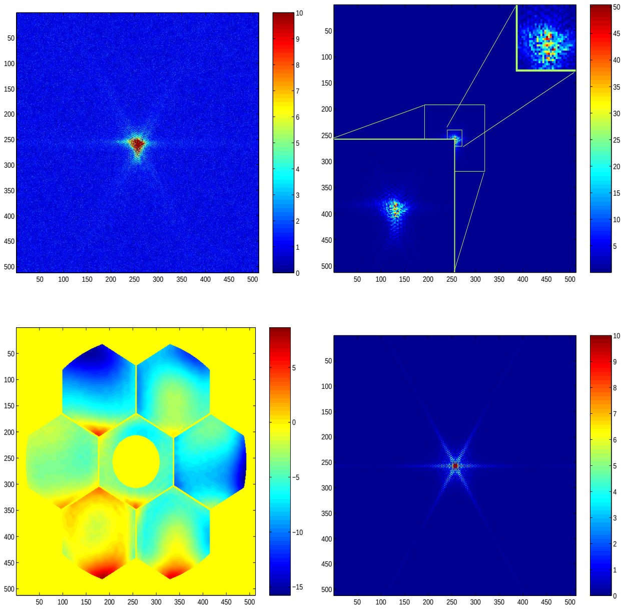
Analysis of Optical Wavefront Reconstruction and Deconvolution in Adaptive Optics

David Russell Luke



A dissertation
submitted in partial fulfillment of
the requirements for the degree of

Doctor of Philosophy

University of Washington

2001

Program Authorized to Offer Degree: Applied Mathematics

University of Washington
Graduate School

This is to certify that I have examined this copy of a doctoral dissertation by

David Russell Luke

and have found that it is complete and satisfactory in all respects,
and that any and all revisions required by the final
examining committee have been made.

Chair of Supervisory Committee:

James V. Burke

Reading Committee:

Professor R. Tyrell Rockafellar

Professor Kenneth Bube

Date:

In presenting this dissertation in partial fulfillment of the requirements for the Doctoral degree at the University of Washington, I agree that the Library shall make its copies freely available for inspection. I further agree that extensive copying of this dissertation is allowable only for scholarly purposes, consistent with "fair use" as prescribed in the U.S. Copyright Law. Requests for copying or reproduction of this dissertation may be referred to Bell and Howell Information and Learning, 300 North Zeeb Road, Ann Arbor, MI 48106-1346, or to the author.

Signature______________________________

Date________________________________

University of Washington

Abstract

Analysis of Optical Wavefront Reconstruction and Deconvolution in Adaptive Optics

by David Russell Luke

Chair of Supervisory Committee:

Professor James V. Burke
Department of Mathematics

It was in the spirit of "reuniting divergent trends by clarifying the common features and interconnections of many distinct and diverse scientific facts" that Courant and Hilbert published their book *Methods of Mathematical Physics* [51]. This thesis is written in the same spirit and with the same goal. We focus our attention on the problem of wavefront reconstruction and deconvolution in adaptive optics. This is an ill-posed, non-linear inverse problem that touches on the theory of harmonic analysis, variational analysis, signal processing, nonconvex optimization, regularization, statistics and probability. Numerical solutions rely on spectral and operator-splitting methods as well as limited memory and multi-resolution techniques. We introduce novel methods for wavefront reconstruction and compare our results against common techniques. Previous work on this problem is reviewed and unified in a non-parametric, analytic framework.

©2001
David Russell Luke

TABLE OF CONTENTS

List of Figures iii

List of Tables vi

Chapter 1: Introduction 1
- 1.1 Literature Review . . . 3

Chapter 2: Optical Imaging 7
- 2.1 The Forward Imaging Model . . . 7
- 2.2 Inverse Problems . . . 20

Chapter 3: Wavefront Reconstruction 28
- 3.1 Geometric Approaches . . . 28
- 3.2 Nonsmooth Analysis . . . 42
- 3.3 Optimization of Smooth Perturbations . . . 67

Chapter 4: Phase Diversity 87
- 4.1 Least Squares Regularization . . . 90
- 4.2 A Statistical Perspective: the Wiener filter . . . 92
- 4.3 Regularization via Parameterization . . . 96

Chapter 5: Numerical Methods 98
- 5.1 Line Search . . . 98
- 5.2 Acceleration Techniques: Limited Memory BFGS with Trust Regions . . . 100
- 5.3 Multi-resolution Analysis . . . 109

Chapter 6: Numerical Results 113
- 6.1 Phase retrieval . . . 113
- 6.2 Multi-resolution Techniques . . . 121
- 6.3 Smooth Objectives . . . 126
- 6.4 Simultaneous Deconvolution and Wavefront Reconstruction . . . 126

Chapter 7: Conclusion 133
- 7.1 Summary . . . 133
- 7.2 Extensions and Future Work . . . 134

Bibliography 135

Appendix A: Convolution and Autocorrelation properties **152**

LIST OF FIGURES

2.1 Model optical system 8

2.2 Rayleigh-Sommerfeld diffraction 10

2.3 Lens model 13

3.1 One dimensional pupil with corresponding image data. Frame (a) is a one dimensional cross section of the amplitude across the aperture shown in Fig.(6.1.a). Frames (b)-(d) are cross sections of the corresponding point-spread functions for the aperture in frame (a) with some unknown phase aberration as well as a known defocus. 29

3.2 Tube constraints. The vertical axis and the axis coming out of the page correspond to the real and imaginary components of the tubes. The horizontal axes correspond to the horizontal axes of Fig.(3.1). Frame (a) represents the constraint set corresponding to Fig.(3.1.a). Frames (b)-(d) represent the constraint sets corresponding to Fig.(3.1.b)-(3.1.d). 30

3.3 Frames (a)-(b) show the behavior of the sequential algorithm in a triangular trap. Frames (c)-(d) show the behavior of the averaged projections algorithm in the same triangular trap. The sum distance error shown in (b) and (d) for the respective algorithms is the performance measure given by equation (3.20). 40

4.1 Comparison of modulus kernel to modulus-squared kernel. The contours are at the same levels for each figure and show that squared-modulus objective has less structure near zero than the modulus objective 89

6.1 Aperture (a) and noiseless image data (b)-(d) for a segmented pupil on a 512 by 512 grid. The 3 diversity images are the optical system's response to a point source at focus, and plus/minus defocus respectively. 114

6.2 Real and imaginary parts, (a) and (b) respectively, of an aberrated wavefront for the segmented pupil recovered from 3 noiseless diversity point source images on a 512 by 512 grid. The wavefront phase is unwrapped (c) and compared to the true phase. The wavefront error (d) is in units of wavelength.115

6.3 Noisy point-spread function (a) for a segmented pupil on a 512 by 512 grid. The recovered point-spread function (b) was first filtered with a Fourier window filter before processing by the wavefront reconstruction algorithm. Frame (c) shows the true, unaberrated point-spread function. 116

6.4 The real and imaginary parts, (a) and (b) respectively, of the aberrated wavefront for the segmented pupil recovered from 3 filtered noisy diversity point source images on a 512 by 512 grid. The wavefront phase (c) is unwrapped and compared to the true phase. The wavefront error (d) is in units of wavelength. The ridges in the wavefront error is due to Gibbs phenomenon associated with the noise filter. 117

6.5 Comparison of algorithms applied to the test problem shown in Fig.(6.1). The algorithms have different objectives, so we compare the behavior of the squared set distance error defined by Eq.(3.31) for each. 118

6.6 Multi-resolution image data. Three levels of multi-resolution windowing operations are depitced for each diversity image. Frame (a) shows close-ups of each of the three resolution levels. The center 32×32 pixels of each diversity image, together with the corresponding low-resolution pupil constraint Fig.(6.7.b) are used to generate the approximate solution shown in Fig.(6.8.a)-(6.8.b). This solution is used to initialize the same problem with the center 128×128 pixels and the correpsonding low-resolution pupil constraint (Fig.(6.7c)) as data. The solution to this problem, shown in Fig.(6.8.c)-(6.8.d), is used to initialize the full resolution problem. The progress of the set distance error versus iteration of the multi-resolution implementation is shown in Fig.(6.9). 122

6.7 Multi-Resolution pupil domain constraints. The lowest resolution pupil constraint (a) corresponds to the 32×32 image data shown in Fig.(6.6). The medium resolution pupil constraint (c) corresponds to the 128×128 image data. 123

6.8 Aberrated wavefront for the segmented pupil recovered from 3 diversity point source images on successively finer grids. The real and imaginary parts of the low resolution wavefront, (a) and (b) respectively, is generated from a truncation of the image data to the center 32 by 32 pixels. This solution is used as a first guess for the next resolution, 128 by 128. The real and imaginary parts of the 128 by 128 pixel resolution, (c) and (d) respectively, is used as a first guess for the full resolution problem shown in Fig.(6.2). . . . 124

6.9 Squared set distance error and corresponding norm of the gradient versus iteration for a multi-resolution implementation of the trust region L-BFGS algorithm Alg.5.2.2. The flat region of the iterations is handled at low resolutions. Only when the estimate is in the neighborhood of a solution does the algorithm switch to higher resolution calculations. 125

6.10 Comparison of performance of a numerical optimization algorithm for solving Pr.(2.80) with $\varphi = \delta$ and the objective E_2 defined by Eq.(4.3) versus the objective E_ϵ defined by Eq.(3.87). The behavior of the squared set distance error E is calculated and plotted for the algorithm with both objectives for comparison. Frames (a)-(d) are at 32×32, 64×64, 256×256, and 512×512 resolution respectively. 127

6.11 Noiseless image data. Three images of the same object taken at 3 different focus settings. 128
6.12 Recovered object (a) and wavefront (c) from noiseless data shown in Fig.(6.11). The true object is shown in frame (b). The true phase (d) is compared to the recovered phase. The error (e) is in units of wavelength. 129
6.13 Noisy image data. Three images of the same object taken at 3 different focus settings. 131
6.14 Recovered object (a) and wavefront (c) from noiseless data shown in Fig.(6.13). The true object is shown in frame (b). The true phase (d) is compared to the recovered phase. The error (e) is in units of wavelength. 132

LIST OF TABLES

6.1 Relative cpu time of projection and analytic algorithms averaged over 30 random trials. The baseline is the LS algorithm. Outliers were not included in the totals for algorithms with an asterisk. 121

ACKNOWLEDGMENTS

"If God is omnipotent, can (S)he create a rock that (S)he cannot lift?"[1]

For their love of life, tenacity of spirit and dedication to each other, I thank my wonderful family, Nick&Nina, Doug&Sarah, Lindsay, Haven&Natalie. Special thanks go to Anja for her tolerance, empathy, support, comfort and love. I would like to thank NASA's Goddard Space Flight Center and the Graduate Student Researchers Program which provided valuable support and technical resources during the course of this work. I am particularly grateful to Rick Lyon for introducing me to the phase retrieval problem and for making his expertise and data so freely available. Tim Murphy deserves all the credit for helping me make the quantum leap from Matlab to "real code". Keshanie and K.K. deserve a great deal of credit for helping me to afford my jet-set lifestyle and for making sure I didn't starve. I am grateful for their help with "the system". I miss my friends and comrades, Arnold, Darryl, Charla&Brian, Yuki, Pat, the Hammond family, Karl, Tom, Don, and Tricia. They have made life in Seattle a pleasure and shared with me the joy of good food. To the scientists and mathematicians who have had the greatest influence on my graduate education, Peter Schmid, Kenneth Bube, Terry Rockafellar, and James Burke, I am eternally grateful. Finally, I would like to thank my advisor, my mentor, and my teacher, James Burke, for the many hours he has spent with me on this work. His sharp eye and dedication have made me a better scientist and mathematician.

---

[1] No.

To Anja, my love

Chapter 1

INTRODUCTION

> *"Since the seventeenth century, physical intuition has served as a vital source for mathematical problems and methods. Recent trends and fashions have, however, weakened the connection between mathematics and physics... This rift is unquestionably a serious threat to science as a whole; the broad stream of scientific development may split into smaller and smaller rivulets and dry out."*
>
> -R. Courant *Methods of Mathematical Physics* [51]

The history of science is filled with misfortunes that have been transformed into scientific triumphs. This thesis is concerned with the analysis of numerical methods for wavefront reconstruction and deconvolution that contributed to the eventual and remarkable successes of NASA's Hubble Space Telescope (HST). Shortly after launch on April 24, 1990, it was discovered that the primary mirror of the HST suffered from a large spherical aberration [35]. Several teams of researchers were dispatched to apply a variety of image processing techniques to the flight data of stellar images in order to identify the aberration and aid in the design of corrective optics. Burrows [34] and Lyon et al. [118] applied parametric techniques; Fienup [64] applied gradient-based algorithms; Fienup et al. [70] and Roddier [155] applied nonparametric projection techniques; Redding *et al* [148] and Meinel et al [121] applied ray tracing and diffraction propagation techniques and Barrett and Sandler [16] applied neural network techniques. The results of all groups were used in conjunction with archival HST manufacturing records to pinpoint the size and source of the error. It wasn't until 1993 that corrective optics were installed. In the meantime, researchers continued with efforts to model the telescope with enough precision to recover unaberrated images through post-processing. In addition to the gross manufacturing errors, researchers were able to identify aberrations due to the polish marks on the primary and secondary mirrors. Again, wavefront reconstruction techniques played an important role in this effort [105]. During this time much was learned about reconstruction algorithms. An important lesson learned from the HST is that relatively simple software can aid and in some cases replace complicated and sensitive optical systems. In 2012 the replacement for the HST, the Next Generation Space Telescope (NGST), will be folded into the nose of a rocket and launched into geosynchronous orbit, far beyond the reach of astronauts. Wavefront reconstruction algorithms will play a central role in maintaining alignment on the NGST [120, 163].

In 2012 the replacement for the HST, the Next Generation Space Telescope (NGST), will be folded into the nose of a rocket and launched into geosynchronous orbit, far beyond the reach of astronauts. Once in orbit, the NGST will unfold a 10 meter-class segmented aperture that will have a resolution several times that of the Hubble. Wavefront reconstruction algorithms will play a central role in maintaining alignment on the NGST [120, 163]. The challenge with such systems rests in compensating for imperfections that can arise due to, among other things, manufacturing error, launch-related alignment shifts, deployment errors, and thermal deformations. *Adaptive optics* systems allow the operator of the instrument to change the configuration in order to compensate for unforeseen imperfections. The simplest example of an adaptive optics control system is the focus on a camera. In this adaptive optics system the operator moves the lens forward and backward in smaller and smaller intervals until the system is identified (that is, the object is in focus) and the data is collected (usually with the push of a button). This simple procedure becomes more complicated when the camera is orbiting the earth, and the operator has no idea what the object she is looking at is supposed to look like. The problem can be understood by the following: you are wearing someone else's glasses and looking at text written in a foreign alphabet - what is the prescription of the glasses and what is the true text?

Strategies for the control problem involve both open and closed-loop control. In open-loop control the system controls are set independent of simultaneous effect on the output. In closed-loop control the controls depend continuously on the output. Closed-loop adaptive optics control systems have been proven in land-based astronomy to compensate for atmospheric turbulence [62]. Most control systems employ dedicated wavefront sensors, such as the Shack-Hartman sensor, to measure the state of the optical system. These sensors are expensive and as prone to error as the imaging device itself. For space applications computational approaches to wavefront sensing have been proposed [116, 146].

Optical wavefront reconstruction is an inverse problem that arises in many applications in physics and engineering. Numerical algorithms for solving this problem have been employed in crystallography, microscopy, optical design and adaptive optics for three decades. The history of the problem goes back much further. The celebrated algorithm of Gerchberg and Saxton [73] demonstrated that practical *numerical* solutions to the two dimensional problems was possible. Since the introduction of the Gerchberg-Saxton algorithm numerous variations have been studied [13,29,39,50,58,68,70,74,110,117,125,131,135,145,172,177,198]. The spectacular success of these algorithms on the HST together with techniques for *simultaneous* phase retrieval and deconvolution developed for use with land-based astronomical observations [37, 75, 101, 115, 139, 140, 144, 167, 184, 186, 187] has lead to the development of software that, in conjunction with simple optical systems, can achieve the same resolution as complicated, expensive and error-prone optical systems [108, 112, 114, 116, 146, 147]. These new computational approaches offer some hope of closing the adaptive optics control loop for space-based astronomy. Computational wavefront sensing enjoys the advantage that dedicated wavefront sensors are not required. The hardware requirements for such systems are relatively simple and have fewer sources for error. Expensive and sensitive hardware is traded for expensive and sensitive software. Software can be modified and improved continuously on the ground. Once hardware is put into orbit it is very difficult and expensive to modify, as the Hubble Space Telescope demonstrated.

1.1 Literature Review

The deconvolution problem involves solving a linear Fredholm integral equation of the first kind where the kernel of the integral operator is compact. This problem has a long history in applied mathematics and appears in many different fields and applications [102, 104, 180]. We review the basic numerical theory in Chapter 4. The problem of simultaneous deconvolution and wavefront reconstruction is equivalent to system identification and image reconstruction. As such, it shares many features of Kalman filtering [98]. We direct most of our attention to studying the hardest part and precursor to the simultaneous problem, wavefront reconstruction.

