation is valid over all distances where the Fraunhofer approximation is applicable, the Fresnel length as a measure of blurring is only valid at short distances (in the Fresnel region). The Fraunhofer or far-field diffraction pattern has an approximate width of $\frac{\lambda z}{D_{tel}}$, but only if no aberrations are present at the imaging aperture.

In contrast to the constant Fresnel length in the Fresnel diffraction region, the Fraunhofer image size is enlarged by the presence of aberrations. For example, under Kolmogorov turbulence, the long-term exposure image is roughly $\frac{\lambda z}{r_0}$ in size. Fried's parameter, r_0, is commonly thought of as the equivalent diameter of an un-aberrated (smaller) imaging aperture.

In summary, the blurring due to diffraction is dependent on several factors. At short distances, within the Fresnel diffraction region, the image blurring is given by the Fresnel length, $\sqrt{\lambda z}$, and is independent of the wavefront at the aperture. At longer distances, the Fraunhofer approximation dominates, and the image size is determined by the wavefront at the imaging aperture, and the size of the aperture.

Fresnel blurring in wavefront sensors

The defocused imaging planes in the geometric and curvature wavefront sensors are displaced from the focus sufficiently to allow the geometric optics approximation to be used. This is equivalent to imaging in the Fresnel region ($z' \ll z_R$), so the Fresnel length is the most appropriate measure of sensor resolution, and represents the limit to the resolution that is achievable in the wavefront sensors. The geometric optics approximation is only valid when applied to image features larger in scale than the Fresnel length, and no longer apply on scales smaller than the Fresnel length.

The Fresnel length is given by $\sqrt{\lambda z'}$, where z' is the virtual propagation distance, which was previously shown to be related to the actual telescope dimensions by $z' \approx \frac{f^2}{l}$. In actual terms, the blurring caused by Fresnel diffraction in the defocused imaging planes is given by re-scaling (see Equation 7.1)

$$\sqrt{\lambda z'}\frac{z}{z'} = \sqrt{\lambda l} \tag{7.39}$$

The direct Fresnel length expression $\sqrt{\lambda z}$ or $\sqrt{\lambda (f-l)}$ is no longer valid because it is larger than the image size, as explained by Equation 7.38. The effect of diffraction, as previously calculated, (represented in Figure 7.17), requires the image to be larger than either of $\sqrt{\lambda z}$ or $\frac{\lambda z}{D}$, an assumption that is no longer valid.

Figure 7.19 demonstrates the decreasing sensor resolution (due to increased blurring along the positive y-axis) against distance. The resolution limits posed by Fraunhofer ($\frac{\lambda z}{D_{tel}}$)

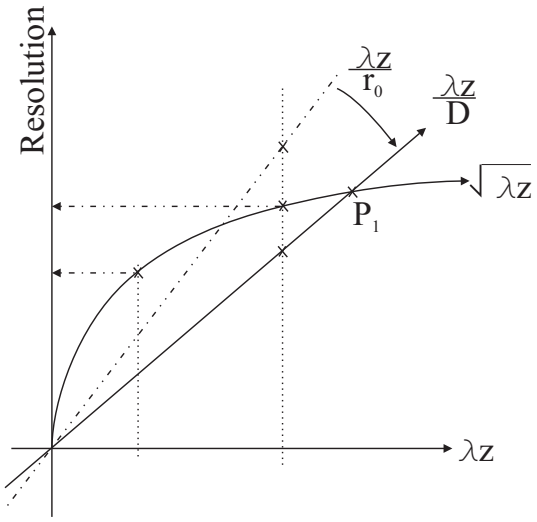


Figure 7.19 Wavefront spatial resolution of the curvature sensor.

and Fresnel ($\sqrt{\lambda z}$) diffraction are shown in solid lines, with the "cross-over" point at the Rayleigh distance $\frac{D^2}{\lambda}$ marked P_1. The propagation distance has to be less than this, and is thus constrained to lie to the left of P_1.

The conventional measure of the (spatial) wavefront resolution of the curvature wavefront sensor [43, 84, 86] is frequently explained by Fraunhofer diffraction only, and is therefore assumed to be limited by the wavefront aberrations at the aperture, as shown with the dotted line ($\frac{\lambda z}{r_0}$). Under closed-loop operation, when the input wavefront is partially compensated (resulting in a larger equivalent r_0), the performance of the curvature sensor then increases.

However, the conventional measure of spatial resolution using Fraunhofer diffraction overestimates the achievable resolution, which is determined by Fresnel diffraction. The optimal propagation distance of curvature sensors is usually closer than $\frac{r_0^2}{\lambda}$ (similar to the Rayleigh distance), the "cross-over" point where the Fresnel length is greater than $\frac{\lambda z}{r_0}$. Furthermore, the diffraction blurring anticipated by $\frac{\lambda z}{r_0}$ is not valid in the Fresnel region. Even in the Fraunhofer region, $\frac{\lambda z}{r_0}$ refers to the approximate width of the long-term exposure image, whereas a short-term exposure image is more appropriate for comparison with the Fresnel length.

Therefore, the resolution achievable in the defocused wavefront sensors is determined by the defocus distance, and is proportional to $\sqrt{\lambda l}$. This is worse than the often cited value determined from Fraunhofer diffraction or the severity of turbulence, r_0. The limit posed by

Fresnel diffraction is not affected by the increased sensitivity during operation in a closed-loop adaptive optics system. However, closed-loop operation can lead to improved performance (Section 7.3.1) by reducing modelling errors in the wavefront sensors.

7.5 Simulations

Section 7.3.1 introduced a treatment of the errors in the curvature sensor. Due to non-linearities and the complexity of the error propagation analysis, a simpler approximate way to compare sensor performance is through simulations of the sensors under various conditions.

Kolmogorov phase-screens are generated independently [42] with turbulence severity for $\frac{D}{r_0}$ ranging from 0.1 to 25. Assuming a telescope diameter of 1m, discretised with 250 pixels, the phase-screens are then propagated forward and backward through free-space to various distances ranging from $\pm 14000m$ to $\pm 200000m$. Each pair of propagated images represent the defocused inputs to the wavefront sensors.

Although both wavefront sensors can work with broadband light, only narrowband light at 600nm is used to reduce the computational effort. At this wavelength, the Rayleigh distance is approximately $1700000m$, and the extreme range of the propagation distance chosen ($200000m$) already suffers from some diffraction, and the propagated image no longer resembles an image of the telescope aperture. Similarly, at the highest phase aberrations ($\frac{D}{r_0} = 25$), the defocused images are too aberrated, and the simulation results are less useful.

In both the geometric and the curvature wavefront sensors, only the first 20 Zernike modes are considered in the simulation, with the remaining higher orders ignored when calculating the phase error. 16 projection angles are used in the geometric sensor, and are more than enough[15] to completely and unambiguously recover the first 20 Zernike modes in the wavefront. Photon noise with Poisson statistics, assuming a mean of 500 photons in each image, is added to the defocused images. As described in previous sections, image normalisation is performed to equalise the intensity in both images. Although this is only necessary in the geometric wavefront sensor, it is also performed in the curvature sensor for consistency, to aid comparison.

At 500 photons, the mean intensity is high enough that the effects of image normalisation is

[15] As explained in the Appendix, more than 10 samples or 5 projections are required for the maximum azimuthal frequency of 5.

minimal. The simpler Gaussian noise model and its corresponding least-mean-square solution is chosen over the optimal Brownian bridge noise model, which requires a maximum-likelihood solution. The inverse wavefront estimation problem is thus performed using direct least-squares matrix inversion in both sensors.