The problem of wavefront reconstruction is a special case of the more general inverse problem of *phase retrieval.* The phase retrieval problem arises in diverse fields such as microscopy [59, 72, 90, 122, 182, 183], holography [67, 176], crystallography [124], neutron radiography [4], optical design [63], adaptive optics and astronomy. Earlier reviews of the phase problem can be found in Ref. [91, 171]. Ref. [124] is an excellent review of the phase problem in X-ray crystallography. The physical setting is discussed in some detail in the following section. The abstract problem is stated as follows: given $a : \mathbb{R}^2 \to \mathbb{R}_+$ and $b : \mathbb{R}^2 \to \mathbb{R}_+$, find $u : \mathbb{R}^2 \to \mathbb{C}$ satisfying $|u| = a$ and $|u^\wedge| = b$. Here $\mathbb{R}_+$ denotes the positive orthant, $\cdot^\wedge$ denotes the Fourier transform, and the modulus is the *pointwise Euclidean* magnitude. Simply stated, the problem is to find the phase of a complex-valued function given its pointwise amplitude and the pointwise amplitude of its Fourier transform, hence the name *phase retrieval.*

Until the 1970's the problem of phase retrieval was thought to be hopeless for a number of reasons. In a letter to A. A. Michelson, Lord Rayleigh stated that the continuous phase retrieval problem in interferometry was in general not possible without *a priori* information on the symmetry of the data [174]. In one dimension it was shown that the discrete problem has a multitude of solutions. Indeed, for a signal that is represented by n terms of the Fourier series expansion there are as many as 2^{n-1} possible solutions to the problem [2,3]. Wolf was among the first to suggest that these obstacles might not be insurmountable [196]. Kano and Wolf [100] followed this claim with a successful analytic reconstruction in a physically nontrivial setting. Their reconstruction was not numerical in nature but depended, rather, on the analytic properties of the continuous Fourier transform. Further efforts were made to broaden the applicability of these results [158]. A the same time Walther and O'Niell provided some hope for the possibility of meaningful solutions in the discrete case, and in some relevant cases uniqueness [136, 192]. Dialetis and Wolf later pointed out, however, that the applicability of the theory for the continuous case was limited [56]. Nevertheless, a number of researchers proposed the addition of constraints to narrow the number of potential solutions for the one dimensional problem [59, 87, 90, 137, 145, 182, 183, 197].

As early as 1972 a practical algorithm was proposed for numerical solutions to the seemingly more difficult two-dimensional problem. In their famous paper, Gerchberg and Saxton [73], independent of previous mathematical results for projections onto convex sets, proposed a simple algorithm for solving phase retrieval problems in two dimensions. In [110] the algorithm was recognized as a projection algorithm. Projection algorithms in convex settings have been well understood since the early 1960's [30, 78, 82, 173, 189, 200, 202, 203].

The phase retrieval problem, however, involves *nonconvex* sets. For this reason, the convergence properties of the Gerchberg-Saxton algorithm and its variants are not completely understood.

In the majority of relevant cases the numerical experience demonstrated that projection-type algorithms converged to correct solutions [65, 66]. It was suggested in [31] that this seeming robustness of numerical methods is due to the factorability (or lack thereof) of related polynomials. Indeed, the solution to the two dimensional phase retrieval problem for a discrete signal that can be represented by a finite Fourier series expansion, that is for a *band-limited* image, if it exists, is almost always unique up to rotations by 180 degrees, linear shifts and multiplication by a unit magnitude complex constant. The proof and details of this result can be found in [85]. While this result is of fundamental importance, it does not apply to many of the algorithms used for phase retrieval, in particular in the presence of noise. Thus, while the uniqueness result above remains valid for band-limited signals, it says nothing about the uniqueness of *approximate* solutions in the event that a true solution does not exist, that is when the feasible set is empty. In the convex setting, when the constraint sets onto which the projections are computed do not intersect, convergence of projection algorithms is an open question [19, 20, 22, 49, 78, 201]. Much less is known about the nonconvex setting where many applications lie [39, 46, 47, 88, 164].

In 1982 Fienup [68] generalized the Gerchberg-Saxton algorithm and analyzed many of its properties, showing, in particular, that the directions of the projections in the generalized Gerchberg-Saxton algorithm are formally similar to directions of steepest descent for a squared set distance metric. We show in Section 3.2.1 that this connection to directions of steepest descent is complicated by the fact that the metric is not everywhere differentiable. In 1985 Barakat and Newsam [14, 15] developed an approach similar to the gradient descent analogy suggested in [68]. They modeled their analysis on the projection theory for convex sets. A well known fact from convex analysis is that the gradient of the squared distance to a convex set is *equivalent* to the direction toward the projection onto the set. To extend this property to the nonconvex sets, Barakat and Newsam require the projection operators to be single-valued, however there is no known example of a nonconvex set for which the projection operator is single-valued [1]. Indeed, we show that the projections in the case of phase retrieval are multi-valued. We show precisely how the multi-valuedness of the projections is related to the nonsmoothness of the squared set distance metric.

In Section 3.3 a smooth error metric is proposed and bounds are derived for the distance between the gradient of the smooth metric and the directions toward the projections. While projection methods often work well in practice, fundamental mathematical questions concerning their convergence remain unresolved. What are often referred to as convergence results for projection algorithms are statements that the error between iterations will not increase [73, 110]. In general, projection algorithms may not converge to the intersection of nonconvex sets. See Ref. [110] and Ref. [50] for a discussion.

We present algorithmic approaches to phase retrieval and simultaneous deconvolution/wavefront reconstruction in a unified analytic framework. For ease of discussion, the problem of wave-

---

[1] The issue of nonuniqueness of the projection operator is not to be confused with the uniqueness of the phase problem. The results of [85] are not effected by the multi-valuedness of the projection operators.

front reconstruction and deconvolution is formulated in the continuum. Results for the discrete case follow easily from these results. The similarity between iterative transform algorithms and line search methods applied to a particular error metric has been known for some time [15, 68, 73]. A precise analysis of this correspondence, however, has proven elusive. The source of the difficulty is the nonconvexity of the underlying sets and the nonsmoothness of the error metric. In this work we detail the connection between geometric and analytic methods for the phase retrieval problem and extend these results to the more general problem of simultaneous deconvolution and phase retrieval. Chapter 2 details the mathematical model for diffraction imaging. In the same chapter the abstract optimization problem associated with wavefront reconstruction is formulated. Chapter 3 is a detailed study of the phase retrieval problem. We study the geometric and analytic properties of the distance function which is at the heart of this problem. A perturbation of the distance function is detailed and the corresponding least squares optimization problem is formulated. The least squares measure is extended to allow adaptive weighting of the errors between measurements. In Chapter 4 the problem of simultaneous wavefront reconstruction and deconvolution is studied. We detail the connection between Tikhonov regularization techniques and optimal filtering for noisy data. In Chapter 5 we outline numerical algorithms including simple line search, and limited memory techniques with trust regions. Basic convergence results are proven. In the same chapter we detail multi-resolution techniques to reduce computational intensity. Numerical results are detailed in Chapter 6.

Chapter 2

OPTICAL IMAGING

2.1 The Forward Imaging Model

The physical setting we consider here is that of a *monochromatic, time harmonic electromagnetic field in a homogeneous, isotropic medium with no charges or currents.* This is depicted as a wave propagating away from some source to the left of the *pupil plane* in Fig.(2.1). By Maxwell's equations, at a given frequency $\omega \in \mathbb{R}_+$, the spatial components of the electric and magnetic fields can be represented as the real part of complex-valued functions $U_\omega : \mathbb{R}^3 \to \mathbb{C}$ satisfying the Helmholtz equation describing the spatial distribution of energy in an expanding wave:

$$(\triangle + k^2 n^2) U_\omega(\boldsymbol{x}) = 0. \tag{2.1}$$

Here $\triangle$ denotes the Laplacian, $n \in \mathbb{R}_+$ is the index of refraction of the medium, and $k \in \mathbb{R}_+$ is the wave number. The wave number is related to the frequency since ω/k is the speed of light. Another quantity that arises is the *wavelength* λ defined as $\lambda = 2\pi/k$. For convenience, let $n = 1$. In all that follows the fields are assumed to be monochromatic (i.e. single frequency), thus we drop the ω subscript from U_ω.

The wave in Fig.(2.1) passes through an *optical system* consisting of apertures, aberrating media such as mirrors and crystal structures, and a focusing lens. The focused wave is imaged onto an array of receptors that measure intensity. The plane in which the receptors lie is referred to as the *image plane.* The *pupil* of the optical system is an abstract designation for intervening media - atmosphere, mirror surfaces, crystal structures, etc. - through which the electromagnetic wave travels before it is finally refocused and projected onto the image plane. The *entrance pupil* is the aperture through which the unaberrated, or *reference* wave enters the optical system. The *exit pupil* is the aperture through which the aberrated wave exits the optical system. In the mathematical model of the optical system, the entrance pupil and exit pupil are collapsed into a single plane with all aberrating effects occurring at what is refered to as the *pupil plane.* The intensity mapping resulting from a point source is the *point-spread function* for the optical system. The electromagnetic field may be written in phasor notation as $U = f \exp[\sqrt{-1}\,\theta]$ where f and θ are real-valued functions. The *phase retrieval* problem involves recovering the phase, θ, of an electromagnetic field in the exit pupil from intensity measurements in the image plane when the source is a point source.

We begin our discussion by building the mathematical model of the optical system and image formation starting with a brief discussion of the fundamentals of diffraction. Diffraction theory models the propagation of a field through a small aperture. The resulting model represents the field on the image plane as an integral operator of the value of the field across

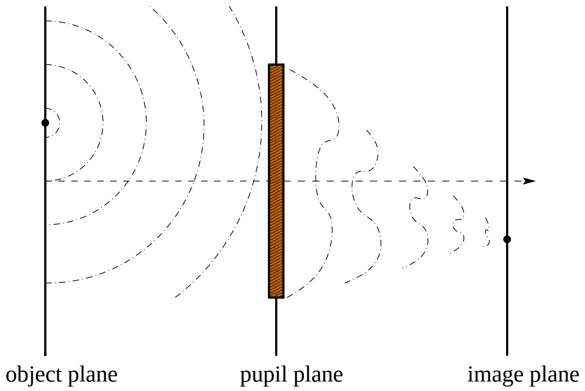


Figure 2.1: Model optical system

the aperture. This is a mathematical formalization of Huygens' Principle, i.e.

> "*'light falling on the aperture* [$\mathbb{A}$] *propagates as if every [surface] element* [dS] *emitted a spherical wave the amplitude and phase of which are given by that of the incident wave* [U]*"* [170].

Boundary conditions at the aperture (*Kirchhoff boundary conditions*) and at infinity (*radiation conditions*) yield approximations to the kernel of the integral operator on the aperture. Two such approximations are derived, the *Fresnel* kernel and the *Fraunhofer* kernel. The Fraunhofer kernel links diffraction theory to the Fourier transform. After deriving this model, we then develop its consequences for fields resulting from a point source, that is, an explicit representation of the point-spread function of the optical system is derived.

2.1.1 Rayleigh-Sommerfeld Diffraction

We now give a terse summary of Rayleigh-Sommerfeld diffraction theory. More detailed developments can be found in [26, 76, 170]. Let Ω be a closed volume in $\mathbb{R}^3$ whose boundary is the orientable closed surface $\mathbb{S}$ and let $\vec{n}$ denote the unit *inward* normal to Ω. Let U, and $\tilde{U}$ be twice continuously differentiable scalar fields mapping Ω and $\mathbb{S}$. By Green's Theorem[1]

$$-\int_{\mathbb{S}} \tilde{U}\frac{\partial U}{\partial \vec{n}} - U\frac{\partial \tilde{U}}{\partial \vec{n}} dS = \int_{\Omega} \tilde{U}\triangle U - U\triangle\tilde{U} dV$$

where $\frac{\partial}{\partial \vec{n}}$ denotes the derivative in the direction of the unit inward normal at $\mathbb{S}$. If both U and $\tilde{U}$ satisfy the Helmholtz equation Eq.(2.1), then

$$-\int_{\mathbb{S}} \tilde{U}\frac{\partial U}{\partial \vec{n}} - U\frac{\partial \tilde{U}}{\partial \vec{n}} dS = 0 \tag{2.2}$$

---

[1] Green's Theorem is usually stated in terms of the unit *outward* normal. In optics, for the derivation of Rayleigh-Sommerfeld diffraction the unit inward normal is usually used.

Let $\mathbb{B}_\epsilon$ denote the Euclidean ball of radius ϵ in $\mathbb{R}^3$ having surface $\mathbb{S}_\epsilon$, and let $\mathbb{B}_\epsilon(\boldsymbol{\xi})$ be the Euclidean ball of radius ϵ centered at $\boldsymbol{\xi}$. Given $\boldsymbol{\xi} \in \operatorname{int}(\Omega)$, choose $\epsilon > 0$ so that $\mathbb{B}_\epsilon(\boldsymbol{\xi}) \subset \operatorname{int}(\Omega)$ and set $\Omega_\epsilon = \Omega \backslash \mathbb{B}_\epsilon(\boldsymbol{\xi})$. Consider the Green's function

$$G_0(\boldsymbol{x};\boldsymbol{\xi}) = \frac{\exp(\sqrt{-1}\,k|\boldsymbol{x}-\boldsymbol{\xi}|)}{|\boldsymbol{x}-\boldsymbol{\xi}|}, \quad \boldsymbol{x} \neq \boldsymbol{\xi} \tag{2.3}$$

where $|\cdot|$ denotes the standard Euclidean norm. The function G_0 is a unit-amplitude spherical wave centered at $\boldsymbol{\xi}$. On Ω_ϵ the scalar field G_0 satisfies the Helmholtz equation

$$(\Delta + k^2)G_0(\boldsymbol{x};\boldsymbol{\xi}) = 4\pi\delta(\boldsymbol{x}-\boldsymbol{\xi}).$$

Thus, as in Eq.(2.2),

$$-\int_{\mathbb{S}+\mathbb{S}_\epsilon} \frac{\exp(\sqrt{-1}\,k|\boldsymbol{x}-\boldsymbol{\xi}|)}{|\boldsymbol{x}-\boldsymbol{\xi}|}\frac{\partial U}{\partial \vec{n}} - U\frac{\partial}{\partial \vec{n}}\frac{\exp(\sqrt{-1}\,k|\boldsymbol{x}-\boldsymbol{\xi}|)}{|\boldsymbol{x}-\boldsymbol{\xi}|}dS = 0. \tag{2.4}$$

The Integral Theorem of Helmholtz and Kirchhoff [26, 170] uses Eq.(2.4) to establish the identity

$$\begin{aligned} U(\boldsymbol{\xi}) &= \lim_{\epsilon\to 0}\frac{-1}{4\pi}\int_{\mathbb{S}_\epsilon} \frac{\exp(\sqrt{-1}\,k|\boldsymbol{x}-\boldsymbol{\xi}|)}{|\boldsymbol{x}-\boldsymbol{\xi}|}\frac{\partial U}{\partial \vec{n}} - U\frac{\partial}{\partial \vec{n}}\frac{\exp(\sqrt{-1}\,k|\boldsymbol{x}-\boldsymbol{\xi}|)}{|\boldsymbol{x}-\boldsymbol{\xi}|}dS \\ &= \frac{1}{4\pi}\int_{\mathbb{S}} \frac{\exp(\sqrt{-1}\,k|\boldsymbol{x}-\boldsymbol{\xi}|)}{|\boldsymbol{x}-\boldsymbol{\xi}|}\frac{\partial U}{\partial \vec{n}} - U\frac{\partial}{\partial \vec{n}}\frac{\exp(\sqrt{-1}\,k|\boldsymbol{x}-\boldsymbol{\xi}|)}{|\boldsymbol{x}-\boldsymbol{\xi}|}dS. \end{aligned} \tag{2.5}$$

Thus, the field at any point $\boldsymbol{\xi}$ can be expressed in terms of the boundary values of the wave on any orientable closed surface surrounding that point.

Rayleigh-Sommerfeld diffraction theory is derived by considering a specific volume Ω and surface $\mathbb{S}$ (see Fig.2.2) together with a particular Green's function G. Let the surface $\mathbb{S}$ be the arbitrarily large half-sphere composed of the hemisphere $\mathbb{E}$ and the disk $\mathbb{D}$ contained in the plane $\mathbb{T}$. The disk $\mathbb{D}$ consists of an annulus $\mathbb{A}'$ with a small opening $\mathbb{A}$. Let $\boldsymbol{x}'$ be an element of the open half-space determined by the plane $\mathbb{T}$ and having empty intersection with Ω. Let $\mathbb{I} \subset \Omega$ be a screen parallel to $\mathbb{T}$ and whose distance from $\mathbb{T}$ equals that of $\boldsymbol{x}'$ to $\mathbb{T}$. The problem is to determine the field U on $\mathbb{I}$ under the assumption that the field propagates only through $\mathbb{A}$.

Consider the field G due to the two mirror point sources, $\boldsymbol{\xi} \in \mathbb{I}$ and $\boldsymbol{x}'$:

$$G(\boldsymbol{x};\boldsymbol{x}',\boldsymbol{\xi}) \equiv G_0(\boldsymbol{x};\boldsymbol{x}') - G_0(\boldsymbol{x};\boldsymbol{\xi}). \tag{2.6}$$

where G_0 is defined in Eq.(2.3) and $|\boldsymbol{x}-\boldsymbol{x}'| = |\boldsymbol{x}-\boldsymbol{\xi}|$ for all $\boldsymbol{x} \in \mathbb{T}$. The field G is the Green's function for a half-space with Dirichlet boundary conditions, that is it satisfies the following conditions:

$$\begin{aligned} (\Delta + k^2)G = 4\pi(\delta(\boldsymbol{x}-\boldsymbol{x}') - \delta(\boldsymbol{x}-\boldsymbol{\xi})) &\quad \text{in} \quad \Omega; \\ G = 0 &\quad \text{on} \quad \mathbb{T}; \\ |\boldsymbol{x}-\boldsymbol{\xi}|\left(\frac{\partial G}{\partial \vec{n}} - \sqrt{-1}\,kG\right) \to 0 &\quad \text{as} \quad |\boldsymbol{x}-\boldsymbol{\xi}| \to \infty. \end{aligned}$$

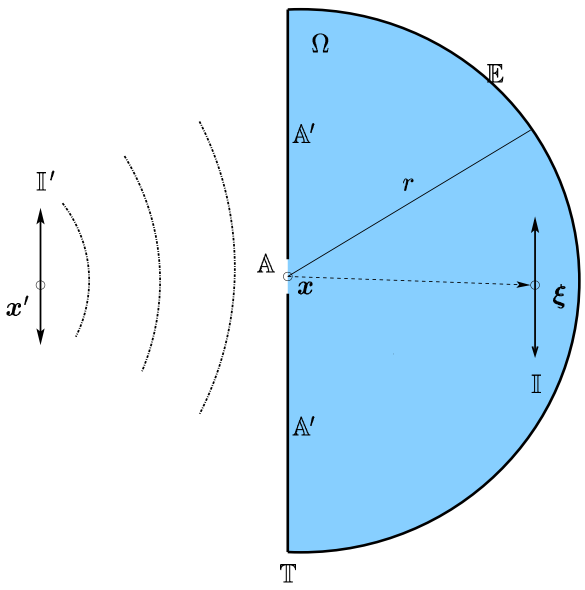


Figure 2.2: Rayleigh-Sommerfeld diffraction

The field G satisfies the conditions required for substitution into Eq.(2.5) in place of G_0 yielding

$$U(\boldsymbol{\xi}) = \frac{1}{4\pi}\int_{\mathbb{S}} G\frac{\partial U}{\partial \vec{n}} - U\frac{\partial G}{\partial \vec{n}} dS. \tag{2.7}$$

While G is identically zero on the plane $\mathbb{T}$ between $\boldsymbol{x}'$ and $\boldsymbol{\xi}$, it's normal derivative is nonzero. We postulate that the unknown field U satisfies the following conditions:

$$U = 0 \quad \text{on} \quad \mathbb{A}'; \tag{2.8}$$

$$|\boldsymbol{x} - \boldsymbol{\xi}|\left(\frac{\partial U}{\partial \vec{n}} - \sqrt{-1}\, kU\right) \to 0 \quad \text{as} \quad |\boldsymbol{x} - \boldsymbol{\xi}| \to \infty \tag{2.9}$$

Condition Eq.(2.8) states that the screen is a nearly "perfect conductor"; Eq.(2.9) is the *Rayleigh-Sommerfeld radiation condition.* In the limit as the radius of the hemisphere $\mathbb{E}$ goes to infinity, Eq.(2.6) and Eq.(2.7) together with the radiation conditions and Eq.(2.8) yield

$$U(\boldsymbol{\xi}) = \frac{1}{4\pi}\int_{\mathbb{A}} -U\frac{\partial G}{\partial \vec{n}} dS. \tag{2.10}$$

Let α map two vectors to the cosine of the angle between them

$$\alpha(\boldsymbol{x}, \boldsymbol{y}) = \frac{\boldsymbol{x}\cdot\boldsymbol{y}}{|\boldsymbol{x}||\boldsymbol{y}|}.$$

If $|\boldsymbol{\xi} - \boldsymbol{x}| \gg \lambda$ on $\mathbb{A}$ then

$$\begin{aligned}\frac{\partial G}{\partial \vec{n}} &= 2\frac{\exp(\sqrt{-1}\,k|\boldsymbol{\xi} - \boldsymbol{x}|)}{|\boldsymbol{\xi} - \boldsymbol{x}|}\left(\sqrt{-1}\,k - \frac{1}{|\boldsymbol{\xi} - \boldsymbol{x}|}\right)\alpha(\vec{n}, \boldsymbol{\xi} - \boldsymbol{x}) \\ &\approx 2k\sqrt{-1}\,\frac{\exp(\sqrt{-1}\,k|\boldsymbol{\xi} - \boldsymbol{x}|)}{|\boldsymbol{\xi} - \boldsymbol{x}|}\alpha(\vec{n}, \boldsymbol{\xi} - \boldsymbol{x}). \end{aligned} \tag{2.11}$$

Substituting Eq.(2.11) into Eq.(2.10) yields the following mathematical formulation of Huygens' principle

$$U(\boldsymbol{\xi}) \approx \int_{\mathbb{A}} U(\boldsymbol{x}) h(\boldsymbol{\xi}; \boldsymbol{x}) dS \tag{2.12}$$

where

$$h(\boldsymbol{\xi}; \boldsymbol{x}) \equiv \frac{\exp(\sqrt{-1}\,k|\boldsymbol{\xi} - \boldsymbol{x}|)}{\sqrt{-1}\,\lambda|\boldsymbol{\xi} - \boldsymbol{x}|}\alpha(\vec{n}, \boldsymbol{\xi} - \boldsymbol{x}).$$

and, again, $\lambda = 2\pi/k$ is the wavelength.

At this point it is useful to introduce into the discussion the *paraxial* or small angle approximation wherein $\alpha(\vec{n}, (\boldsymbol{\xi} - \boldsymbol{x}) \approx 1$. For this we establish reference coordinates (x_1, x_2, x_3) relative to the plane $\mathbb{T}$ centered on the region $\mathbb{A}$. Let the x_3-axis be perpendicular to $\mathbb{T}$ and $\mathbb{I}$, with the origin at the center of the region $\mathbb{A}$. Let $\boldsymbol{x} \in \mathbb{A}$. Denote the distance between $\mathbb{I}$ and $\mathbb{A}$ by ξ_3, and let $\boldsymbol{\xi} \in \mathbb{I}$ satisfy $|(x_1 - \xi_1, x_2 - \xi_2, 0)| \ll \xi_3$. Then $\alpha(\vec{n}, \boldsymbol{\xi} - \boldsymbol{x}) \approx 1$ and the kernel of the Rayleigh-Sommerfeld diffraction integral is $h(\boldsymbol{x}; \boldsymbol{\xi}) \approx \frac{\exp(\sqrt{-1}\,k|\boldsymbol{\xi} - \boldsymbol{x}|)}{\sqrt{-1}\,\lambda|\boldsymbol{\xi} - \boldsymbol{x}|}$. Using the binomial expansion, in the region where both $|\xi_1 - x_1| \ll \xi_3$ and $|\xi_2 - x_2| \ll \xi_3$, yields

$$|\boldsymbol{\xi} - \boldsymbol{x}| \approx \xi_3\left[1 + \frac{1}{2\xi_3^2}(\xi_1 - x_1)^2 + \frac{1}{2\xi_3^2}(\xi_2 - x_2)^2\right]. \tag{2.13}$$

Using this approximation and neglecting the quadratics in the denominator, the kernel h reduces to the well known *Fresnel* kernel

$$h_{Fre}(\boldsymbol{\xi}; \boldsymbol{x}) = \frac{\exp(\sqrt{-1}\,k\xi_3)}{\sqrt{-1}\,\lambda\xi_3}\exp\left(\frac{\sqrt{-1}\,k}{2\xi_3}\left((\xi_1 - x_1)^2 + (\xi_2 - x_2)^2\right)\right). \tag{2.14}$$

This kernel exactly satisfies what is known as the *parabolic* wave equation

$$\left[\frac{\partial}{\partial \xi_3} - \frac{\sqrt{-1}}{2k}\triangle_t - ik\right] h_{Fre} = 0 \tag{2.15}$$

where $\triangle_t$ is the Laplacian in the $\xi_1\xi_2$−plane, *i.e* $\triangle_t = \frac{\partial^2}{\partial \xi_1^2} + \frac{\partial^2}{\partial \xi_2^2}$. By substituting h_{Fre} into Eq.(2.12), we obtain the Fresnel diffraction field

$$U_{Fre}(\boldsymbol{\xi}) = \int_{\mathbb{A}} U(\boldsymbol{x}) h_{Fre}(\boldsymbol{\xi}; \boldsymbol{x})\, dx_1\, dx_2. \tag{2.16}$$

This field also satisfies Eq.(2.15).

If the aperture is small compared to the image ($x_1, x_2 \ll \xi_1, \xi_2$, as is the case in diffraction imaging) one can expand the quadratic in the Fresnel kernel Eq.(2.14) and neglect quadratic terms in x_1 and x_2:

$$\begin{aligned}(\xi_1 - x_1)^2 + (\xi_2 - x_2)^2 &= \xi_1^2 + \xi_2^2 - 2(x_1\xi_1 + x_2\xi_2) + x_1^2 + x_2^2 \\ &\approx \xi_1^2 + \xi_2^2 - 2(x_1\xi_1 + x_2\xi_2).\end{aligned}$$

With this approximation Eq.(2.14) reduces to

$$h_{Fra}(\boldsymbol{\xi}; \boldsymbol{x}) = \frac{\exp(\sqrt{-1}\, k\xi_3)}{\sqrt{-1}\, \lambda\xi_3} \exp\left(\frac{\sqrt{-1}\, k}{2\xi_3}(\xi_1^2 + \xi_2^2)\right) \exp\left(\frac{\sqrt{-1}\, k}{\xi_3}(x_1\xi_1 + x_2\xi_2)\right). \tag{2.17}$$

This is known as the *Fraunhofer* approximation of the Fresnel diffraction field.

The *Fraunhofer transform* of a field $U(\boldsymbol{x})$ across an aperture $\mathbb{A}$ is given by

$$U_{Fra}(\boldsymbol{\xi}) = \int_{\mathbb{A}} U(\boldsymbol{x}) h_{Fra}(\boldsymbol{\xi}; \boldsymbol{x})\, dx_1\, dx_2. \tag{2.18}$$

Close examination of Eq.(2.18) reveals a relationship between the Fraunhofer transform and the Fourier transform. For $u : \mathbb{R}^n \to \mathbb{C}$, let $\wedge$ denote the Fourier transform defined by[2]:

$$u^{\wedge}(\boldsymbol{\xi}) \equiv \int_{\mathbb{R}^n} u(\boldsymbol{x}) \exp(-2\pi\sqrt{-1}\, \boldsymbol{x} \cdot \boldsymbol{\xi})\, d\boldsymbol{x}. \tag{2.19}$$

Let $\mathcal{X}_{\mathbb{A}}$ denote the indicator function for the region $\mathbb{A}$:

$$\mathcal{X}_{\mathbb{A}}(\boldsymbol{x}) \equiv \left\{ \begin{array}{ll} 1 & \text{for } \boldsymbol{x} \in \mathbb{A} \\ 0 & \text{for } \boldsymbol{x} \notin \mathbb{A} \end{array} \right. . \tag{2.20}$$

Assume $U \in L^1 \cap L^2[\mathbb{R}^3, \mathbb{C}]$, then

$$\begin{aligned}U_{Fra}(\boldsymbol{\xi}) &= \int_{\mathbb{A}} U(\boldsymbol{x}) h_{Fra}(\boldsymbol{\xi}; \boldsymbol{x}) dx_1 dx_2 \\ &= C(\boldsymbol{\xi})[\mathcal{X}_{\mathbb{A}} U]^{\wedge_T}(\hat{\xi}_1, \hat{\xi}_2, \xi_3).\end{aligned}$$

Here $\hat{\xi}_i = \frac{1}{\lambda\xi_3}\xi_i$ for $i = 1, 2$, $[\cdot]^{\wedge_T}$ denotes the Fourier transform with respect to the (x_1, x_2) coordinates, and

$$C(\boldsymbol{\xi}) = \frac{\exp(\sqrt{-1}\, k\xi_3)}{\sqrt{-1}\, \lambda\xi_3} \exp\left(\frac{\sqrt{-1}\, k}{2\xi_3}(\xi_1^2 + \xi_2^2)\right).$$

2.1.2 Diffraction Imaging with a Lens

Based on these integral approximations to the field U on the image plane $\mathbb{I}$ we now derive the associated Green's function of the optical system with a lens. We begin with a brief discussion motivating the mathematical model for a thin lens using the paraxial approximation [26, Ch.4] [76, Chapter 5].

---

[2] Note that this definition is valid only for functions in $L^1 \cap L^2$. In Section 2.2 we make use of the extension of this transform to functions on L^2, the Fourier-Plancherel transform.

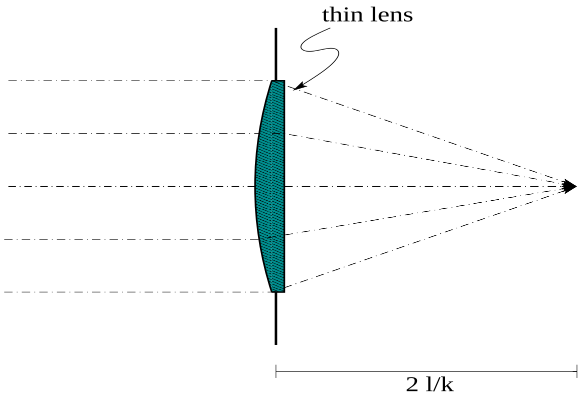


Figure 2.3: Lens model

A lens is modeled from a geometric optics perspective. Under this interpretation a wave propagates along rays orthogonal to its level surface, or in mathematical parlance, along the characteristics of the Helmholtz equation Eq.(2.1). The phase, θ of the complex phasor representation[3], of a wave describes the geometric shape of the level surface, and thus the orientation of the rays along which the wave travels. A lens is a (thin) piece of glass or some other transparent material with a different index of refraction (depending on the wave number k) than the surrounding medium. Physically a lens changes the path [26, Ch.3] of the wave without altering it's amplitude, that is it changes the geometric path of propagation. This is modeled as a change in the direction of the rays, or equivalently a change in the phase θ of the wave across the lens.

For instance, the direction of propagation of the wave described by the Fresnel kernel h_{Fre} (Eq.(2.14)) is parabolic with axis of symmetry along the ξ_3 axis[4]. Suppose we place a lens shown in Fig.(2.3) at the pupil plane of our model optical system Fig.(2.1) with axis of symmetry in the x_3 direction centered at $x_1 = x_2 = 0$. Suppose further that this lens is designed to change the direction of propagation of the field in a parabolic fashion across the aperture $\mathbb{A}$. In the complex phasor representation of the wave, this physical affect is modeled by the addition of a complex phase term to the on the support of the lens. We represent such a lens by the function

$$\phi(\boldsymbol{x}) = \exp\left(\mathcal{X}_{\mathbb{A}}(\boldsymbol{x})\frac{-\sqrt{-1}\,k}{2l}(x_1^2 + x_2^2)\right). \tag{2.21}$$

where k is the wave number and l is a scaling that describes the curvature of the lens. All rays parallel to the axis of symmetry of the lens and passing through the lens will cross the x_3 axis at the point $(0, 0, 2l/k)$. Notice that the lens does not change the amplitude of the wave.

---

[3] The phase function $\theta : \mathbb{R}^3 \to \mathbb{R}$, sometimes called the *eikonal*, satisfies the eikonal equation [26, Ch.3].

[4] The amplitude of the wave is constant in the x_1x_2-plane.

The Fraunhofer approximation Eq.(2.17) to the Fresnel kernel Eq.(2.14) also arises in model of optical systems with a lens. To see this consider a wavefront of the form h_{Fre} at the x_1x_2-plane. The field immediately after the lens is given by multiplying h_{Fre} by the lens Eq.(2.21). If $l = \xi_3$ this multiplication yields the identity $\phi h_{Fre} = h_{Fra}$.