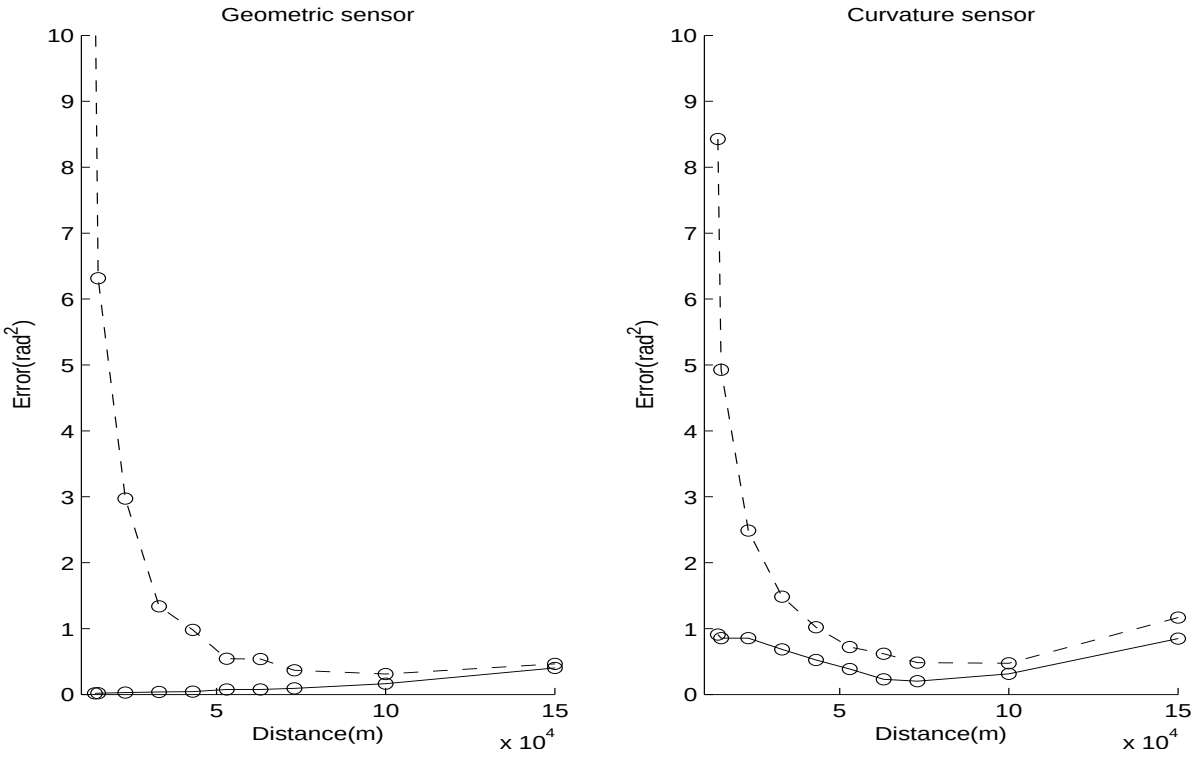


Figure 7.20 The geometric and curvature wavefront sensors at $\frac{D}{r_0} = 2$, without photon noise (solid line) and with photon noise (dashed line). The datapoints corresponding to the propagation distances used in the simulation are marked with circles.

Figure 7.20 shows the wavefront estimation error in both wavefront sensors for the first 20 Zernike modes with and without photon noise. In the absence of photon noise, the only sources of error are modelling errors (only in the curvature sensor) and the lowered resolution due to Fresnel diffraction (both types of errors increase with distance). This holds for the geometric sensor in the simulation results, but not for the curvature sensor because of modelling errors. To keep the curvature sensor comparable to the geometric sensor, a matrix is used to describe linear relationship between the input wavefront and curvature sensor signal, so the edge signal is not explicitly modelled. As a result, the curvature sensor under-estimates the wavefront slope, the largest phase term.

When Poisson noise (with a mean of 500 photons in each image) is added, the errors in both wavefront sensors are dominated by photon noise. The sensitivity of the sensors in-

creases with propagation distance, so the error decreases with distance. With the precision of the sensors reduced by photon noise, the large resulting error reduces the relative error contribution from the effects of diffraction. Thus, at low photon counts, Fresnel diffraction is not an important factor in determining the resolution-precision trade-off in the geometric and curvature wavefront sensors. In contrast, in the Shack-Hartmann sensor, as explained in Section 6.3.2, both resolution and precision can have significant effects on the combined sensor error, so the trade-off between resolution and precision is an important concern.

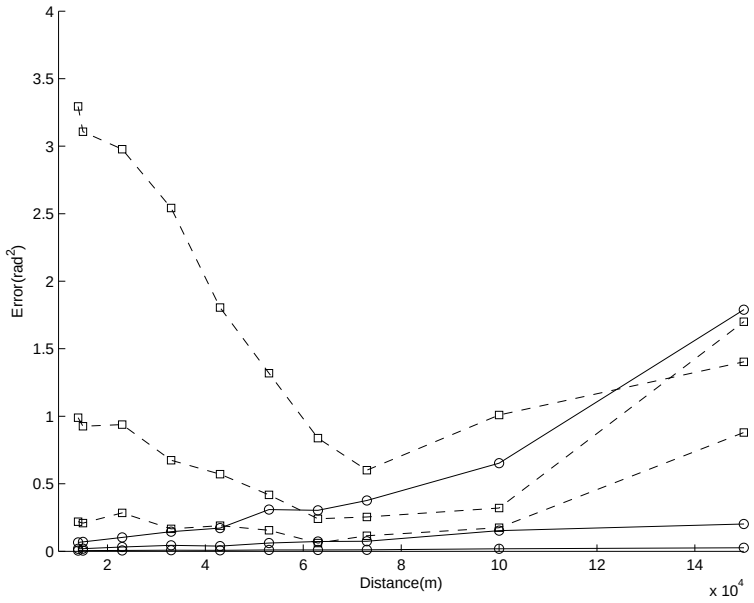


Figure 7.21 Errors in the geometric wavefront sensor (solid lines) and curvature sensor (dashed lines) without photon noise, with increasing turbulence levels of $\frac{D}{r_0}$=0.5, 1, and 2. Each datapoint is also marked with a circle.

Figure 7.21 compares the errors in the geometric wavefront sensor with the curvature wavefront sensor. The curves approximate the sensor estimation error and loss in resolution by measuring the total error without photon noise (effectively reproducing Figure 7.20 without photon noise, and for a wider range of turbulence levels). While the error curves in both sensors increase with the turbulence level, the geometric wavefront sensor always out-performs the curvature sensor with a lower error at most distances. At larger distances and turbulence wavefronts, the geometric wavefront sensor seems to under-perform the curvature sensor. This is due to the assumptions of geometric optics breaking down (refer to the commentary to Figure 7.6 on ray crossings), and stronger diffraction effects.

In the presence of photon noise (Figure 7.22), no discernable difference (within the simulation tolerance) between the two sensors can be observed. Although the geometric sensor also out-performs the curvature sensor at higher turbulence levels (not shown here), the

results are inconclusive there because of the severe diffraction and ray crossing effects.