We now detail Kirchhoff's diffraction theory for the following imaging model, based on Huygens' principle Eq.(2.12), with diffraction kernel h_{Fre} and a lens of the form Eq.(2.21):

$$U(\boldsymbol{\xi}) \approx \int_{\mathbb{A}} U(\boldsymbol{x})\phi(\boldsymbol{x})h_{Fre}(\boldsymbol{\xi};\boldsymbol{x})d\boldsymbol{x}. \tag{2.22}$$

The derivation of Eq.(2.12) requires the conditions Eq.(2.8) and Eq.(2.9), where U satisfies Eq.(2.1). Here we encounter the difficulty that we have not specified any boundary conditions on the region $\mathbb{A}$, without which we cannot obtain an explicit approximation for U at $\boldsymbol{\xi}$. Kirchhoff's diffraction theory is based on the conditions Eq.(2.8) and Eq.(2.9) together with the additional boundary condition

$$U(\boldsymbol{x}) = G_0(\boldsymbol{x};\boldsymbol{x}') \quad \text{for } \boldsymbol{x}' \notin \Omega, \quad \text{and } \boldsymbol{x} \in \mathbb{A}$$

where G_0 is given by Eq.(2.3). Since $\boldsymbol{x}'$ enters as a parameter on the right hand side, we write the field U on $\mathbb{A}$ satisfying the above equation as

$$U(\boldsymbol{x};\boldsymbol{x}') = G_0(\boldsymbol{x};\boldsymbol{x}') \quad \text{for } \boldsymbol{x}' \notin \Omega, \quad \text{and } \boldsymbol{x} \in \mathbb{A}. \tag{2.23}$$

Similarly, we write $U(\boldsymbol{\xi}) = U(\boldsymbol{\xi};\boldsymbol{x}')$ to indicate that the field U on $\mathbb{I}$ is also parameterized by the location of the point source $\boldsymbol{x}'$. Conditions Eq.(2.8) and Eq.(2.23) are called *Kirchhoff's boundary conditions.* The function satisfying Eq.(2.1), Eq.(2.8)-(2.9) and Eq.(2.23) is very special indeed. For most applications, however, it is sufficient to approximate the field U by the field that would result from a point source at $\boldsymbol{x}'$ in the absence of the screen $\mathbb{S}$, that is $U(\cdot;\boldsymbol{x}') \approx G_0(\cdot;\boldsymbol{x}')$ everywhere to the left of the screen *except* on $\mathbb{A}'$ where $U(\cdot;\boldsymbol{x}') = 0$. The justification of such an approximation is beyond the scope of this work. There is a vast classical literature surrounding this problem. Interested readers are referred to [26, Chapter 11] and references therein.

Assume next that $\boldsymbol{x}'$ satisfies $|(x_1', x_2', 0)| \ll x_3'$ where $x_3' = \text{dist}(\boldsymbol{x}', \mathbb{T})$. Then, as in the derivation of the Fresnel kernel Eq.(2.14), the field at any point $\boldsymbol{x} \in \mathbb{A}$ can be approximated by

$$U(\boldsymbol{x};\boldsymbol{x}') \approx \frac{\exp(\sqrt{-1}\; kx_3')}{\sqrt{-1}\; \lambda x_3'} \exp\left(\frac{\sqrt{-1}\; k}{2x_3'}\left((x_1 - x_1')^2 + (x_2 - x_2')^2\right)\right) \tag{2.24}$$

Substituting Eq.(2.24) into Eq.(2.22) with the lens Eq.(2.21) yields

$$\begin{aligned} U(\boldsymbol{\xi};\boldsymbol{x}') \;\; \equiv \;\; & \frac{\exp(ik(x_3' + \xi_3))}{-\lambda^2 x_3'\xi_3}\tilde{C}(\boldsymbol{\xi})\tilde{C}(\boldsymbol{x}') \times \\ & \iint_{-\infty}^{\infty} \mathcal{X}_{\mathbb{A}}(\boldsymbol{x}) \exp\left(\frac{\sqrt{-1}\; k}{2}(1/x_3' + 1/\xi_3 - 1/l)(x_1^2 + x_2^2)\right) \times \\ & \quad \exp\left(\frac{-2\pi\sqrt{-1}}{\lambda x_3'\xi_3}(x_1'\xi_3 + \xi_1 x_3',\; x_2'\xi_3 + \xi_2 x_3') \cdot (x_1, x_2)\right) dx_1 dx_2. \end{aligned} \tag{2.25}$$

Here $\tilde{C}(\boldsymbol{\xi}) = \exp\left(\frac{\sqrt{-1}\,k}{2\xi_3}(\xi_1^2+\xi_2^2)\right)$ and likewise for $\tilde{C}(\boldsymbol{x}')$. When the *lens law* [76, Eq.(5-30)] is satisfied, i.e. when

$$1/x_3' + 1/\xi_3 - 1/l = 0, \tag{2.26}$$

then the rays along which the light wave travels depend only linearly on the coordinates in the aperture $\mathbb{A}$. The field at $\mathbb{I}$ is said to be *in focus* when the lens law is satisfied since this plane (where the receptors lie) coincides with the level surface of the wave[5]. We consider only those points $(\xi_1, \xi_2, \xi_3) \in \mathbb{I}$ and $(x_1', x_2', x_3') \in \mathbb{I}'$ for which

$$\xi_3 \gg \frac{k(\xi_1^2+\xi_2^2)}{2} \quad \text{and} \quad x_3' \gg \frac{k({x_1'}^2+{x_2'}^2)}{2} \tag{2.27}$$

where $\mathbb{I}$ and $\mathbb{I}'$ is the planes depicted in Fig.(2.2). Then, as with the Fraunhofer approximation, the $\tilde{C}(\cdot)$ factors are nearly unity. Thus, when Eq.(2.27) and the lens law Eq.(2.26) hold,

$$\begin{aligned} U(\boldsymbol{\xi};\boldsymbol{x}') \;\approx\; & \frac{-\exp(ik(x_3'+\xi_3))}{\lambda^2 x_3'\xi_3} \times \\ & \iint_{-\infty}^{\infty} \mathcal{X}_{\mathbb{A}}(\boldsymbol{x}) \exp\left(\frac{-2\pi\sqrt{-1}}{\lambda x_3'\xi_3}(\xi_3 x_1' + x_3'\xi_1,\ \xi_3 x_2' + x_3'\xi_2)\cdot(x_1,x_2)\right) dx_1 dx_2. \end{aligned} \tag{2.28}$$

The field $U(\boldsymbol{\xi};\boldsymbol{x}')$ is the field at the image plane of a diffractive optical system with a lens due to a point source at $\boldsymbol{x}'$. Define the change of variables

$$\hat{\boldsymbol{x}} = \frac{\xi_3}{x_3'}\boldsymbol{x}' \quad \text{and} \quad \hat{\boldsymbol{\xi}} = \hat{\boldsymbol{x}} + \boldsymbol{\xi}$$

to obtain the following Fourier transform representation

$$\begin{aligned} U(\boldsymbol{\xi};\boldsymbol{x}') \;&\approx\; c\iint_{-\infty}^{\infty} \mathcal{X}_{\mathbb{A}}(\boldsymbol{x}) \exp\left(\frac{-2\pi i}{\lambda\xi_3}(\hat{\xi}_1,\hat{\xi}_2)\cdot(x_1,x_2)\right) dx_1 dx_2 \\ &=\; c\mathcal{X}_{\mathbb{A}}^{\wedge_T}\left(\frac{\hat{\boldsymbol{\xi}}}{\lambda\xi_3}\right) \end{aligned} \tag{2.29}$$

where $c = \frac{-\exp(ik(x_3'+\xi_3))}{\lambda^2 x_3'\xi_3}$ and, again, $\cdot^{\wedge_T}$ denotes the Fourier transform in the $\hat{\xi}_1\hat{\xi}_2$-plane.

The image ψ due to an *extended* source φ in the object plane $\mathbb{I}'$ is given by the superposition of the optical system's response to point sources

$$\psi(\boldsymbol{\xi}) = \int_{\mathbb{R}^2} U\left(\boldsymbol{\xi};\boldsymbol{x}'\right)\varphi(\boldsymbol{x}')dx_1'dx_2'$$

If every point in the support of the source in the object plane satisfies $|(x_1', x_2', 0)| \ll x_3'$, as we have been assuming all along, then we can approximate the dependence of $U(\boldsymbol{\xi};\boldsymbol{x}')$ on $\boldsymbol{x}'$ by $U(\boldsymbol{\xi};\boldsymbol{x}') \approx U(\boldsymbol{\xi}) = U\left(\hat{\boldsymbol{\xi}} - \hat{\boldsymbol{x}}\right)$. This approximation implies that the system's response to a point source $U(\boldsymbol{\xi};\boldsymbol{x}')$ remains invariant under translation of the source in the $x_1'x_2'$-plane[6].

---

[5] Note that the lens law depends entirely on the parabolic approximation to the incident wavefront given by Eq.(2.24).

[6] Regions in the $x_1'x_2'$-plane over which this approximation are employed are called *isoplanatic patches.*

The superposition is thus represented by the two dimensional convolution, denoted $*$,

$$\begin{aligned}\psi(\hat{\boldsymbol{\xi}}) &= \int_{\mathbb{R}^2} U\left(\hat{\boldsymbol{\xi}} - \hat{\boldsymbol{x}}\right)\hat{\varphi}(\hat{\boldsymbol{x}})d\hat{x}_1 d\hat{x}_2 \\ &= U * \hat{\varphi}(\hat{\boldsymbol{\xi}}) \end{aligned} \tag{2.30}$$

where $\hat{\varphi}(\hat{\boldsymbol{x}}) = \left(\frac{x_3'}{\xi_3}\right)^2 \varphi\left(\frac{x_3'}{\xi_3}\hat{\boldsymbol{x}}\right) = \left(\frac{x_3'}{\xi_3}\right)^2 \varphi\left(\boldsymbol{x}'\right)$.

2.1.3 Incoherent fields

The last piece of physics to be added to the mathematical model is the fact that what is actually measured in many optical devices is the *intensity* of an *incoherent* field. In this setting, these are statistical properties of waves. In the interest of brevity, our discussion is terse. Interested readers are referred to [77]. In Eq.(2.1) we have only accounted for the spatial component of a time harmonic wave. The entire wave is of the form $U_\omega(\boldsymbol{x}, t) = U(\boldsymbol{x})\exp(\sqrt{-1}\ \omega t)$, (see Eq.(2.1) for fixed frequency ω. Define the *mutual coherence function*, Γ, to be the cross correlation of light at $\boldsymbol{x}$ and $\boldsymbol{y}$

$$\Gamma(\boldsymbol{x}, \boldsymbol{y}, \tau) \equiv \left\langle\!\left\langle U_\omega(\boldsymbol{x}, \cdot + \tau),\ \overline{U}_\omega(\boldsymbol{y}, \cdot)\right\rangle\!\right\rangle . \tag{2.31}$$

where $\overline{U}_\omega$ denotes the complex conjugate and $\langle\langle \cdot,\ \cdot\rangle\rangle$ denotes an infinite time average

$$\left\langle\!\left\langle U_\omega(\boldsymbol{x}, \cdot + \tau),\ \overline{U}_\omega(\boldsymbol{y}, \cdot)\right\rangle\!\right\rangle \equiv \lim_{T\to\infty} \frac{1}{T}\int_{-T/2}^{T/2} U_\omega(\boldsymbol{x}, t+\tau)\overline{U}_\omega(\boldsymbol{y}, t)dt.$$

The normalized mutual coherence function evaluated at $\tau = 0$ measures the *spatial coherence* of the light. The *mutual intensity* of the light at $\boldsymbol{x}$ and $\boldsymbol{y}$ is defined by

$$J(\boldsymbol{x}, \boldsymbol{y}) \equiv \Gamma(\boldsymbol{x}, \boldsymbol{y}, 0).$$

The (coincident) intensity is simply the modulus squared of the wave at the point $\boldsymbol{x}$:

$$J(\boldsymbol{x}, \boldsymbol{x}) = \left\langle\!\left\langle U_\omega(\boldsymbol{x}, \cdot), \overline{U}_\omega(\boldsymbol{x}, \cdot)\right\rangle\!\right\rangle = |U(\boldsymbol{x})|^2.$$

The intensity of the image ψ in Eq.(2.30) at a point $\hat{\boldsymbol{\xi}}$ is thus given by

$$|\psi|^2 = \left|U * \hat{\varphi}(\hat{\boldsymbol{\xi}})\right|^2 \tag{2.32}$$

Rearranging the integrals yields

$$\left|\psi(\hat{\boldsymbol{\xi}})\right|^2 = \int_{\mathbb{R}^2}\int_{\mathbb{R}^2} U(\hat{\boldsymbol{\xi}} - \hat{\boldsymbol{x}})\overline{U}(\hat{\boldsymbol{\xi}} - \hat{\boldsymbol{y}})\hat{\varphi}(\hat{\boldsymbol{x}})\overline{\hat{\varphi}}(\hat{\boldsymbol{y}})d\hat{x}_1 d\hat{x}_2\, d\hat{y}_1 d\hat{y}_2.$$

However, the resolution of our optical system in the image plane is such that what is observed is the time averaged quantity

$$\left|\psi(\hat{\boldsymbol{\xi}})\right|^2 = \int_{\mathbb{R}^2}\int_{\mathbb{R}^2} J(\hat{\boldsymbol{\xi}} - \hat{\boldsymbol{x}}, \hat{\boldsymbol{\xi}} - \hat{\boldsymbol{y}})\hat{\varphi}(\hat{\boldsymbol{x}})\overline{\hat{\varphi}}(\hat{\boldsymbol{y}})d\hat{x}_1 d\hat{x}_2\, d\hat{y}_1 d\hat{y}_2. \tag{2.33}$$

If in addition the resolution of the optical system has a resolution in the image plane that is coarser than the spatial coherence of the light, then the light is said to be *incoherent.* The mutual intensity corresponding to incoherence can be approximated by

$$J(\hat{\boldsymbol{\xi}}, \hat{\boldsymbol{\eta}}) \approx \hat{c}|U(\hat{\boldsymbol{\xi}})|^2\delta(\hat{\boldsymbol{x}} - \hat{\boldsymbol{\eta}}) \tag{2.34}$$

where $\hat{c}$ is some real constant. For a detailed discussion of this intricate theory see [77, Section 5.5]. Substituting Eq.(2.34) into Eq.(2.33) yields

$$\left|\psi(\hat{\boldsymbol{\xi}})\right|^2 \approx \hat{c}\,|U|^2 * \left|\hat{\varphi}(\hat{\boldsymbol{\xi}})\right|^2. \tag{2.35}$$

If the coherence of the light is resolvable, then one must work with the less convenient representation of Eq.(2.32).

2.1.4 Rescaling the model

We assume that $\lambda\xi_3 = 1$ (this is equivalent to resizing the aperture). The contribution of the x_3' component to the field U given in Eq.(2.29) is just a scalar multiple. This is normalized so that the scaling in Eq.(2.35) is unity,

$$\hat{c}|U|^2 * |\varphi|^2(\boldsymbol{\xi}) = \left|\tilde{c}\mathcal{X}_{\mathbb{A}}^{\wedge}\right|^2 * |\varphi|^2(\boldsymbol{\xi}) \tag{2.36}$$

where $\tilde{c} = -\exp\left(ik(x_3' + \xi_3)\right)$. We represent the field at the exit pupil of the optical system, that is on the right "side" of the pupil plane of the imaging system in Fig.(2.1), by the function $u: \mathbb{R}^2 \to \mathbb{C}$. In Eq.(2.36) this field is known

$$u = \tilde{c}\mathcal{X}_{\mathbb{A}} \quad \text{and} \quad U = u^{\wedge}. \tag{2.37}$$

We show in the next section that the field u is not always of this form.

The normalized mathematical model for the intensity mapping in the focal plane of a diffracted, incoherent, monochromatic, far-field electromagnetic field Eq.(2.35) becomes

$$|\psi|^2(\boldsymbol{\xi}) \approx \left|u^{\wedge}\right|^2 * |\varphi|^2(\boldsymbol{\xi}). \tag{2.38}$$

The kernel of the convolution $|u^{\wedge}|^2$ is known as the *point-spread function* of the idealized optical system of Fig.(2.1). This kernel characterizes the optical system.

2.1.5 Aberrated Optical Systems

It has been assumed that the optical system is in the far-field (with respect to some source) of a homogeneous medium, thus the wave at the entrance pupil, that is on the *left* side of the pupil plane, is characterized by a constant amplitude plane wave, $\arg(u_-) \equiv 0$ and $|u_-| = \mathit{const}$ across the aperture $\mathbb{A}$, where u_- indicates the field at the entrance pupil. This is often called the *reference* wave. In most applications, however, the assumption of homogeneity is not correct for the field at the exit pupil. Inhomogeneities in the media cause deviations in the true wave from the reference wave. There are two types of deviations from the reference wave. We refer to deviations in the phase as *phase aberrations* and deviations from the amplitude as the *throughput* of the optical system. Deviations may occur at any point along the path of propagation and can be caused by an intervening medium such as atmosphere, crystal structure, or mirror surface. In geometric optics, the wave is assumed to travel along rays normal to the wavefront. The phase represents differences in the optical path length along different rays. The locations of the deviations along the rays are not important. Accordingly, all deviations are taken to occur at the pupil plane depicted in Fig.(2.1).