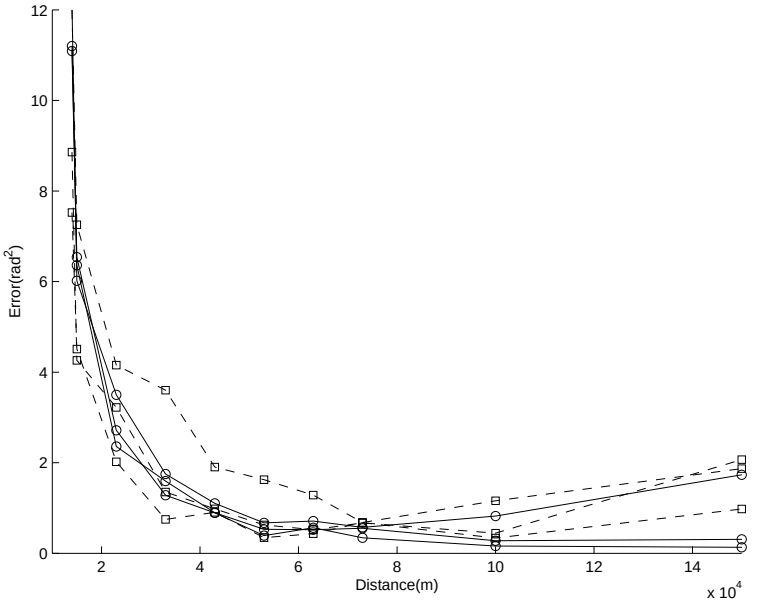


Figure 7.22 Errors in the geometric wavefront sensor (solid lines) and curvature sensor (dashed lines) with photon noise, with increasing turbulence levels of $\frac{D}{r_0}$=0.5, 1, and 2. Each datapoint is also marked with a circle.

Figure 7.23 shows the total estimation errors for the geometric wavefront sensor with photon noise (mean 500 photons) at different levels of turbulence. The error in the geometric wavefront sensor increases with turbulence level ($\frac{D}{r_0}$). This is consistent with geometric optics where increasing the phase aberration (and consequently the wavefront slope) requires the propagation distance z to be decreased proportionately, in order to keep the propagated image constant. There appears to be an optimal propagation distance where the error is lowest —beyond that, the error increases again because the image is too distorted from the diffraction and ray crossings.

7.6 Conclusion

The geometric sensor uses geometric optics to estimate wavefront from defocused images through ray-tracing. The position and displacement of the light rays are recovered using histogram specification, and used to infer the wavefront at the optical aperture. The algorithm assumes that the wavefront is small enough, so that no light rays cross path within the propagation region. The histogram specification step also requires the intensity in both defocused images to be equal, and requires the images to be normalised in the presence of noise. Image normalisation modifies the noise statistics into a Brownian bridge. In the

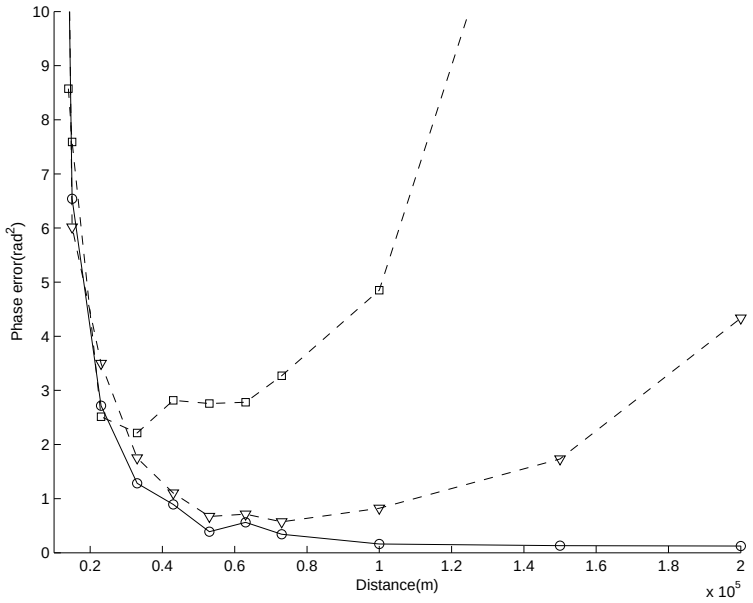


Figure 7.23 The effect of increasing turbulence on the estimation error in the geometric wavefront sensor, for $\frac{D}{r_0}$=0.5 (solid line, circular points), 2 (dashed line, triangular points), and 4 (dot-dashed line, square points).

simulations performed in this section, the precise statistics of sensor noise is not important, so the least-squares approach is chosen because of its simplicity and robustness.

The curvature sensor is an approximation to the geometric wavefront sensor. Using a few simplifying assumptions, the difference between two defocused images are used as an estimate of the wavefront curvature. Extended analyses of the curvature sensor have focused on the lower order errors in the intensity and wavefront propagation equations. Due to the complexity of the analyses, a direct simulation is used to compare the geometric sensor with the curvature sensor.

It was found that the geometric sensor achieves lower wavefront estimation errors compared to the curvature sensor in the absence of noise. This reflects the more accurate geometric optics algorithm in the geometric wavefront sensor[16]. In the presence of photon noise, there may be some improvements, but the major factor determining performance, sensitivity, is common to both wavefront sensors, so no major improvement can be seen. Perhaps simulations with larger wavefronts (and shorter propagation distances to ensure that the geometric optics approximations are met) will show a difference in performance.

The effects of diffraction (loss in sensor resolution) within the geometric optics region were

[16]Note however that there are some simulation modelling errors in the curvature sensor.

found to be negligible compared to the effects of photon noise (reduced sensor precision). Outside of the geometric region, where Fraunhofer diffraction dominates (large z), or where there were too many ray crossings (large wavefronts, $\frac{D}{r_0}$), the geometric optics approximation breaks down, and the error increases quickly.

Chapter 8

Conclusion

This thesis examined four main types of wavefront sensors in adaptive optics systems. A uniform description of the sensors was provided and the fundamental performance limits of the sensors were compared using a geometric optics model.

8.1 Summary

Atmospheric turbulence distorts the images collected by astronomical imaging telescopes, degrading resolution. Real-time adaptive optics systems detect the wavefront aberrations introduced by atmospheric turbulence and correct them using a deformable mirror, in a closed-loop system. Due to the relative youth of this field, many possible designs for wavefront sensors exist, but have not been examined and compared in great detail. This thesis proposes a unified framework for presenting the operation of wavefront sensors to allow a uniform comparison of the wavefront sensors.

Chapters 1 to 4 introduced various concepts and mathematical tools used in the subsequent chapters.

The quad-cell is an image displacement estimator consisting of intensity detectors arranged in a 2x2 array. It is often employed at the focal plane, where image displacement corresponds to the aberration wavefront slope at the optical aperture. The main sources of noise examined in the quad-cell are instrument read noise and photon noise. After developing the Strehl ratio for measuring quad-cell performance, several different methods for quantifying the performance of the quad-cell are compared. Compared to the fundamental limit posed

by the Cramer-Rao bound, slope estimation with the quad-cell is an attractive trade-off given its simplicity and cost. In the Shack-Hartmann and pyramid wavefront sensors, the quad-cell arrangement is used to estimate wavefront aberrations.

Due to the duality between the imaging and aperture planes, there is a fundamental trade-off between resolution and precision in the Shack-Hartmann and pyramid wavefront sensors. The trade-off is described in terms of the Fourier transform and shown with simulations in Section 6.3.2. The resolution is determined by the wavefront sub-division operation, which separates a wavefront into smaller sections. Within each section, the precision of the wavefront estimate is determined by a local slope sensing operation (using the quad-cell).

By comparing sensor operations in the dual imaging planes, a comparison of the precision of the two sensors is made based on the quad-cell analysis. The crucial difference in the order of the sub-division operation leads to a theoretically higher performance from the pyramid wavefront sensor. Simulations within a range of operating conditions show better performance from the pyramid wavefront sensor. From a practical standpoint, the pyramid sensor also allows more flexibility in adjusting the sensor resolution and precision.