A simple example of a phase aberration is defocus which can be modeled by use of a lens as in Eq.(2.21). The field due to a defocused generalized pupil function is given by Eq.(2.25) where the lens law Eq.(2.26) is not satisfied, i.e. $1/z_0 + 1/\zeta - 1/l = \epsilon,\ 0 < |\epsilon| \ll 1$. It often happens, however, that the aberration is unknown. Defocus is added to an optical system to improve signal to noise ratios in the tails of images. Defocus is also used to stabilize numerical schemes for recovering $\arg(u)$ (phase retrieval) and φ (deconvolution) from the image ψ.

The throughput of the optical system is affected by the mounts and bolts used to hold optical mirrors in place as well as the support of the aperture. These objects change the amplitude of the wave as it propagates through the system and are modeled by the amplitude of the field u.

The function u accounting for all of the above aberrations is refered to as the *generalized pupil function.* The generalized pupil function uniquely characterizes the optical system. For a perfect, deviation-free normalized optical system (where in particular $\lambda\xi_3 = 1$) the generalized pupil function is given by $u = \tilde{c}\mathcal{X}_{\mathbb{A}}$ as in Eq.(2.37). For a field with deviations from the reference wave u_-, i.e. with phase aberration $\theta(\boldsymbol{x})$ and throughput $A(\boldsymbol{x})$ the generalized pupil function can be represented in complex phasor form by

$$u[A(\boldsymbol{x}), \theta(\boldsymbol{x})] = A(\boldsymbol{x}) \exp[\sqrt{-1}\ \theta(\boldsymbol{x})]. \tag{2.39}$$

The corresponding imaging model for an aberrated optical system is the same as Eq.(2.38).

2.1.6 Notation and Summary

We now establish the notation which will be used throughout the remainder of this work and summarize the above results with the new notation. Since the third spatial dimension, x_3' and ξ_3, only determines relative scalings and magnification factors in the image plane and the pupil planes, we will only be interested in the behavior of the fields in the x_1x_2-plane (respectively the $\xi_1\xi_2$-plane). From this point forward the fields are therefore described as mappings on $\mathbb{R}^2$. Rather than defining a new variable for the intensity of the image and object, we reassign the variables ψ and φ to represent rescaled amplitude mappings instead of complex scalar waves:

$$|\psi| \to \psi : \ \mathbb{R}^2 \to \mathbb{R}_+ \quad \text{and} \quad |\tilde{\varphi}|^2 \to \varphi : \ \mathbb{R}^2 \to \mathbb{R}_+ \, .$$

The imaging model thus takes the form

$$\psi^2(\boldsymbol{\xi}) \approx \left|u^\wedge\right|^2 * \varphi(\boldsymbol{\xi}). \tag{2.40}$$

2.2 Inverse Problems

In the previous sections we have taken great care to develop the *forward* model for image formation. We now turn our attention to the *inverse* problem. If u is known and φ unknown, Eq.(2.40) is a Fredholm integral equation of the first kind. Recovering φ from u and ψ^2 is called, for good reason, *deconvolution.* The *phase retrieval* problem arises when the amplitude of the generalized pupil function u is known, but the phase aberration, θ in Eq.(2.39), is unknown. When both the object φ and the phase aberrations in u are unknown[7] one is faced with the problem of *simultaneous* deconvolution and phase retrieval. We discuss the simultaneous problem since deconvolution and phase retrieval are special cases.

Suppose that the amplitude $|u|$ is known and satisfies the equation

$$A = |u| \tag{2.41}$$

where $A : \mathbb{R}^2 \to \mathbb{R}_+$ is known. This is often modeled as an indicator function for the aperture, $A = \mathcal{X}_{\mathbb{A}}$. For the purposes of this work it is only necessary to note that $A \in \mathbb{U}_+$ where $\mathbb{U}_+$ is a cone of nonnegative functions to be explicitly defined below. According to the uniqueness results proved in [85], for discrete band-limited images in two dimensions, if a solution to the phase retrieval problem exists, knowledge about both $|u|$ and $|u^\wedge|$ uniquely characterizes u and thus the optical system, up to a complex constant, linear shifts and rotations by 180 degrees.

In most cases there is no closed-form analytic solution to the phase retrieval problem. Notable exceptions were first recognized in [56, 100, 158]. In numerical approaches, the problem is further constrained by the addition of known phase aberrations to the system. The corresponding images are called *diversity images.* The problem is then to find the unknown phase common to all images given the amplitude constraints. For $m = 1, \ldots, M$ let $\tilde{\theta}_m : \mathbb{R}^2 \to \mathbb{R}$ denote a known phase aberration added to the system across the aperture. The corresponding diversity images are denoted by $\psi_m : \mathbb{R}^2 \to \mathbb{R}$. These images are approximated by

$$\psi_m^2 \approx |\mathcal{P}_m[u]|^2 * \varphi \tag{2.42}$$

where $\mathcal{P}_m$ is defined by

$$\mathcal{P}_m[u] \equiv \left[u \exp[\sqrt{-1}\, \tilde{\theta}_m]\right]^\wedge . \tag{2.43}$$

Therefore the mth aberrated point-spread function is $|\mathcal{P}_m[u]|^2$. The phase retrieval problem for M diversity images is formulated as a system of nonlinear equations

$$\begin{pmatrix} A^2 \\ \psi_1^2 \\ \vdots \\ \psi_M^2 \end{pmatrix} = \begin{pmatrix} |u|^2 \\ |\mathcal{P}_1[u]|^2 * \varphi \\ \vdots \\ |\mathcal{P}_M[u]|^2 * \varphi \end{pmatrix} . \tag{2.44}$$

---

[7] This is a common situation in land-based astronomical observation where the earth's atmosphere introduces unknown phase aberrations during observations.

We indulge in a little notational abstraction that will come in handy later. Define the convolution operator $\mathcal{K}_m$ by

$$\mathcal{K}_m[u]\varphi = \psi_m^2, \tag{2.45}$$

where

$$\mathcal{K}_m[u]\varphi \equiv |\mathcal{P}_m[u]|^2 * \varphi. \tag{2.46}$$

In order to easily include the pupil constraint into the system of operator equations above, define the constant operator

$$\mathcal{K}_0\varphi \equiv |\mathcal{P}_0[u]|^2 \tag{2.47}$$

where

$$\mathcal{P}_0 \equiv \mathcal{I} \tag{2.48}$$

for the identity operator $\mathcal{I}$. Define the the aperture constraint Eq.(2.41) to be ψ_0:

$$\psi_0 \equiv A.$$

Thus for all φ,

$$|\boldsymbol{u}|^2 = A^2 \qquad \longleftrightarrow \qquad \mathcal{K}_0[\boldsymbol{u}] = \psi_0^2 \tag{2.49}$$

For the vector $\boldsymbol{v} = (v_1, \ldots, v_n)$ let $\boldsymbol{v}^{\cdot m} = (v_1^m, \ldots, v_n^m)$. The phase diversity problem is formulated as a system of $M+1$ operator equations

$$\boldsymbol{\psi}^{\cdot 2} = \boldsymbol{\mathcal{K}}[u]\varphi \tag{2.50}$$

where $\boldsymbol{\mathcal{K}}[u] : \mathbb{X} \to \mathbb{X}^M$ and $\boldsymbol{\psi} \in \mathbb{X}^M$ for $\varphi \in \mathbb{X}$, some set to be explicitly defined below. Here $\boldsymbol{\psi}$ is a vector of images and $\boldsymbol{\mathcal{K}}[\boldsymbol{u}]$ is a linear operator parameterized by the function u

$$\boldsymbol{\psi} \equiv \begin{pmatrix} \psi_0 \\ \vdots \\ \psi_M \end{pmatrix} \quad \text{and} \quad \boldsymbol{\mathcal{K}} = \begin{pmatrix} \mathcal{K}_0 \\ \vdots \\ \mathcal{K}_M \end{pmatrix}. \tag{2.51}$$

Dual Representations

The system of equations given by Eq.(2.44) is diagonalized by transforming the equation to it's Fourier dual. The *Fourier dual* to an equation is defined as the Fourier transform of both sides of the equation. For example, by the convolution theorem any convolution operator, $\mathcal{G}$, with kernel $g \in L^1$, is associated with a dual multiplication operator G, with "kernel" $g^\wedge$, defined by the Fourier dual to the corresponding operator equation:

$$\mathcal{G}\varphi \equiv g * \varphi = \psi \qquad \overset{\wedge}{\longleftrightarrow} \qquad G\varphi^\wedge \equiv \left(g^\wedge\right)\left(\varphi^\wedge\right) = \psi^\wedge.$$

By the convolution theorem, the dual operator to the convolution operator with kernel $|\mathcal{P}_m[u]|$ in Eq.(2.44) is diagonal with Hermitian kernel. A *Hermitian* function $v : \mathbb{R}^n \to \mathbb{C}$ satisfies $v(\boldsymbol{x}) = \overline{v}(-\boldsymbol{x})$. Equivalently, v is Hermitian if and only if $v^\wedge$ is a real-valued function.

Denote the adjoint to the Fourier-Plancherel transform by $\vee$. Using the identity $v^{\wedge\wedge} = v(-\boldsymbol{x}) = v^{\vee\vee}$, it is straight forward to verify that, for any complex-valued scalar functions $v, w \in (L^1 \cap L^2)[\mathbb{R}^n, \mathbb{C}]$,

$$\left((v^{\wedge}) \, (w^{\vee}) \right)^{\wedge} \; = v \star w, \tag{2.52}$$

where $\star$ is the correlation operator defined by

$$v \star w(\boldsymbol{x}) \equiv \int_{\mathbb{R}^n} v(\boldsymbol{x}') w(\boldsymbol{x} + \boldsymbol{x}') d\boldsymbol{x}'. \tag{2.53}$$

The associate operation to the $\star$ operator is denoted by $\square$ and is defined by

$$v \square w(\boldsymbol{x}) \equiv \int_{\mathbb{R}^n} v(\boldsymbol{x}') w(\boldsymbol{x}' - \boldsymbol{x}) d\boldsymbol{x}'. \tag{2.54}$$

For a listing of relations of these operators see Appendix A.

For $m \neq 0$ The Fourier dual to the system of equations Eq.(2.44) is

$$\begin{aligned} [|\mathcal{P}_m[u]|^2]^{\wedge} &= \left[\mathcal{P}_m[u] \cdot \overline{\mathcal{P}_m[u]} \right]^{\wedge} \\ &= (\mathcal{P}_m[u])^{\vee} \star \left(\overline{\mathcal{P}_m[u]} \right)^{\wedge} \\ &= (\mathcal{P}_m[u])^{\vee} \star \overline{(\mathcal{P}_m[u])^{\vee}} \\ &= \left(u \exp[\sqrt{-1}\, \tilde{\theta}_m] \right) \star \left(\overline{u \exp[\sqrt{-1}\, \tilde{\theta}_m]} \right). \end{aligned} \tag{2.55}$$

The Fourier dual of the convolution operator is a multiplication operator denoted by K_m

$$\mathcal{K}_m[u]^{\wedge} = K_m[u]$$

where

$$K_m[u] \varphi^{\wedge} = \left[\mathcal{P}_m[u] \star \overline{\mathcal{P}_m[u]} \right] \cdot \varphi^{\wedge}. \tag{2.56}$$

For $m = 0$, the Fourier dual of $\mathcal{K}_0$ is a constant operator with "kernel" defined by

$$K_0[u] \varphi^{\wedge} = [|u|^2]^{\wedge} = u^{\vee} \star \overline{u^{\vee}}. \tag{2.57}$$

The Fourier dual of (2.50) is the diagonalized system of operator equations

$$\boldsymbol{K}[u] \varphi^{\wedge} = [\boldsymbol{\psi}^{\cdot 2}]^{\wedge} \tag{2.58}$$

with

$$[\boldsymbol{\psi}^{\cdot 2}]^{\wedge} \equiv \begin{pmatrix} [\psi_0^2]^{\wedge} \\ \vdots \\ [\psi_M^2]^{\wedge} \end{pmatrix} \quad \text{and} \quad \boldsymbol{K} \equiv \begin{pmatrix} K_0 \\ \vdots \\ K_M \end{pmatrix}. \tag{2.59}$$

The diagonalization of the convolution operator is a crucial property for numerical solutions. Note also that the kernel of K_m is Hermitian for all m.

Using the Fourier dual representation one can also write the point-spread function as a quadratic in the dual function $v_m : \mathbb{R}^2 \to \mathbb{C}$ where $v_m \in L^2[\mathbb{R}^2, \mathbb{C}]$

$$v_m = \mathcal{P}_m[u].$$

The image data is simply the pointwise magnitude of v_m convolved against the source φ:

$$|v_m|^2 * \varphi = \psi_m^2. \tag{2.60}$$

The mth operator equation in Eq.(2.44) has the following dual representation

$$|v_m|^2 * \varphi = \psi_m^2. \tag{2.61}$$

By the convolution theorem and equation (2.53), each v_m must satisfy

$$v_m \star \overline{v_m} = \left[A^2\right]^\vee \tag{2.62}$$

It is useful to interpret the functions v_m above in terms of wave propagation. In geometric optics v_m represents the distribution of *ray* components, *i.e.* the directions of propagation, of the wave u across $\mathbb{A}$ through a lens with known aberration $\exp[\sqrt{-1}\ \tilde{\theta}_m]$. In studies of wave propagation in which the Wigner distribution plays a role, the domain of interest is the product space including the physical domain and the spatial frequency domain. The wavefront exists in the physical domain and the distribution of rays normal to the wavefront exists in the spatial frequency domain. This product space is called *phase space.* For a general review of this theory see [17, 18, 194, 195].

A Partial Differential Equation Perspective

It is interesting to consider the problem of wavefront reconstruction in the context of partial differential equations with non-standard data. To see this we separate the aberration free kernel Eq.(2.28) into "slow" and "fast" components

$$U(\boldsymbol{x};\boldsymbol{x}') \approx \exp[\sqrt{-1}\ kx_3']\tilde{U}(\boldsymbol{x};\boldsymbol{x}'). \tag{2.63}$$

Here

$$\tilde{U}(\boldsymbol{x};\boldsymbol{x}') = \frac{1}{\sqrt{-1}\ \lambda x_3'} \exp\left\{\frac{\sqrt{-1}\ k}{2x_3'}[(x_1 - x_1')^2 + (x_2 - x_2')^2]\right\}, \tag{2.64}$$

and satisfies the paraxial wave equation on $\mathbb{A}$

$$\left[\frac{\partial}{\partial x_3} - \frac{\sqrt{-1}}{2k}\triangle_t\right]\tilde{U}(\boldsymbol{x},\boldsymbol{x}') = 0. \tag{2.65}$$

Assume that the kernel with phase aberrations also satisfies Eq.(2.65). The kernel then satisfies the following partial differential equation with nonstandard data:

$$\begin{aligned} \left[\frac{\partial}{\partial \xi_3} - \frac{\sqrt{-1}}{2k}\triangle_t\right] U(\boldsymbol{\xi}) &= 0 && (2.66)\\ |U(\boldsymbol{\xi})| &= \psi(\boldsymbol{\xi}),\ \xi_3 = \text{constant} && (2.67)\\ U \star_t \overline{U}(\boldsymbol{\xi}) &= [A^2]^{\wedge_t}(\boldsymbol{\xi}), && (2.68) \end{aligned}$$

where the t subscript indicates that the operator is only with respect to the transverse coordinates. Rather than solving this partial differential equation directly one seeks solutions to the related optimization problem where the governing Eq.(2.66) provides consistency conditions that further constrain the problem.

Teague [179] has suggested representing aberrations as phase factors in the scalar wavefront represented in Eq.(2.24). The wavefront in region $\mathbb{A}$ is then given by

$$U(\boldsymbol{x}) = I(\boldsymbol{x}) \exp[\sqrt{-1}\, \theta(\boldsymbol{x})] \tag{2.69}$$

where $I(\boldsymbol{x}) \equiv |U(\boldsymbol{x})|^2$. If we take U to be the paraxial approximation to the wave given by Eq.(2.64), then U satisfies the paraxial wave Eq.(2.65). Substituting Eq.(2.69) into Eq.(2.65) and taking real and imaginary parts yields the following system of non-linear equations for the intensity and the phase on the support of U

$$\begin{aligned} k\frac{\partial I}{\partial x_3} &= -\nabla_t \cdot I \nabla_t \theta && (2.70) \\ 2kI^2 \frac{\partial \theta}{\partial x_3} &= \frac{1}{2} I \triangle_t I - \frac{1}{4}(\nabla_t I)^2 - I^2(\nabla_t \theta)^2 && (2.71) \end{aligned}$$

where, again, $\nabla_t = (\frac{\partial}{\partial x_1}, \frac{\partial}{\partial x_2})$ and $\triangle_t = \nabla_t \cdot \nabla_t$.