The geometric and curvature wavefront sensors are the other pair of wavefront sensors compared in this thesis because of their similarities. The sensor inputs consist of two opposing equally defocused images. The geometric sensor is shown to be a geometric optics model which recovers the wavefront aberration at the optical aperture by ray tracing. The more popular curvature wavefront sensor is shown to be an approximation to the geometric wavefront sensor. The simpler algorithm in the curvature sensor is at a cost to estimation performance due to curvature signal displacement and mis-matched aperture edge signals.

An analysis of the effect of photon noise on the measurement precision of the geometric wavefront sensor, resulting in a Brownian noise model, is presented. Diffraction also limits performance by reducing the sensor resolution. The conventional Fraunhofer diffraction model is shown to over-estimate the performance achievable in the defocused sensors, and the Fresnel diffraction model is suggested as a replacement.

Some observations are also presented on the implementation of the geometric wavefront sensor, based on image recovery through projections.

8.2 Future work

Further extensions to much of the ideas presented in this thesis would be helpful in resolving several outstanding issues. In this section, I suggest some of the more interesting and potentially fruitful areas of discussion.

The quad-cell is the one of the longest known slope detector and is well-understood. Even then, new interpretations of the slope detection operation and novel variations on the quad-cell theme continue to be implemented, as seen in the pyramid wavefront sensor. The image truncation on the boundaries of the quad-cell perform a spatial filtering operation, but issues of aliasing and truncation introduced in Section 4.3.1 and Section 5.1.1 remain under-explored.

The closed-loop model presented in Chapter 6 is very much simplified in order to contrast the Shack-Hartmann and pyramid wavefront sensors. A detailed model of the dynamically compensated system, incorporating atmospheric statistics and control systems analysis, would help characterise the compensated output over time. The modelling process would involve estimating the relative contribution of each parameter in the system, and knowing which ones are not important, and could safely be ignored.

The degradation in resolution due to diffraction effects in the defocused wavefront sensors have been explained using Fresnel diffraction. Using the Fourier transform analysis, and a more precise definition of resolution, it may be possible to quantify the effects of diffraction on sensor resolution. On a more practical note, simulations of Fresnel diffraction require a discretised approximation of the Fresnel kernel and propagated field. For a fixed pixel size, there is a limit to the shortest possible propagation distance that can be simulated. Simple techniques to shorten this constraint would have been useful in the simulation for Chapter 7.

The behaviour of the geometric sensor in the presence of photon noise was simulated with a high number of photons, approximating Gaussian white noise. In the extreme, with low photon counts, the geometric sensor is much more unpredictable. Unlike the three other fully linear sensors, at low photon count levels, the presence of each single individual photon can have wildly different effects on the sensor output.

In the geometric wavefront sensor, the observed property of the Zernike polynomials under projection, presented in the Appendix of this thesis, is also an unsolved conjecture. A proof of this conjecture would complete the modal wavefront analysis of the geometric wavefront

sensor.

8.2.1 Unification of wavefront sensors

The chapter layout of this thesis reflects the similarity between pairs of wavefront sensors —the Shack-Hartmann and the pyramid wavefront sensors that form a dual Fourier pair, and the geometric and curvature sensor that are based on the same inputs. A theoretical framework linking any of the Shack-Hartmann or pyramid sensors with the geometric or curvature sensors would complete a link forming an series of transformations between any two sensor.

One possible direction here would be to focus on the similarities between the Shack-Hartmann and curvature wavefront sensors. At the same time, this complements the experimental comparisons reported by Rigaut et. al [81].

In the Shack-Hartmann sensor, a wavefront is first subdivided into smaller sections. The mean slope within each section is estimated using a quad-cell positioned at the focal plane. By defocusing the measurement plane, as shown in Figure 8.1(a), then recombining (reversing the subdivision operation) the quad-cell detectors to form an imaging array, we obtain the curvature sensor.

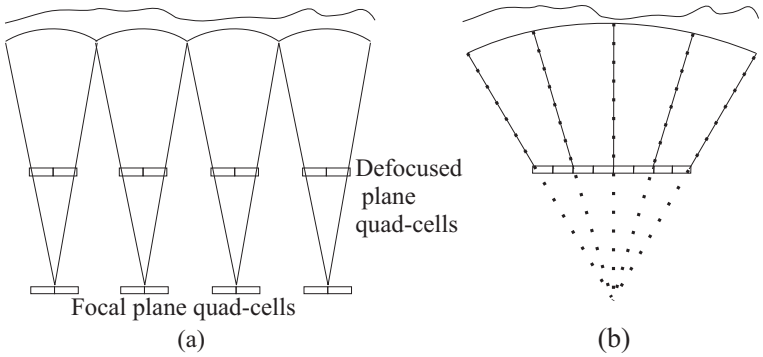


Figure 8.1 A side-by-side comparison of the Shack-Hartmann (a) and curvature (b) wavefront sensor.

In the curvature sensor, the wavefront subdivision operation is now implicit in image formation, which localises the wavefront signal. The quad-cell slope detection equation needs to be updated to take into account any intensity "spill-over" from the newly joined neighbouring detector elements. The intensity level in each detector element results from the gain or loss of light to and from its neighbours. Any change in the intensity thus arises from the difference in the wavefront slope at the pixel boundaries, proportional to the mean

curvature of the wavefront within the pixel or sub-aperture.

Further analyses could also incorporate the scidar and slodar techniques into this wavefront sensor framework, since they have very similar optical arrangements.

Appendix

8.3 Projections of Zernike polynomials

During my analyses of the geometric wavefront sensor involving the radon transform, some useful properties of the Zernike polynomials were observed. The properties of the Zernike polynomials under projection may find wider application in many projection based imaging techniques like computed tomography imaging or magnetic resonance imaging.

It was found that the rotational invariance of the Zernike polynomials translates to a projection direction invariance after a radon transform. Furthermore, all Zernike polynomials within the same radial order seem to possess the same primary projection function. Since the recovery of any image (within a circular support region) from their projections can be described by its Zernike polynomial representation, the properties of the Zernike polynomials can be used simplify the inverse problem by reducing it into smaller sub-problems for each Zernike radial group.

The properties observed here represent a special case of Cormack's projection functions [19], which examined the projection of radially symmetrical functions, their inverses, and the uniqueness of the solution. Indeed, in a subsequent paper [20](Part 2), Cormack described the Zernike polynomials and demonstrated in an experiment their projection solutions. However, the conventions used to describe the Zernike polynomials were slightly different from those adopted here.

The definitions for the Zernike polynomials in Equation 2.78, adopted from Noll [65], is reproduced here for reference.

$$Z_i(r,\theta) = \begin{cases} \sqrt{n+1}R_n^0(r) & \text{if m = 0,} \\ \sqrt{n+1}R_n^m(r)\sqrt{2}\cos(m\theta) & \text{if } m \neq 0, \text{ and i is even,} \\ \sqrt{n+1}R_n^m(r)\sqrt{2}\sin(m\theta) & \text{if } m \neq 0, \text{ and i is odd,} \end{cases} \tag{8.1}$$

where

$$R_n^m(r) = \sum_{s=0}^{\frac{n-m}{2}} \frac{(-1)^s(n-s)!}{s![\frac{n+m}{2}-s]![\frac{n-m}{2}-s]!} r^{n-2s} \tag{8.2}$$

for $0 \leq r \leq 1$ and non-negative integral values of n and m, with $m \leq n$ and $n-|m|$ being even.