The common assumption of adaptive optics employed in the phase retrieval problem is that the intensity is constant across $\mathbb{A}$. Thus I in Eq.(2.71) and Eq.(2.70) is known in transverse directions. In the plane containing $\mathbb{A}$ the normalized intensity I is given by the indicator function $\mathcal{X}$. Substituting this into Eq.(2.70) implies that $\frac{\partial I}{\partial x_3}$ is of the form

$$k\frac{\partial I}{\partial x_3} = f(\boldsymbol{x}) + \delta_{\mathbb{A}}(\boldsymbol{x}) g(\boldsymbol{x}) \tag{2.72}$$

for "smooth" functions f and g and the delta distribution $\delta_{\mathbb{A}}$ for the jump at the boundary of $\mathbb{A}$ (denoted by bdy $\mathbb{A}$). Equating terms in Eq.(2.72) and Eq.(2.70) yields the following boundary value problem for Poisson's equation

$$\begin{aligned} \triangle_t \theta &= -f \text{ on the interior of } \mathcal{A} \\ \nabla_t \theta &= -g \text{ on bdy } \mathbb{A}. && (2.73) \end{aligned}$$

It is possible to obtain direct measurements of $\frac{\partial I}{\partial x_3}$, but this information requires hardware modifications [79, 80]. In the absence of such information the above analysis provides consistency conditions on θ that can be used to constrain optimization schemes. In particular, conservation of energy requires that

$$\begin{aligned} \int_{\mathbb{A}} \frac{\partial}{\partial x_3} I \, dx_1 dx_2 &= 0 && (2.74) \\ \Rightarrow \int_{\mathbb{A}} \triangle_t \theta \, dx_1 dx_2 &= 0. && (2.75) \end{aligned}$$

By seeking solutions that in some sense satisfy Eq.(2.75), issues about existence and uniqueness of the boundary value problem Eq.(2.73) are exchanged for questions of existence and uniqueness to the related optimization problem.

An Optimization Perspective

In the presence of noise it is likely that an exact solution to the system of equations given by Eq.(2.44) does not exist[8]. One therefore seeks the best estimate, u_*, for a given performance measure, ρ . While many different algorithms can be applied to recover the best estimate u_* numerically, it is our view that they all address some type of optimization problem. The method by which the best estimate is found involves some sort of optimality principle that depends on the formulation of the underlying optimization problem. Before stating this optimization problem, some remarks about the spaces in which the operators lie are necessary.

To establish a well-posed optimization problem the domain must be closed. The Fourier transform defined by Eq.(2.19) is only valid on $L^1 \cap L^2$, which is not closed. This technicality is avoided by defining the corresponding transform on L^2. The *Fourier-Plancherel* transform is the unique L^2 limit of the Fourier transform of elements in $L^1 \cap L^2$ [99]. All of the properties of the standard Fourier transform hold for this extended definition. In addition to being closed, the space L^2 has the advantage of being a Hilbert space. In all of the following, the transforms $\mathcal{P}_m : L^2[\mathbb{R}^2, \mathbb{C}] \to L^2[\mathbb{R}^2, \mathbb{C}]$ are defined by

$$\mathcal{P}_m[u] \equiv \left[u \exp[\sqrt{-1}\, \tilde{\theta}_m] \right]^{\wedge},$$

where $\wedge$ indicates the Fourier-Plancherel transform. The transform $\mathcal{P}_m$ is a unitary bounded linear operator with adjoint denoted by $\mathcal{P}_m^*$ with $\mathcal{P}_m^* = \mathcal{P}_m^{-1}$.

It will be convenient to represent the fields as mappings into $\mathbb{R}^2$ rather than $\mathbb{C}$. Define the transformation $\mathcal{R} : \mathbb{R}^2 \to \mathbb{C}$ by

$$\mathcal{R}(\boldsymbol{v}) \equiv v_1 + \sqrt{-1}\, v_2,$$

where $\boldsymbol{v} = (v_1, v_2) \in \mathbb{R}^2$. The adjoint of $\mathcal{R}$ with respect to the real inner product for $v, v' \in \mathbb{C}$

$$\langle v, v' \rangle = \mathrm{Re}\, (\overline{v'} v)$$

is given by

$$\mathcal{R}^*(v) = \begin{pmatrix} \mathrm{Re}\, v \\ \mathrm{Im}\, v \end{pmatrix}.$$

The mapping $\mathcal{R}$ is a unitary bounded linear operator with $\mathcal{R}^{-1} = \mathcal{R}^*$. Our discussion switches frequently between finite dimensional and infinite dimensional settings. Therefore, it is convenient to think of $\mathcal{R}$ as a mapping from $L^2[\mathbb{R}^2, \mathbb{R}^2]$ to $L^2[\mathbb{R}^2, \mathbb{C}]$. Whenever there is chance for confusion, square brackets are used to indicate a mapping, e.g.

$$\mathcal{R}[\boldsymbol{v}] \equiv \mathcal{R}(\boldsymbol{v}(\cdot)) \tag{2.76}$$

for $\boldsymbol{v} : \mathbb{R}^2 \to \mathbb{R}^2$.

Using this notation, we equivalently write the field at the exit pupil as the function $\boldsymbol{u} : \mathbb{R}^2 \to \mathbb{R}^2$

$$\boldsymbol{u} = \mathcal{R}^*[u].$$

---

[8] The uniqueness results studied in [85] therefore do not apply.

The imaging equation Eq.(2.42) is equivalently written as

$$\psi^2(\boldsymbol{\xi}) \approx |\mathcal{F}_m[\boldsymbol{u}]|^2 * \varphi(\boldsymbol{\xi}). \tag{2.77}$$

where

$$\mathcal{F}_m[\boldsymbol{u}] \equiv \mathcal{R}^*\left[\mathcal{P}_m[\mathcal{R}[\boldsymbol{u}]]\right]. \tag{2.78}$$

In general $|\cdot|$ denotes the pointwise magnitude where the finite dimensional 2-norm is assumed. The modulus, $|v|$, of a function $v : \mathbb{R}^2 \to \mathbb{C}$, is used interchangeably with the pointwise Euclidean norm $|\boldsymbol{v}|$ of the function $\boldsymbol{v} : \mathbb{R}^2 \to \mathbb{R}^2$. Unless indicated otherwise, $\|\cdot\|$ denotes the L^2 operator norm. Since both $\mathcal{P}_m$ and $\mathcal{R}$ are unitary bounded linear operators, $\mathcal{F}_m$ also has this property. The adjoint is denoted by $\mathcal{F}_m^*$ with $\mathcal{F}_m^* = \mathcal{F}_m^{-1}$. For convenience define

$$\mathcal{F}_0 \equiv \mathcal{I} \tag{2.79}$$

where $\mathcal{I}$ is the identity operator.

The general optimization problem is formulated in terms of the functions $\boldsymbol{u}$ and φ.

$$\begin{aligned} \text{minimize} \quad & \sum_{m=0}^{M} \rho\left[\psi_m,\ \mathcal{K}_m[\boldsymbol{u}]\varphi\ \right] \\ \text{over} \quad & \boldsymbol{u} \in L^2[\mathbb{R}^2],\ \varphi \in L^2[\mathbb{R}^2, \mathbb{R}_+]. \end{aligned} \tag{2.80}$$

The data, ψ_m and A, belong to $\mathbb{U}_+$, a set of functions on the unit sphere (*i.e.* the data is normalized) for which the Fourier transform is well defined and whose tails tend to zero sufficiently fast. Alternatively, the optimization problem can be formulated in terms of the Fourier dual

$$\text{minimize} \quad \sum_{m=0}^{M} \rho\left[\psi_m^\wedge,\ K_m[\boldsymbol{u}]\varphi^\wedge\ \right] \tag{2.81}$$

$$\text{over} \quad \boldsymbol{u} \in L^2[\mathbb{R}^2],\ \varphi^\wedge \in (L^2[\mathbb{R}^2, \mathbb{R}_+])^\wedge. \tag{2.82}$$

Here $(L^2[\mathbb{R}^2, \mathbb{R}_+])^\wedge$ is the Fourier dual to $L^2[\mathbb{R}^2, \mathbb{R}_+]$, *i.e.* the space of Hermitian functions on $\mathbb{R}^2$. The data, $\psi_m^\wedge$ also belongs to $(\mathbb{U}_+)^\wedge$. The rest of this work is dedicated to numerical methods for solving the above optimization problems.

For easy reference, the following is assumed throughout:

Hypothesis 2.2.1 *Let* $\boldsymbol{u} = (u_{re}\ , u_{im}\ )$ *where* u_{re} *and* $u_{im} \in L^2[\mathbb{R}^2, \mathbb{R}]$. *Assume that* ψ_m *satisfies* $\psi_m \in \mathbb{U}_+$ *for* $m = 0, 1, \ldots, M$, *where* $\mathbb{U}_+$ *is the cone of nonnegative functions given by*

$$\mathbb{U}_+ = \left\{ v \in L^1[\mathbb{R}^2, \mathbb{R}_+] \cap L^2[\mathbb{R}^2, \mathbb{R}_+] \cap L^\infty[\mathbb{R}^2, \mathbb{R}_+] \text{ such that } |v(\boldsymbol{x})| \to 0 \ \text{ as } |\boldsymbol{x}| \to \infty \right\}.$$

The remainder of this work is devoted to the study of numerical methods for the solution of the above optimization problem.

Chapter 3

WAVEFRONT RECONSTRUCTION

The wavefront reconstruction problem or phase retrieval is fundamental to the more general problem of *simultaneous* wavefront reconstruction and deconvolution. This chapter is devoted to a careful study of wavefront reconstruction. Numerical approaches to this problem are divided into geometric and analytic methods which lead to seemingly different algorithms. In this chapter we show that the approaches are essentially the same. Furthermore, analytic approaches are well known and have several numerical advantages not available to standard geometric approaches. These issues are discussed in Chapter 5. The purpose of this chapter is to precisely characterize the correspondence between projections of geometrical algorithms and the subdifferential of the squared set distance operator.

3.1 Geometric Approaches

Projection algorithms, such as iterative transform methods, are central to current numerical techniques for solving the phase retrieval problem [13, 29, 52, 68, 73, 110, 125, 172, 200, 202]. Much is known about projections onto convex sets [21, 30, 78, 82, 173, 189, 203]. However, the problem of phase retrieval involves projections onto nonconvex sets. It is shown below that as a consequence of nonconvexity the projections can be multi-valued. This is the principle obstacle to proving the convergence of projection-type algorithms. For special classes of nonconvex sets a convergence theory can be provided [43, 50]. The nonconvex sets considered here do not belong to these classes. The geometric analysis of Ref. [50] applies to the phase retrieval problem although it requires assumptions that are difficult to satisfy. A convergence theory for generalized projection algorithms is developed in [15], however there are no known nonconvex sets to which their hypotheses apply. In particular, the hypotheses required in Proposition 2 of Ref. [15] are not satisfied in the case of phase retrieval. In this section the theory of nonconvex projections is reviewed.

3.1.1 General Theory

Simply stated, iterative transform methods adjust the phase of the current estimate, $\boldsymbol{u}^{(\nu)}$ or $\mathcal{F}_m[\boldsymbol{u}^{(\nu)}]$, at iteration ν and replace the magnitude with the known pointwise magnitude. It is straight forward to show that this operation is a projection.

The amplitude data for a one dimensional example is depicted in Fig. (3.1). The functions satisfying the data belong to sets that are collections of functions that lie on the surface of the tube-like structures depicted in Fig. (3.2).

Given $\psi_m \in L^1[\mathbb{R}^2, \mathbb{R}_+] \cap L^2[\mathbb{R}^2, \mathbb{R}_+] \cap L^\infty[\mathbb{R}^2, \mathbb{R}_+]$, $\psi_m \not\equiv 0$ the mathematical description of these sets is

$$\mathbb{Q}_m \equiv \left\{ \boldsymbol{u} \in L^2[\mathbb{R}^2, \mathbb{R}^2] \;\mid\; |\mathcal{F}_m[\boldsymbol{u}]| = \psi_m \ a.e. \right\} \tag{3.1}$$

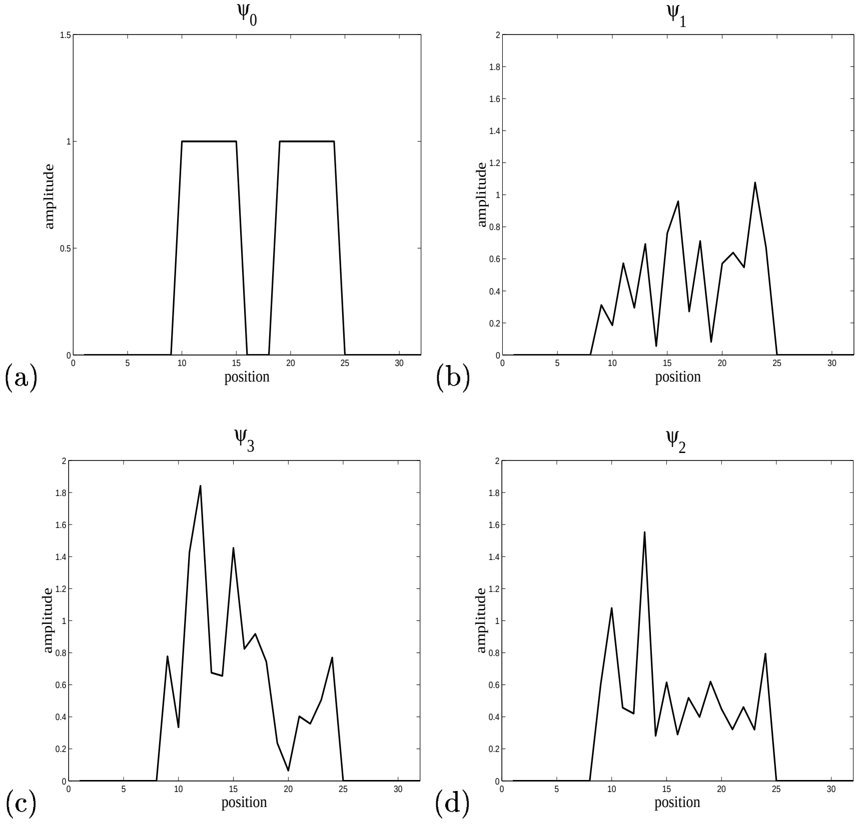


Figure 3.1: One dimensional pupil with corresponding image data. Frame (a) is a one dimensional cross section of the amplitude across the aperture shown in Fig.(6.1.a). Frames (b)-(d) are cross sections of the corresponding point-spread functions for the aperture in frame (a) with some unknown phase aberration as well as a known defocus.

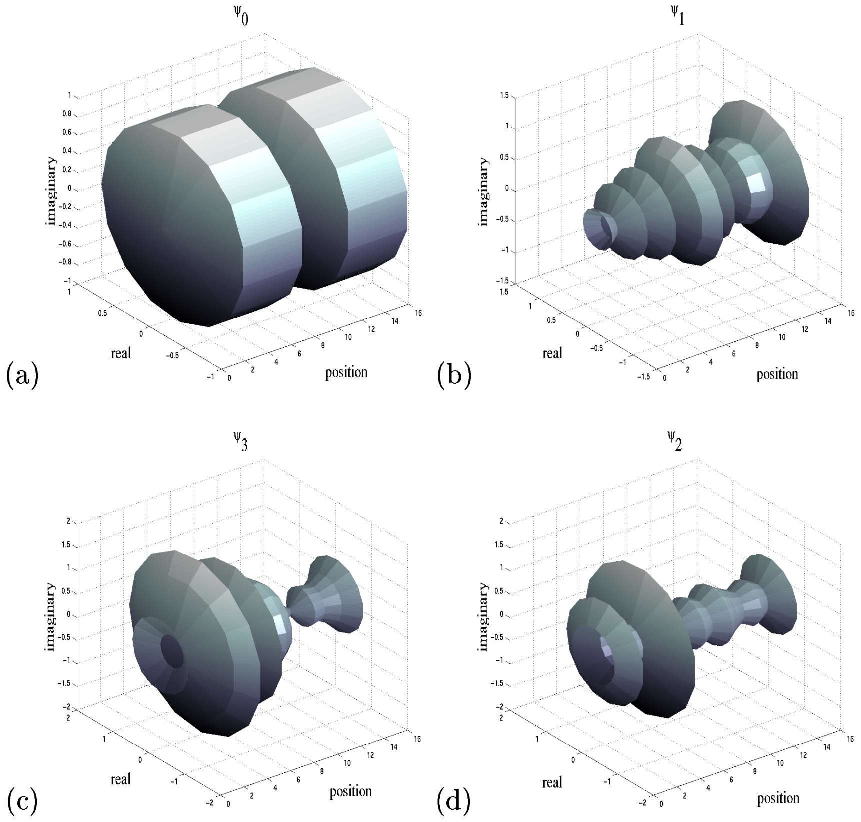


Figure 3.2: Tube constraints. The vertical axis and the axis coming out of the page correspond to the real and imaginary components of the tubes. The horizontal axes correspond to the horizontal axes of Fig.(3.1). Frame (a) represents the constraint set corresponding to Fig.(3.1.a). Frames (b)-(d) represent the constraint sets corresponding to Fig.(3.1.b)-(3.1.d).