The polar coordinates (r,θ) can be converted back and forth to rectangular coordinates (x,y). Let the x-axis be parallel to the line along azimuthal angle 0, and the y-axis to $\frac{\pi}{2}$ radians. The radon transform of a particular Zernike polynomial $Z_i(x,y)$ is defined to be

$$\zeta_i(u,\phi) = \int_{-\infty}^{\infty} Z_i(x,y)\, dv \tag{8.3}$$

where the projection is taken along the v-axis corresponding to the (parallel to a line at) angle ϕ. The u-axis is orthogonal to the v-axis and lies along $\phi + \frac{\pi}{2}$.

For example, integrating along the y-axis corresponds to $\phi = \frac{\pi}{2}$.

$$\zeta_i(u,\frac{\pi}{2}) = \int_{-\infty}^{\infty} Z_i(x,y)\, dy, \text{ for } u = x \tag{8.4}$$

Depending on the symmetry of the Zernike polynomials, their projections are given by

$$\begin{aligned}
\zeta_i(u,\frac{\pi}{2}) &= \int_{-\infty}^{0} Z_i(x,y)\,dy + \int_{0}^{\infty} Z_i(x,y)\,dy \\
&= \int_{0}^{\infty} Z_i(x,-y)\,dy + \int_{0}^{\infty} Z_i(x,y)\,dy \\
&= \begin{cases} \zeta'_{(n,0)}(u) & \text{if m = 0,} \\ \zeta'_{(n,m)}(u) & \text{if } m \neq 0\text{, and i is even,} \\ 0 & \text{if } m \neq 0\text{, and i is odd,} \end{cases}
\end{aligned} \tag{8.5}$$

For odd i and $m \neq 0$, $Z_i(x,-y) = -Z_i(x,y)$, so the projection along the y-axis is 0. For even i, or when $m = 0$, $Z_i(x,-y) = Z_i(x,y)$, no cancellation occurs, and the resulting "primary projection" is named ζ' with the corresponding radial order n and azimuthal frequency m as subscripts.

For any arbitrary rotation angle ϕ, a Zernike polynomial can be expressed in terms of the sinusoidal and cosinusoidal Zernike pairs with the same radial order and azimuthal frequency (this is trivially true when $m = 0$). Consequently, the projection of a Zernike polynomial along any arbitrary angle is a weighted sum of the sinusoidal projection (always 0) and the cosinusoidal projection. For all even i, and corresponding pair $i \pm 1$[1], this is

$$\begin{aligned}
\zeta_i(u,\phi) &= \int_{-\infty}^{\infty} Z_i(r,\theta)\,dv \\
&= \int_{-\infty}^{\infty} Z_i(r,\theta+\phi)\,dx \\
&= \int_{-\infty}^{\infty} Z_i\left(r,\theta+\phi-\frac{\pi}{2}\right)\,dy \\
&= \int_{-\infty}^{\infty} \cos\left(m(\phi-\frac{\pi}{2})\right) Z_i(r,\theta) - \sin\left(m(\phi-\frac{\pi}{2})\right) Z_{i\pm1}(r,\theta)\,dy \\
&= \cos\left(m(\phi-\frac{\pi}{2})\right) \int_{-\infty}^{\infty} Z_i(r,\theta)\,dy - 0 \\
&= \zeta'_{(n,m)}(u)\cos(m(\phi-\frac{\pi}{2}))
\end{aligned} \tag{8.6}$$

Figure 8.3 shows the projection of astigmatism, the 5th Zernike polynomial, over several

[1]Examples of the Zernike pairs are the tip/tilt terms 2 and 3, the astigmatic terms 6 and 5, or the coma terms 8 and 7.

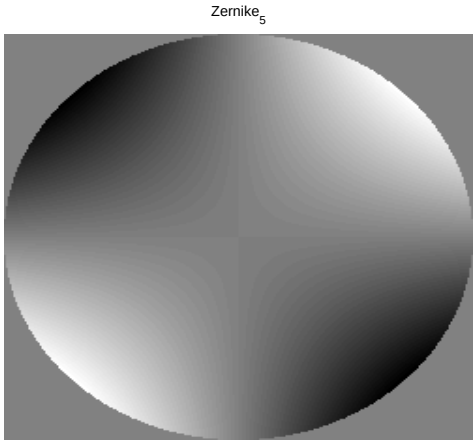


Figure 8.2 The 5^{th} Zernike polynomial, corresponding to astigmatism.

angles. The projection functions are identical over all angles to within a scale factor. This confirms the result from Equation 8.6.

The coefficients of the original pair of Zernike polynomials can be derived by fitting the projections to $\cos\left(m(\phi - \frac{\pi}{2})\right)$. To recover a Zernike pair with azimuthal frequency M, the Nyquist limit requires *more than* $2M$ samples over a revolution of projections. Since the projections at angles ϕ and $\phi + \frac{\pi}{2}$ are the same (reflections), this requires more than M equally spaced projections in the radon transform.

We now examine the primary projection, which, for all cosinusoidal terms, is given by

$$\begin{aligned} \zeta'_{(n,m)}(u) &= \int_{-\infty}^{\infty} Z_i(x,y)\, dy \\ &= \int_{-\infty}^{\infty} \sqrt{n+1} R_n^m(r)\cos(m\theta) \left\{ \begin{array}{c} 1 \quad \text{if m = 0,} \\ \sqrt{2} \text{ if } m \neq 0 \end{array} \right\} dy \end{aligned} \tag{8.7}$$

Ignoring all constant scale factors, within the same radial order, the primary projection

$$\int_{-\infty}^{\infty} R_n^m(r)\cos(m\theta)\, dy \tag{8.8}$$

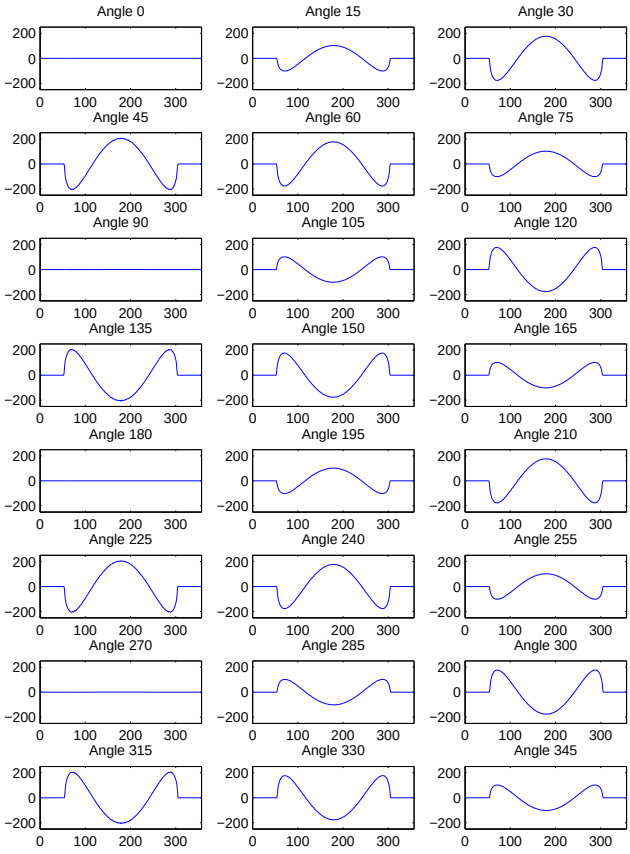


Figure 8.3 The projection of Z_5 (see Figure 8.2) over one revolution, showing the invariance of the projection image (to within a scale factor). Note also that the top-half of the plots are the same (they are actually reflected across the y-axis) as the plots in the bottom-half, since they are simply projections in opposite directions.

is identical for all m, so the primary projection is in fact parametrised only by n, $\zeta'_n(u)$. As shown in Table 8.1, the projections are given by Chebyshev Polynomials of the 2nd kind.