Property 3.1.1 *The sets* $\mathbb{Q}_m$ *defined by Eq.(3.1) are neither weakly closed nor convex in* $L^2[\mathbb{R}^2, \mathbb{R}^2]$ *whenever* ψ_m *is not identically zero.*

PROOF: First, we show that the set $\mathbb{Q}_0$ is not convex. If $\boldsymbol{u}$ belongs to $\mathbb{Q}_0$, then so does $\boldsymbol{u}' = -\boldsymbol{u}$. Thus for any non-trivial convex combination of $\boldsymbol{u}$ and $\boldsymbol{u}'$,

$$\boldsymbol{u}'' \equiv \lambda \boldsymbol{u} + (1-\lambda)\boldsymbol{u}' = (2\lambda - 1)\boldsymbol{u},$$

for $\lambda \in (0,1)$, the function $\boldsymbol{u}''$ does not belong to $\mathbb{Q}_0$ since

$$|\boldsymbol{u}''(\boldsymbol{x})| = |(2\lambda-1)|\psi_0(\boldsymbol{x}) < \psi_0(\boldsymbol{x}) \ \forall\ \boldsymbol{x} \text{ such that } \psi_0(\boldsymbol{x}) > 0 \text{ and } \lambda \in (0,1).$$

Next we show that $\mathbb{Q}_0$ is not weakly closed. Choose $\boldsymbol{u} \in \mathbb{Q}_0$ and define the sequence $\{\boldsymbol{u}_n\}$ by

$$\boldsymbol{u}_n(\boldsymbol{x}) = \mathcal{R}^*(\mathcal{R}(\boldsymbol{u}(\boldsymbol{x}))\exp[-2\pi\sqrt{-1}\ \boldsymbol{n}\cdot\boldsymbol{x}])$$

where $\boldsymbol{n} = (n, n)$.

Clearly $\boldsymbol{u}_n \in \mathbb{Q}_0$ for all n. Set

$$\hat{\boldsymbol{u}} = \mathcal{R}^*[\mathcal{R}[\boldsymbol{u}]^\wedge] \quad \text{and} \quad \hat{\boldsymbol{u}}_n = \mathcal{R}^*[\mathcal{R}[\boldsymbol{u}_n]^\wedge].$$

The transformed sequence $\{\hat{\boldsymbol{u}}_n\}$ is related to the Fourier transform of $\mathcal{R}[\boldsymbol{u}]$ by

$$\mathcal{R}[\hat{\boldsymbol{u}}_n] = \left[\mathcal{R}(\boldsymbol{u}(\boldsymbol{x}))\exp[-2\pi\sqrt{-1}\ \boldsymbol{n}\cdot\boldsymbol{x}]\right]^\wedge = \mathcal{R}[\hat{\boldsymbol{u}}](\boldsymbol{\xi}+\boldsymbol{n}),$$

For any $\boldsymbol{u}' \in L^2[\mathbb{R}^2, \mathbb{R}^2]$ the standard inner product in L^2 yields

$$\begin{aligned} \langle \boldsymbol{u}_n, \boldsymbol{u}' \rangle &= \langle \mathcal{R}[\boldsymbol{u}]\exp[-2\pi\sqrt{-1}\ \boldsymbol{n}\cdot\boldsymbol{x}], \mathcal{R}[\boldsymbol{u}'] \rangle \\ &= [\mathcal{R}[\boldsymbol{u}]\overline{\mathcal{R}[\boldsymbol{u}']}]^\wedge(\boldsymbol{n}). \end{aligned}$$

By the Riemann-Lebesgue Lemma [99, page 297]

$$[\mathcal{R}[\boldsymbol{u}]\overline{\mathcal{R}[\boldsymbol{u}']}]^\wedge(\boldsymbol{n}) \to 0 \quad \text{as} \quad \boldsymbol{n} \to \infty.$$

But for all n, $\|\boldsymbol{u}_n\| = \|\boldsymbol{u}\| \neq 0$.

The same properties also hold for the sets $\mathbb{Q}_m$ for $m = 1, 2, \ldots, M$ since $\mathcal{F}_m$ is a bijective linear operator. □

The true generalized pupil function must satisfy all the constraints simultaneously, *i.e.* it lies in the intersection of the sets $\mathbb{Q}_0 \cap \mathbb{Q}_1 \cap \cdots \cap \mathbb{Q}_M$, assuming that this intersection is nonempty. Projection methods are common techniques for finding such intersections in the convex setting. The Gerchberg-Saxton algorithm, discussed later in this section, is a well known projection algorithm that has been successfully applied to the nonconvex problem of phase retrieval. However, due to the nonconvexity of these sets it does not always converge.

We now develop the projection theory for sets of the form Eq.(3.1). Let $\mathbb{X}$ be a metric space with metric $\rho : \mathbb{X} \to \mathbb{R}_+$ and let $\mathbb{Q} \subset \mathbb{X}$. Define the distance of a point $x \in \mathbb{X}$ to the set $\mathbb{Q}$ by

$$\operatorname{dist}(x; \mathbb{Q}) = \inf_{u \in \mathbb{Q}} \rho(x, u). \tag{3.2}$$

We assume that the metric ρ is the Euclidean norm in $\mathbb{R}^n$ and the L^2-norm in L^2. Suppose $\mathbb{Q} \subset \mathbb{X}$ is closed and define the projection operator, $\Pi_{\mathbb{Q}}[v]$, to be the possibly multi-valued mapping that sends every point of $\mathbb{X}$ to the set of nearest points in $\mathbb{Q}$:

$$\Pi_{\mathbb{Q}}[v] = \arg\min_{u\in\mathbb{Q}} \|v-u\| = \{\bar{u} \in \mathbb{Q} : \|v-\bar{u}\| = \inf_{u\in\mathbb{Q}} \|v-u\|\}. \tag{3.3}$$

Proof of the existence of projections in metric spaces is a classical result dating back to the late 1950's [142, 143]. For a survey and bibliography see Ref. [185]. This theory is cited in Ref. [50] with specific application to phase retrieval. Since the sets in the phase retrieval problem are not weakly closed, however, the general theory does not apply [141]. Fortunately, we are able to provide a simple *constructive* proof of existence while at the same time providing a precise formulation of the projections. We construct the projection operator after a brief review of the general theory and its limitations with regard to phase retrieval. The formulation agrees for the most part with what has heretofore been called the projection in the literature. While it is elementary, we are not aware of any other proof of the existence of this specific projection, much less its precise characterization.

The following discussion is limited to a general Hilbert space setting, which is the implied setting for electromagnetic applications.

Definition 3.1.2 *Let $\mathbb{X}$ denote a general Hilbert space.*

(i) A set $\mathbb{Q} \subset \mathbb{X}$ is called boundedly compact *if $\mathbb{Q} \cap \mathbb{B}$ is empty or compact for each closed ball $\mathbb{B}$.*

(ii) A set $\mathbb{Q} \subset \mathbb{X}$ is called approximatively compact *if for any $u \in \mathbb{Q}$ each minimizing sequence $(v_\nu) \subset \mathbb{Q}$ has a subsequence converging to an element of $\mathbb{Q}$.*

(iii) A set $\mathbb{Q} \subset \mathbb{X}$ is said to be proximinal *if every point of $\mathbb{X}$ has at least one projection onto $\mathbb{Q}$.*

Lemma 3.1.3 (Efimov and Stechkin [60]) *Let $\mathbb{Q}$ be a nonempty subset of $\mathbb{X}$. Each of the following implies the next.*

(i) $\mathbb{Q}$ is boundedly compact;

(ii) $\mathbb{Q}$ is approximatively compact;

(iii) $\mathbb{Q}$ is proximinal;

(iv) $\mathbb{Q}$ is closed.

Property 3.1.4 *The sets $\mathbb{Q}_i$ defined by Eq. (3.1) for $i = 0, \dots, M$ are boundedly compact.*

PROOF: This follows from the fact that the unit ball $\mathbb{B}$ is weakly compact in any L^p space, for $1 < p < \infty$. □

Lemma 3.1.3 and Property 3.1.4 are cited as proof of the existence of projections onto sets of the form $\mathbb{Q}_m$. Since the sets are not weakly closed, however, the existence of a minimizing sequence in the definition of approximatively compact sets is an open question. We do not attempt to address this issue here.

We are interested in computing the projection onto sets of the form

$$\mathbb{Q}[b] \equiv \left\{ \boldsymbol{u} \in L^2[\mathbb{R}^2, \mathbb{R}^2] \mid |\boldsymbol{u}| = b \ a.e. \right\}, \tag{3.4}$$

where $b \in L^2[\mathbb{R}^2, \mathbb{R}_+]$ with $b \not\equiv 0$. We show that the projection of $\boldsymbol{u} \in L^2[\mathbb{R}^2, \mathbb{R}^2]$ onto $\mathbb{Q}[b]$ is precisely the set

$$\Pi[\boldsymbol{u}; b] = \{\pi[\boldsymbol{u}; b, \theta] \mid \theta \text{ measurable}\}$$

where the functions $\pi[\boldsymbol{u}; b, \theta] : \mathbb{R}^2 \to \mathbb{R}^2$ are given by

$$\pi[\boldsymbol{u}; b, \theta](\boldsymbol{x}) = \begin{cases} b(\boldsymbol{x}) \frac{\boldsymbol{u}(\boldsymbol{x})}{|\boldsymbol{u}(\boldsymbol{x})|} & \text{for } \boldsymbol{u}(\boldsymbol{x}) \neq 0 \\ b(\boldsymbol{x}) \mathcal{R}^* \exp[\sqrt{-1}\, \theta(\boldsymbol{x})] & \text{for } \boldsymbol{u}(\boldsymbol{x}) = 0 \end{cases}, \tag{3.5}$$

for $\theta : \mathbb{R}^2 \to \mathbb{R}$ Lebesgue measurable. Indeed, the proof shows that the set $\Pi[\boldsymbol{u}; b]$ is precisely the set of all functions in $\mathbb{Q}[b]$ that attain the pointwise distance of $\boldsymbol{u}(\boldsymbol{x})$ to $b(\boldsymbol{x})\mathbb{S}$ a.e. on $\mathbb{R}^2$.

Theorem 3.1.5 *For every $b \in L^2[\mathbb{R}^2, \mathbb{R}_+]$ and $\boldsymbol{u}$, $\boldsymbol{v} \in L^2[\mathbb{R}^2, \mathbb{R}^2]$, we have*

$$\begin{aligned} \boldsymbol{v} \in \Pi[\boldsymbol{u}; b] \quad &\Longleftrightarrow \quad |\boldsymbol{v}(\boldsymbol{x}) - \boldsymbol{u}(\boldsymbol{x})| = \mathit{dist}(\boldsymbol{u}(\boldsymbol{x}); b(\boldsymbol{x})\mathbb{S}) \ a.e., && (3.6) \\ \Pi_{\mathbb{Q}[b]}[\boldsymbol{u}] \quad &= \quad \Pi[\boldsymbol{u}; b], \quad \mathit{and} && (3.7) \\ \mathit{dist}(\boldsymbol{u}; \mathbb{Q}[b]) \quad &= \quad \| \, |\boldsymbol{u}| - b \|. && (3.8) \end{aligned}$$

PROOF: We first show Eq.(3.6). Let $\boldsymbol{u} \in L^2[\mathbb{R}^2, \mathbb{R}^2]$ and $b \in L^2[\mathbb{R}^2, \mathbb{R}_+]$ be given. Observe that if $\pi[\boldsymbol{u}; b, \theta] \in \Pi[\boldsymbol{u}; b]$, then $\pi[\boldsymbol{u}; b, \theta] \in \mathbb{Q}[b]$ and

$$\pi[\boldsymbol{u}; b, \theta](\boldsymbol{x}) \in \underset{\boldsymbol{w} \in b(\boldsymbol{x})\mathbb{S}}{\arg\min} |\boldsymbol{u}(\boldsymbol{x}) - \boldsymbol{w}| \quad \forall \boldsymbol{x} \in \mathbb{R}^2.$$

That is, the function $\pi[\boldsymbol{u}; b, \theta]$ attains the pointwise distance of $\boldsymbol{u}(\boldsymbol{x})$ to the set $b(\boldsymbol{x})\mathbb{S}$ on $\mathbb{R}^2$. Conversely, suppose that $\boldsymbol{v} \in L^2[\mathbb{R}^2, \mathbb{R}^2]$ attains the pointwise distance of $\boldsymbol{u}(\boldsymbol{x})$ to the set $b(\boldsymbol{x})\mathbb{S}$ on $\mathbb{R}^2$. Then by [159, Corollary 1.9.e] there exists a complex measurable function $\alpha : \mathbb{R}^2 \to \mathbb{C}$ such that $|\alpha(\boldsymbol{x})| = 1$ for all $\boldsymbol{x} \in \mathbb{R}^2$ and $\mathcal{R}[\boldsymbol{v}] = \alpha|\boldsymbol{v}|$. Define the measurable function $\theta : \mathbb{R}^2 \to \mathbb{R}$ by $\theta = \cos^{-1}(Re(\alpha))$ where we take the principle branch of $\cos^{-1}$. Then $\alpha = \exp[\sqrt{-1}\, \theta]$. Consequently

$$\boldsymbol{v}(\boldsymbol{x}) = \begin{cases} b(\boldsymbol{x}) \frac{\boldsymbol{u}(\boldsymbol{x})}{|\boldsymbol{u}(\boldsymbol{x})|} & , \ \boldsymbol{u}(\boldsymbol{x}) \neq 0 \\ b(\boldsymbol{x}) \mathcal{R}^*[\exp[\sqrt{-1}\, \theta(\boldsymbol{x})]] & , \ \boldsymbol{u}(\boldsymbol{x}) = 0 \end{cases},$$

which implies that $\boldsymbol{v} \in \Pi[\boldsymbol{u}; b]$. Therefore Eq.(3.6) holds.

We now show that $\Pi[\boldsymbol{u}; b] \subset \Pi_{\mathbb{Q}[b]}[\boldsymbol{u}]$. Choose $\pi[\boldsymbol{u}; b, \theta] \in \Pi[\boldsymbol{u}; b]$ for some Lebesgue measurable $\theta : \mathbb{R}^2 \to \mathbb{R}$, and let $\boldsymbol{v} \in \mathbb{Q}[b]$ with $\boldsymbol{v} \notin \Pi[\boldsymbol{u}; b]$. Clearly, $\pi[\boldsymbol{u}; b, \theta] \in \mathbb{Q}[b]$. Moreover, since $\boldsymbol{v} \notin \Pi[\boldsymbol{u}; b]$ there must exist a set of positive measure $\mathbb{Y} \subset \mathbb{R}^2$ on which $\boldsymbol{v}$ does not attain the pointwise distance of $\boldsymbol{u}(\boldsymbol{x})$ to $b(\boldsymbol{x})\mathbb{S}$, that is,

$$\begin{aligned} |\boldsymbol{u}(\boldsymbol{x}) - \pi[\boldsymbol{u}; b, \theta](\boldsymbol{x})| &= \min_{\boldsymbol{w} \in b(\boldsymbol{x})\mathbb{S}} |\boldsymbol{u}(\boldsymbol{x}) - \boldsymbol{w}| \\ &< |\boldsymbol{u}(\boldsymbol{x}) - \boldsymbol{v}(\boldsymbol{x})|, \qquad \forall \boldsymbol{x} \in \mathbb{Y}. \end{aligned}$$

Therefore,

$$\begin{aligned} \|\boldsymbol{u} - \pi[\boldsymbol{u}; b, \theta]\|^2 &= \int_{\mathbb{R}^2} \min_{\boldsymbol{w} \in b(\boldsymbol{x})\mathbb{S}} |\boldsymbol{u}(\boldsymbol{x}) - \boldsymbol{w}|^2 \, d\boldsymbol{x} \\ &< \int_{\mathbb{R}^2 \setminus \mathbb{Y}} \min_{\boldsymbol{w} \in b(\boldsymbol{x})\mathbb{S}} |\boldsymbol{u}(\boldsymbol{x}) - \boldsymbol{w}|^2 \, d\boldsymbol{x} \; + \; \int_{\mathbb{Y}} |\boldsymbol{u}(\boldsymbol{x}) - \boldsymbol{v}(\boldsymbol{x})|^2 \, d\boldsymbol{x} \\ &\leq \int_{\mathbb{R}^2} |\boldsymbol{u}(\boldsymbol{x}) - \boldsymbol{v}(\boldsymbol{x})|^2 \, d\boldsymbol{x} \\ &= \|\boldsymbol{u} - \boldsymbol{v}\|^2, \end{aligned}$$

where the strict inequality follows from the fact that the set $\mathbb{Y}$ has positive measure and $\boldsymbol{v} \notin \Pi[\boldsymbol{u}; b]$. Hence $\pi[\boldsymbol{u}; b, \theta] \in \Pi_{\mathbb{Q}[b]}[\boldsymbol{u}]$.