I am not aware of any analytical proof for this assertion. However, using symbolic integration techniques, this observation has been verified up to at least $n = 100$. I propose the conjecture that all Zernike polynomials with the same radial order have the same primary projection function (ignoring the 1 or $\sqrt{2}$ scale factor in Equation 8.7), regardless of azimuthal frequency. An exception to this are polynomials with odd i (as has been shown, these have projections of zero), using the Noll [65] numbering convention. Perhaps some of the identities and reasoning in [20] could be used to prove this.

n	Polynomial number	Projection $\frac{\int_{-\infty}^{\infty} R_n^m(r)\cos(m\theta)dy}{\sqrt{1-x^2}}$
0	1	2
1	2-3	$2x$
2	4-6	$-\frac{2}{3}+\frac{8}{3}x^2$
3	7-10	$-2x+4x^3$
4	11-15	$\frac{2}{5}-\frac{24}{5}x^2+\frac{32}{5}x^4$
5	16-21	$2x-\frac{32}{3}x^3+\frac{32}{3}x^5$
6	22-28	$-\frac{2}{7}+\frac{48}{7}x^2-\frac{160}{7}x^4+\frac{128}{7}x^6$
7	29-36	$-2x+20x^3-48x^5+32x^7$
8	37-45	$\frac{2}{9}-\frac{80}{9}x^2+\frac{160}{3}x^4-\frac{896}{9}x^6+\frac{512}{9}x^8$
9	46-55	$2x-32x^3+\frac{672}{5}x^5-\frac{1024}{5}x^7+\frac{512}{5}x^9$
10	56-66	$-\frac{2}{11}+\frac{120}{11}x^2-\frac{1120}{11}x^4+\frac{3584}{11}x^6-\frac{4608}{11}x^8+\frac{2048}{11}x^{10}$

Table 8.1 Prime projections of the Zernike polynomials.

8.3.1 Projection functions of Zernike polynomials

Table 8.1 shows the symbolically computed projection functions within some radial order n, and their corresponding Zernike polynomials, up to $n = 10$. The symbolic integration involved in the projection is partly simplified using the Chebyshev identity.

$$\begin{aligned}\int_{-\infty}^{\infty} R_n^m(r)\cos(m\theta)\,dy &= \int_{-\infty}^{\infty} R_n^m(r)\left(2\cos\theta\cos((m-1)\theta)-\cos((m-2)\theta)\right)\,dy \\ &= 2x\int_{-\infty}^{\infty}\frac{R_n^m(r)}{r}\cos((m-1)\theta)\,dy \\ &\quad -\int_{-\infty}^{\infty} R_n^m(r)\cos((m-2)\theta)\,dy \end{aligned} \tag{8.9}$$

8.3.2 Final thoughts

The chief disadvantage of this method is that the tabulated polynomials need to be discretised into a matrix and least-squares inverted in order to obtain the Zernike decomposition of an image from its projections. An analytical expression for Table 8.1 would result in a more practical inversion process.

References

[1] CCD Primer (http://www.ing.iac.es/~smt/CCD_Primer/CCD_Primer.htm).

[2] Gemini Observatory: exploring the universe from both hemispheres (http://www.gemini.edu).

[3] Large Binocular Telescope Observatory (http://medusa.as.arizona.edu/lbto/).

[4] Main Hubble Page (http://hubble.nasa.gov/).

[5] The James Webb Space Telescope (http://www.jwst.nasa.gov/).

[6] The Very Large Telescope Project (http://www.eso.org/projects/vlt/).

[7] W. M. Keck Observatory (http://www.keckobservatory.org/).

[8] Howard Anton and Chris Rorres. *Elementary Linear Algebra*. John Wiley and Sons, 7th edition, 1994.

[9] H. W. Babcock. The possibility of compensatinhg astronomical seeing. *Publications of the Astronomical Society of the Pacific*, 65:229–236, 1953.

[10] P. A. Bakut, V. E. Kirakosyants, V. A. Loginov, C. J. Solomon, and J. C. Dainty. Optimal wavefront reconstruction from a shack-hartmann sensor by use of a bayesian algorithm. *Opt. Comm.*, 104:10–15, 1994.

[11] Salvador Bara, Susana Rios, and Eva Acosta. Integral evaluation of the modal phase coefficients in curvature sensing: Albrecht's cubatures. *J. Opt. Soc. Am. A*, 13:1467–1474, July 1996.

[12] Jeffrey D. Barchers, David L. Fried, and Donald J. Link. Evaluation of the performance of hartmann sensors in strong scintillation. *Appl. Opt.*, 41(6):1012–1021, February 2002.

[13] Jacques M. Beckers. Interpretation of out-of-focus star images in terms of wave-front curvature. *J. Opt. Soc. Am. A,* 11(1):425–427, 1994.

[14] Richard Berry and James Burnell. *The Handbook of Astronomical Image Processing*. Willmann-Bell, 2000.

[15] Ronald Newbold Bracewell. *Fourier analysis and imaging*. Kluwer Academic/Plenum Publishers, 1st edition, 2003.

[16] Genrui Cao and Xin Yu. Accuracy analysis of a hartmann-shack wavefront sensor operated with a faint object. *Opt. Eng.*, 33(7):2331–2335, July 1994.

[17] Richard M. Clare and Richard G. Lane. Comparison of wavefront sensing with the shack-hartmann and pyramid sensors. *Proceedings of SPIE,* 5490:1211–1222, 2004.

[18] Richard M. Clare and Richard G. Lane. Wavefront sensing from subdivision of the focal plane with a lenslet array. *Appl. Opt.*, 43:4080–4087, 2004.

[19] A. M. Cormack. Representation of a function by its line integrals, with some radiological applications. *Journal of Applied Physics*, 34(9):2722–2727, 1963.

[20] A. M. Cormack. Representation of a function by its line integrals, with some radiological applications. ii. *Journal of Applied Physics*, 35(10):2908–2913, 1964.

[21] Joana B. Costa. Modulation effect of the atmosphere in a pyramid wave-front sensor. *Appl. Opt.*, 44(1):60–66, January 2005.

[22] S. Esposito and A. Riccardi. Pyramid wavefront sensor behaviour in partial correction adaptive optic systems. *A&A.*, 369:L9–L12, 2001.

[23] G. F. Franklin et. al. *Feedback control of dynamic systems*. Prentice-Hall, 4th edition, 2002.

[24] J. R. Fienup. Phase retrieval algorithms: a comparison. *Appl. Opt.*, 21:2758–2769, 1982.

[25] Ralf C. Flicker. Sequence of phase correction in multiconjugate adaptive optics. *Opt. Lett.*, 26(22):1743–1745, 2001.

[26] L. M. Foucault. Mmoire sur la construction des tlescopes en verre argent. *Ann. Obs. Imp. Paris*, 5:197–237, 1859.

[27] R. Foy, A. Migus, F. Biraben, G. Grynberg, P. R. McCullough, and M. Tallon. The polychromatic artificial sodium star: A new concept for correcting the atmospheric tilt. *Astronomy and Astrophysics Supplement Series*, 111:569–578, 1995.

[28] David L. Fried. Anisoplanatism in adaptive optics. *J. Opt. Soc. Am.*, 72(1):52–61, January 1982.

[29] David L. Fried and John F. Belsher. Analysis of fundamental limits to artificial-guide-star adaptive-optics-system performance for astronomical imaging. *J. Opt. Soc. Am. A*, 11(1):277–287, January 1994.

[30] B. Roy Frieden. *Physics from Fisher information: A unification*. Cambridge University Press, New York, 1998.

[31] Christ Ftaclas and Alex Kostinski. Curvature sensors, adaptive optics and neumann boundary conditions. *Appl. Opt.*, 40(4):435–438, 2001.

[32] R. W. Gerchberg and W. O. Saxton. A practical algorithm for the determination of the phase from image and diffraction plane pictures. *Optik*, 35:237–246, 1972.

[33] Andreas Glindemann. Improved performance of adaptive optics in the visible. *J. Opt. Soc. Am. A*, 11(4):1370–1375, 1994.

[34] R. A. Gonsalves. Nonisoplanatic imaging by phase diversity. *Opt. Lett.*, 19(7):493–495, 1993.

[35] Robert A. Gonsalves. Phase retrieval and diversity in adaptive optics. *Opt. Lett.*, 21(5):829–832, 1982.

[36] Robert A. Gonsalves. Small-phase solution to the phase-retrieval problem. *Opt. Lett.*, 26(10):684–685, 2001.

[37] Rafael C. Gonzalez and Richard E. Woods. *Digital Image Processing*. Addison-Wesley Pub, 3 edition, 1992.

[38] J. Goodman. *Introduction to Fourier Optics*. McGraw-Hill, New York, 1996.

[39] T. E. Gureyev and K. A. Nugent. Phase retrieval with the transport-of-intensity equation. ii. orthogonal series solution for nonuniform illumination. *J. Opt. Soc. Am. A*, 13(8):1670–1682, 1996.

[40] T. E. Gureyev, A. Roberts, and K. A. Nugent. Phase retrieval with the transport-of-intensity equation: matrix solution with use of zernike polynomials. *J. Opt. Soc. Am. A*, 12(9):1932–1941, 1995.

[41] Inwoo Han. New method for estimating wavefront from curvature signal by curve fitting. *Opt. Eng.*, 34(4):1232–1237, 1995.

[42] C. M. Harding, R. A. Johnston, and R. G. Lane. Fast simulation of a kolmogorov phase screen. *Appl. Opt.*, 38(11):2161–2170, 1999.

[43] Paul Hickson. Wave-front curvature sensing from a single defocused image. *J. Opt. Soc. Am. A*, 11(5):1667–1673, May 1994.

[44] Hirofumi Horikawa, Naoshi Baba, Masashi Ohtsubo, Yuji Norimoto, Tetsuo Nishimura, and Noriaki Miura. Wind-flow measurement over the subaru telescope. *Appl. Opt.*, 43(15):3097–3102, 2004.

[45] Kazuichi Ichikawa, Adolf W Lohmann, and Mitsuo Takeda. Phase retrieval based on the irradiance transport equation and the fourier transform method: experiments. *Appl. Opt.*, 27(16):3433–3436, 1988.

[46] Ignacio Iglesias, Roberto Ragazzoni, Yves Julien, and Pablo Artal. Extended source pyramid wave-front sensor for the human eye. *Optics Express*, 10(9):419–428, 2002.

[47] Roy Irwan. *Wavefront sensing in Astronomical Imaging*. PhD thesis, University of Canterbury, 1999.

[48] Stuart M. Jefferies, Michael Lloyd-Hart, E. Keith Hege, and James Georges. Sensing wave-front amplitude and phase with phase diversity. *Appl. Opt.*, 41(11):2095–2102, April 2002.

[49] Dustin C. Johnston, Brent L. Ellerbroek, and Stephen M. Pompea. Curvature sensing analysis. *SPIE Proceedings - Adaptive Optics in Astronomy*, pages 528–538, 1994.

[50] Dustin C. Johnston and Byron M. Welsh. Analysis of multiconjugate adaptive optics. *J. Opt. Soc. Am. A*, 11(1):394–408, January 1994.

[51] Steven M. Kay. *Fundamentals of Statistical Signal Processing: estimation theory*. Prentice Hall, 1993.

[52] Richard L. Kendrick, D. S. Acton, and A. L. Duncan. Phase-diversity wave-front sensor for imaging systems. *Appl. Opt.*, 33(27):6533–6546, September 1994.

[53] A. N. Kolmogorov. *(Russian) Dokl. Akad, Nauk SSSR*, 30:229, 1941.

[54] A. N. Kolmogorov. *The local structure of turbulence in incompressible viscous fluids for very large reynolds' numbers*. Wiley-Interscience, New York, 1961.

[55] R. G. Lane. Methods for maximum-likelihood deconvolution. *J. Opt. Soc. Am. A*, 13(10):1992–1998, October 1996.

[56] N. F. Law and R. G. Lane. Wavefront estimation at low light levels. *Opt. Comm.*, 126:19–24, 1996.

[57] David J. Lee, Michael C. Roggemann, and Byron M. Welsh. Cramer-rao analysis of phase-diverse wave-front sensing. *J. Opt. Soc. Am. A*, 16(5):1005–1015, 1999.

[58] Michael Lloyd-Hart and N. Mark Milton. Fundamental limits on isoplanatic correction with multiconjugate adaptive optics. *J. Opt. Soc. Am. A*, 20(10):1949–1957, October 2003.

[59] D. Russell Luke, James V. Burke, and Richard G. Lyon. Optical wavefront reconstruction: Theory and numerical methods. *SIAM Review*, 44(2):169–224, 2002.

[60] R. P. Millane. Phase retrieval in crystallography and optics. *J. Opt. Soc. Am. A*, 7:394–411, 1990.

[61] Mark Milman, David Redding, and Laura Needels. Analysis of curvature sensing for large-aperture adaptive optics systems. *J. Opt. Soc. Am. A*, 13(6):1226–1238, 1996.

[62] Peter W. Milonni. Adaptive optics for astronomy. *American Journal of Physics*, 67(6):476–485, June 1999.

[63] Peter W. Milonni, Robert Q. Fugate, and John M. Telle. Analysis of measured photon returns from sodium beacons. *J. Opt. Soc. Am. A*, 15(1):217–233, 1998.

[64] Isaac Newton. *Opticks or a treatise of the reflexions, refractions, inflexions and colours of light*. 1704.

[65] Robert J. Noll. Zernike polynomials and atmospheric turbulence. *J. Opt. Soc. Am.*, 66(3):207–211, 1976.

[66] Scot S. Olivier and Donald T. Gavel. Tip-tilt compensation for astronomical imaging. *J. Opt. Soc. Am. A*, 11(1):368–378, 1994.

[67] Scot S. Olivier, Claire E. Max, Donald T. Gavel, and James M. Brase. Tip-tilt compensation: Resolution limits for ground-based telescopes using laser guide star adaptive optics. *The Astrophysical Journal*, 407:428–439, 1993.

[68] Athanasios Papoulis. *Systems and Transforms with Applications in Optics*. McGraw-Hill Book Company, 1968.

[69] Ronald R. Parenti and Richard J. Sasiela. Laser-guide-star systems for astronomical applications. *J. Opt. Soc. Am. A*, 11(1):288–309, 1994.

[70] R. G. Paxman and J. R. Fienup. Optical misalignment sensing and image reconstruction using phase diversity. *J. Opt. Soc. Am. A*, 5(6):914–923, June 1988.

[71] Maria Petrou and Pedro Garcia Sevilla. *Image Processing, Dealing with Texture*. John Wiley and Sons, Ltd, 1st edition, 2006.

[72] Ben C. Platt and Roland Shack. History and principles of shack-hartmann wavefront sensing. *Journal of Refractive Surgery*, 17:S573–S577, 2001.

[73] Lisa A. Poyneer and Bruce Macintosh. Spatially filtered wave-front sensor for high-order adaptive optics. *J. Opt. Soc. Am. A*, 21(5):810–819, 2004.

[74] J. Primot, G. Rousset, and J. C. Fontanella. Deconvolution from wavefront sensing: a new technique for compensating turbulence degraded images. *J. Opt. Soc. Am. A*, 9:1598–1608, 1990.

[75] R. Ragazzoni and J. Farinato. Sensitivity of a pyramidic wave front sensor in closed loop adaptive optics. *A&A.*, 350:L23–L26, 1999.

[76] Roberto Ragazzoni. Pupil plane wavefront sensing with an oscillating prism. *Journal of Modern Optics*, 43(2):289–293, 1996.

[77] Roberto Ragazzoni, Emiliano Diolaiti, and Elise Vernet. A pyramid wavefront sensor with no dynamic modulation. *Opt. Comm.*, 208:51–60, 2002.

[78] Roberto Ragazzoni, Enrico Marchetti, and Gianpaolo Valente. Adaptive-optics corrections available for the whole sky. *Nature*, 403:54–56, January 2000.

[79] W.H. Richardson. Bayesian-based iterative method of image restoration. *J. Opt. Soc. Am.*, 62(1):55, 1972.

[80] F. Rigaut and E. Gendron. Laser guide star in adaptive optics: the tilt determination problem. *A&A.*, 261:677–684, 1992.

[81] Francois Rigaut, Brent L. Ellerbroek, and Malcolm J. Northcott. Comparison of curvature-based and shack-hartmann-based adaptive optics for the gemini telescope. *Appl. Opt.*, 36(13):2856–2868, 1997.

[82] Claude Roddier and Francois Roddier. Combined approach to the hubble space telescope wave-front distortion analysis. *Appl. Opt.*, 32(16):2992–3008, June 1993.

[83] Claude Roddier and Francois Roddier. Wave-front reconstruction from defocused images and the testing of ground-based optical telescopes. *J. Opt. Soc. Am. A*, 10(11):2277–2287, 1993.

[84] F. Roddier. Error propagation in a closed-loop adaptive optics system: a comparison between shack-hartmann and curvature wave-front sensors. *Opt. Comm.*, 113:357–359, 1995.

[85] F. Roddier, editor. *Adaptive Optics in Astronomy*. Cambridge University Press, 1999.

[86] Francois Roddier. Curvature sensing and compensation: a new concept in adaptive optics. *Appl. Opt.*, 27(7):1223–1225, 1988.

[87] Francois Roddier. Wavefront sensing and the irradiance transport equation. *Appl. Opt.*, 29(10):1402–1403, 1990.

[88] Francois Roddier, Malcolm Northcott, and J. Elon Graves. A simple low-order adaptive optics system for near-infrared applications. *Publications of the Astronomical Society of the Pacific*, 103:131–149, January 1991.

[89] M. C. Roggemann and B. Welsh. *Imaging through turbulence*. The CRC Press, 1st edition, 1996.

[90] Claude E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27:379–423, 1948.

[91] M. Soto and E. Acosta. Performance analysis of curvature sensors: optimum positioning of the measurement planes. *Optics Express*, 11(20):2577–2588, 2003.

[92] W. H. Southwell. Wave-front analyzer using a maximum likelihood algorithm. *J. Opt. Soc. Am.*, 67(3):396–399, 1976.

[93] W. H. Southwell. Wave-front estimation from wave-front slope measurements. *J. Opt. Soc. Am.*, 70(8):998–1006, August 1980.

[94] Jakob J. Stamnes. *Waves in Focal Regions*. IOP Publishing Limited, 1986.

[95] N. Streibl. Phase imaging by the transport equation of intensity. *Opt. Comm.*, 49(1):6–9, 1984.

[96] Michael Reed Teague. Irradiance moments: their propagation and use for unique retrieval of phase. *J. Opt. Soc. Am.*, 72(9):1199–1209, 1982.

[97] Michael Reed Teague. Deterministic phase retrieval: a green's function solution. *J. Opt. Soc. Am.*, 73(11):1434–1441, 1983.

[98] Michael Reed Teague. Image formation in terms of the transport equation. *J. Opt. Soc. Am. A*, 2(11):2019–2026, 1985.

[99] Stephen F. Tonkin. *Practical Amateur Spectroscopy (Patrick Moore's Practical Astronomy)*. Springer-Verlag London Ltd, 2002.

[100] Glenn A. Tyler and David L. Fried. Image-position error associated with a quadrant detector. *J. Opt. Soc. Am. A*, 72(6):804–808, June 1982.

[101] Robert K. Tyson. *Introduction to Adaptive Optics*. The International Society for Optical Engineering, 2000.

[102] M. A. van Dam and R. G. Lane. Direct wavefront sensing using geometric optics. In *High Resolution Wavefront Control: Methods, Devices and Applications IV, Proceedings of SPIE*, volume 4825, 2002.

[103] M. A. van Dam and R. G. Lane. Extended analysis of curvature sensing. *J. Opt. Soc. Am. A*, 19(7):1390–1397, July 2002.

[104] M. A. van Dam and R. G. Lane. Tip/tilt estimation from defocused images. *J. Opt. Soc. Am. A*, 19(4):745–752, Apr 2002.

[105] M. A. van Dam and R. G. Lane. Wave-front sensing from defocused images using wave-front slopes. *Appl. Opt.*, 41(26):5497–5502, September 2002.

[106] Christophe Verinaud. On the nature of the measurements provided by a pyramid wave-front sensor. *Opt. Comm.*, 233:27–38, 2004.

[107] Edward P. Wallner. Optimal wave-front correction using slope measurements. *J. Opt. Soc. Am.*, 73(12):1771–1776, 1983.

[108] Brian D. Warner. *A practical guide to lightcurve photometry and analysis*. Springer Science+Business Media Inc, 1st edition, 2006.

[109] Byron M. Welsh, Brent L. Ellerbroek, Michael C. Roggemann, and Timothy L. Pennington. Fundamental performance comparison of a hartmann and a shearing interferometer wave-front sensor. *Appl. Opt.*, 34(21):4186–4195, 1995.

[110] Byron M. Welsh and Chester S. Gardner. Performance analysis of adaptive-optics systems using laser guide stars and slope sensors. *J. Opt. Soc. Am. A*, 6(12):1913–1923, 1989.

[111] R. Gale Wilson. Wavefront-error evaluation by mathematical analysis of experimental foucault-test data. *Appl. Opt.*, 14(9):2286–2297, 1975.

[112] Kim A. Winick. Cramer-rao lower bounds on the performance of charge-coupled-device optical position estimators. *J. Opt. Soc. Am. A*, 3(11):1809–1815, November 1986.

[113] Carl Witthoft. Wavefront sensor noise reduction and dynamic range expansion by means of optical image intensification. *Opt. Eng.*, 29(10):1233–1238, October 1990.

[114] Simon C. Woods and Alan H. Greenaway. Wave-front sensing by use of a green's function solution to the intensity transport equation. *J. Opt. Soc. Am. A*, 20(3):508–512, 2003.

[115] H. T. Yura. Short-term average optical-beam spread in a turbulent medium. *J. Opt. Soc. Am.*, 63(5):567–572, 1973.

[116] Fabio E. Zocchi. A simple analytical model of adaptive optics for direct detection free-space optical communication. *Opt. Comm.*, 248:359–374, 2005.