Conversely, if $\boldsymbol{v} \in \Pi_{\mathbb{Q}[b]}[\boldsymbol{u}]$, then, in particular $\boldsymbol{v} \in \mathbb{Q}[b]$. If $\boldsymbol{v} \notin \Pi[\boldsymbol{u}; b]$, then, as above, there is a set of positive measure on which $\boldsymbol{v}$ does not attain the pointwise distance to the set $b(\boldsymbol{x})\mathbb{S}$ which implies the contradiction $\|\boldsymbol{u} - \pi[\boldsymbol{u}; b, \theta]\| < \|\boldsymbol{u} - \boldsymbol{v}\|$ for any function $\pi[\boldsymbol{u}; b, \theta] \in \Pi[\boldsymbol{u}; b]$. Thus we have established Eq.(3.7).

We now show Eq.(3.8). Choose $\pi[\boldsymbol{u}; b, \theta]$ from $\Pi_{\mathbb{Q}[b]}[\boldsymbol{u}]$. Then

$$\begin{aligned} \operatorname{dist}^2(\boldsymbol{u}; \mathbb{Q}[b]) &= \|\boldsymbol{u} - \pi[\boldsymbol{u}; b, \theta]\|^2 \\ &= \int |\, \boldsymbol{u}(\boldsymbol{x}) - \pi[\boldsymbol{u}; b, \theta](\boldsymbol{x})|^2 \; d\boldsymbol{x} \\ &= \int \left| (|\boldsymbol{u}(\boldsymbol{x})| - b(\boldsymbol{x})) \, \frac{\boldsymbol{u}(\boldsymbol{x})}{|\boldsymbol{u}(\boldsymbol{x})|} \mathcal{X}_{\mathrm{supp}\,(\boldsymbol{u})}(\boldsymbol{x}) \right|^2 \; d\boldsymbol{x} \\ &\qquad\qquad + \int \left| \, b(\boldsymbol{x}) \mathcal{R}^* \left[\exp[\sqrt{-1}\, \theta(\boldsymbol{x})] \right] \right|^2 \left(1 - \mathcal{X}_{\mathrm{supp}\,(\boldsymbol{u})}(\boldsymbol{x})\right) \; d\boldsymbol{x} \\ &= \int | \; |\boldsymbol{u}(\boldsymbol{x})| - b(\boldsymbol{x})|^2 \mathcal{X}_{\mathrm{supp}\,(\boldsymbol{u})}(\boldsymbol{x}) + |b(\boldsymbol{x})|^2 \left(1 - \mathcal{X}_{\mathrm{supp}\,(\boldsymbol{u})}(\boldsymbol{x})\right) d\boldsymbol{x} \\ &= \|\, |\boldsymbol{u}| - b\|^2 \end{aligned}$$

□

As an elementary consequence of Theorem 3.1.5 we are also able to characterize the projection onto the sets $\mathbb{Q}_m$ defined in Eq.(3.1). In order to define $\Pi_{\mathbb{Q}_m}$, we must first define transforms of multi-valued mappings. The definitions employed here are stated in terms of measure spaces in anticipation of Section 3.2.5. Here and throughout multi-valued mappings are indicated with a double arrow $\rightrightarrows$. Let $(\Omega, \mathcal{A}, \mu)$ be a measure space with the general σ-algebra $\mathcal{A}$ and measure μ. Let f be a function of Ω into $\mathbb{R}^n$, *i.e.* $f = (f_1, \ldots, f_n)$, $f_i : \Omega \to \mathbb{R}$, for $i = 1, \ldots, n$. The integral $\int f d\mu$ is defined by the vector

$$\left(\int f_1, d\mu, \ldots, \int f_n d\mu \right). \tag{3.9}$$

When the measure space is a Lebesgue measure space this will be denoted by $(\Omega, \mathcal{M}, \mu_L)$ where $\mathcal{M}$ is the Lebesgue σ-algebra and μ_L is Lebesgue measure [99]. The σ-algebra of Borel sets is denoted by $\mathcal{B}$.

Let $F : \Omega \rightrightarrows \mathbb{R}^n$. Denote by $\mathcal{S}(F)$ the set of μ-integrable functions $f : \Omega \to \mathbb{R}^n$ that satisfy $f(x) \in F(x)$ a.e. in Ω $(x \in \Omega)$. We call the set $\mathcal{S}(F)$ *integrable* selections of F.

Definition 3.1.6 (Integrals of multi-valued functions) *The set*

$$\left\{ \int f d\mu \mid f \in \mathcal{S}(F) \right\}$$

is the integral *of the multi-valued mapping* $F : \Omega \rightrightarrows \mathbb{R}^n$ *and is denoted by* $\int F d\mu$ *or* $\int F$.

For $b \in L^1[\mathbb{R}^2, \mathbb{R}_+] \cap L^2[\mathbb{R}^2, \mathbb{R}_+]$, by Theorem 3.1.5

$$\Pi_{\mathbb{Q}[b]}[\boldsymbol{u}] = \mathcal{S}\left(b(\cdot)\Pi_{\mathbb{S}}(\boldsymbol{u}(\cdot))\right) \tag{3.10}$$

where the projection $\Pi_{b\mathbb{S}} : \mathbb{R}^2 \rightrightarrows \mathbb{R}^2$ onto the sphere of radius b, $b\mathbb{S}$ is defined by

$$\Pi_{b\mathbb{S}}(\boldsymbol{v}) = \begin{cases} b\frac{\boldsymbol{v}}{|\boldsymbol{v}|} & \text{for } \boldsymbol{v} \neq 0 \\ b\mathbb{S} & \text{for } \boldsymbol{v} = 0 \end{cases}. \tag{3.11}$$

The $\mathcal{F}_m$-transform of $\Pi_{\mathbb{Q}[b]}[\boldsymbol{u}]$ is thus the $\mathcal{F}_m$-transform of all $\boldsymbol{v} \in \Pi_{\mathbb{Q}[b]}[\boldsymbol{u}]$ and is written $\mathcal{F}_m\left[\Pi_{\mathbb{Q}[b]}[\boldsymbol{u}]\right]$.

Corollary 3.1.7 *Let the set* $\mathbb{Q}_m$ *be defined as in Eq.(3.1) and let the operators* $\Pi_{\mathbb{Q}_m}$ *and* $\mathcal{F}_m$ *be as defined in Eq.(3.3) and Eq.(2.43), respectively. Then*

$$\Pi_{\mathbb{Q}_m}[\boldsymbol{u}] = \mathcal{F}_m^*\left[\Pi_{\mathbb{Q}[\psi_m]}[\mathcal{F}_m[\boldsymbol{u}]]\right] \tag{3.12}$$

and

$$dist(\boldsymbol{u}; \mathbb{Q}_m) = \| \, |\mathcal{F}_m[\boldsymbol{u}]| - \psi_m \|$$

for all $\boldsymbol{u} \in L^2[\mathbb{R}^2, \mathbb{R}^2]$.

PROOF: Since the operator $\mathcal{F}_m$ is unitary and surjective, we have

$$\begin{aligned} \inf_{\boldsymbol{w} \in \mathbb{Q}_m} \|\boldsymbol{u} - \boldsymbol{w}\| &= \inf_{\boldsymbol{w} \in \mathbb{Q}_m} \|\mathcal{F}_m[\boldsymbol{u}] - \mathcal{F}_m[\boldsymbol{w}]\| \\ &= \inf_{\boldsymbol{v} \in \mathbb{Q}[\psi_m]} \|\mathcal{F}_m[\boldsymbol{u}] - \boldsymbol{v}\|. \end{aligned}$$

Therefore,

$$\begin{gathered}\boldsymbol{w}' \in \underset{\boldsymbol{w}\in\mathbb{Q}_m}{\arg\min}\|\boldsymbol{u}-\boldsymbol{w}\| \\ \Longleftrightarrow \\ \mathcal{F}_m[\boldsymbol{w}'] \in \Pi_{\mathbb{Q}[\psi_m]}[\mathcal{F}_m[\boldsymbol{u}]] \\ \Longleftrightarrow \\ \boldsymbol{w}' \in \mathcal{F}_m^*\left[\Pi_{\mathbb{Q}[\psi_m]}[\mathcal{F}_m[\boldsymbol{u}]]\right],\end{gathered}$$

since $\mathcal{F}_m^* = \mathcal{F}_m^{-1}$.

Finally, since $\mathcal{F}_m$ is unitary, we obtain from Theorem 3.1.5 that

$$\begin{aligned}\operatorname{dist}(\boldsymbol{u};\mathbb{Q}_m) &= \|\boldsymbol{u}-\mathcal{F}_m^*\left[\pi[\mathcal{F}_m[\boldsymbol{u}];\psi_m,\theta]\right]\| \\ &= \|\mathcal{F}_m[\boldsymbol{u}]-\pi[\mathcal{F}_m[\boldsymbol{u}];\psi_m,\theta]\| \\ &= \|\,|\mathcal{F}_m[\boldsymbol{u}]|-\psi_m\|\end{aligned}$$

for any $\pi[\mathcal{F}_m[\boldsymbol{u}];b,\theta] \in \Pi_{\mathbb{Q}[\psi_m]}[\mathcal{F}_m[\boldsymbol{u}]]$. □

3.1.2 Projection Algorithms

A general framework for projection algorithms can be found in Ref. [21] which considers sequences of weighted relaxed projections of the form

$$\boldsymbol{u}^{(\nu+1)} \in \left(\sum_{m=0}^{M}\gamma_m^{(\nu)}\left[(1-\alpha_m^{(\nu)})\mathcal{I}+\alpha_m^{(\nu)}\Pi_{\mathbb{Q}_m}\right]\right)[\boldsymbol{u}^{(\nu)}]. \tag{3.13}$$

Here $\mathcal{I}$ is the identity mapping, $\alpha_m^{(\nu)}$ is a relaxation parameter usually in the interval $[0,2]$, and the weights $\gamma_m^{(\nu)}$ are non-negative scalars summing to one. General results for these types of algorithms apply only to convex sets. In the convex setting the inclusion in Alg.(3.13) is an equality since projections onto convex sets are single-valued. In the nonconvex setting this is not the case.

The Gerchberg-Saxton algorithm [73] and its variants can be viewed as an instance of Alg.(3.13). To see this, define the set of *active indices* at iteration ν by

$$\mathbb{J}^{(\nu)} \equiv \{j\in 0,\dots,M \mid \gamma_m^{(\nu)}>0\}.$$

Index m is *active at iteration* ν if $\gamma_m^{(\nu)}>0$, *i.e.* if $m\in\mathbb{J}^{(\nu)}$. Suppose $\mathbb{J}^{(\nu)}$ consists of the single element $\{\nu \bmod (M+1)\}$ for $\nu\geq 0$. In this case the weights $\gamma_m^{(\nu)}$ are given by

$$\gamma_m^{(\nu)} = \begin{cases}1 & \text{if } m\in\mathbb{J}^{(\nu)},\ \textit{i.e.} \text{ if } m=\nu \mod (M+1) \\ 0 & \text{otherwise}\end{cases}, \qquad m=0,1,\dots,M.$$

This is an instance of what is called a *cyclic* projection algorithm [21]. Projections onto the sets $\mathbb{Q}_m$ are calculated one at a time in a sequential manner. Thus $M+1$ iterations of this cyclic algorithm are the same as one iteration of the following sequential projection algorithm, known in the optics community as the *iterative transform algorithm*:

$$\boldsymbol{u}^{(\nu+1)} \in \left(\prod_{m=0}^{M}\left[(1-\alpha_m^{(\nu)})\mathcal{I}+\alpha_m^{(\nu)}\Pi_{\mathbb{Q}_m}\right]\right)[\boldsymbol{u}^{(\nu)}]. \tag{3.14}$$

The Gerchberg-Saxton algorithm [73] is obtained by setting $M = 1$ and $\alpha = 1$. Variants of this algorithm [117, 125] involve increasing the number of diversity images, *i.e.* $M > 1$, and adjusting the relaxation parameters $\alpha_m^{(\nu)}$. Convergence results often cited for the Gerchberg-Saxton algorithm refer to the observation that the set distance error, defined as the sum of the distances of an iterate $\boldsymbol{u}^{(\nu)}$ to two constraint sets, $\mathbb{Q}_0$ and $\mathbb{Q}_1$, will not increase as the iteration proceeds [68, 73, 110]. For $M > 1$, this may not be the case. That is, the set distance error can *increase*. In all cases, the algorithm may fail to converge due to the nonconvexity of the sets $\mathbb{Q}_m$ (see Levi and Stark [110] for an example of this behavior).

In our analysis it is convenient to use the change of variables

$$\lambda^{(\nu)}\beta_m^{(\nu)} \equiv \gamma_m^{(\nu)}\alpha_m^{(\nu)} \tag{3.15}$$

to rewrite Alg.(3.13) as

$$\boldsymbol{u}^{(\nu+1)} \in \left(\mathcal{I}-\lambda^{(\nu)}\mathcal{G}^{(\nu)}\right)[\boldsymbol{u}^{(\nu)}]. \tag{3.16}$$

where for all ν the operators $\mathcal{G}^{(\nu)} : L^2 \to L^2$ are given by

$$\mathcal{G}^{(\nu)} \equiv \sum_{m=0}^{M}\mathcal{G}_m^{(\nu)} \tag{3.17}$$

where

$$\mathcal{G}_m^{(\nu)} \equiv \beta_m^{(\nu)}\left(\mathcal{I}-\Pi_{\mathbb{Q}_m}\right). \tag{3.18}$$

In Alg.(3.16) the non-negative weights $\beta_m^{(\nu)}$ do not necessarily sum to 1, and the parameters $\lambda^{(\nu)}$ are to be interpreted as a *step length*. This formulation of the projection algorithm is shown in Section 3.2 to be equivalent to a steepest descent algorithm for a weighted squared distance function. To our knowledge, the multi-valued nature of the projections has not been adequately addressed in the numerical theory for the phase retrieval problem. Insufficient attention to this detail can result in unstable numerical calculations. This is discussed in Chapter 6. Several authors have proposed extensions to projection algorithms to overcome stagnation [68, 177]. These methods are a valuable topic for further study, however in order to illustrate the comparison between geometric methods and analytic methods studied in the following sections we restrict our attention to simple projection algorithms of the form Alg.(3.13) and Alg.(3.14)

3.1.3 Traps, holes, and monotonic decrease

We finish this discussion of geometric approaches with an illustration of what can go wrong with projection algorithms. We also establish conditions that weights $\gamma_m^{(\nu)}$ and relaxation parameters $\alpha_m^{(\nu)}$ must satisfy to ensure decrease of the set distance error in Alg(3.13) for the case of two sets. The purpose of this discussion is to illustrate the fact that, for the geometric approach, prescriptions for *optimal* parameters are difficult and cumbersome.

Let $T_i^{(\nu)} = (1-\alpha_i^{(\nu)})I + \alpha_i^{(\nu)}\Pi_{\mathbb{Q}_i}^{(\nu)}$. Note that the fixed point of $T_i^{(\nu)}$ is the same as that of $\Pi_{\mathbb{Q}_i}^{(\nu)}$. Letting $T^{(\nu)} = T_1^{(\nu)}T_2^{(\nu)}\cdots T_M^{(\nu)}$ we rewrite (3.14) compactly as

$$\boldsymbol{u}^{(\nu+1)} \in \ T^{(\nu)}[\boldsymbol{u}^{(\nu)}]. \tag{3.19}$$

A point $\boldsymbol{u}$ that is a fixed point of the algorithm, but not in $\mathbb{Q}_0$ is called a *trap*. Here $\boldsymbol{u}$ is a fixed point of the algorithm, but not of the individual operators T_i. This is illustrated in the following example.

Example 3.1.8 *A simple example of a trap (see figure 3.3 below) can be shown by considering a hole in the union of the sets* $\mathbb{Q}_1$, $\mathbb{Q}_2$, *and* $\mathbb{Q}_3$ *in the shape of a triangle. Suppose these nonconvex sets have some non-trivial intersection far away from the hole shown in figure (3.3). For starting points near the interior of the hole the unrelaxed, intermittent singular implementation of Alg. (3.13) converges to a cycle of projections rotating between the sides of the triangle. The unrelaxed, evenly averaged* ($\lambda_i = 1/3 \ \forall \ i$) *version of Alg.(3.13) converges to a point on the interior of the triangle (in this example the point* $(1/4, 1/4)$*). For this example we define the set distance error* $\rho[\cdot, \mathbb{Q}_0] : \ \mathbb{X} \rightarrow \mathbb{R}$ *as the sum of set distance measures,*

$$\rho[\boldsymbol{u}, \mathbb{Q}_0] \ = \ \sum_{i=1}^{M} dist(\boldsymbol{u}, \mathbb{Q}_i), \tag{3.20}$$

where $\mathbb{Q}_0 \ = \ \bigcap_{i=1}^{M} \mathbb{Q}_i$. *The point on the triangle that* locally *minimizes the sum of set distances is the point* $(0, 0)$, *for a minimum value of* $1/\sqrt{2}$. *In the case shown in figure (3.3) both algorithms* diverge *from the locally optimal point causing the set distance error to* increase. *Note, however, that if Alg.(3.14) were started so that the cycle ended at the point* $(0, 0)$ *then the cyclic algorithm would converge to the correct local minimum. Alg. (3.13), on the other hand converges to the local minimum of the* squared *set distance error defined by*

$$\rho^2[\boldsymbol{u}, \mathbb{Q}_0] \ = \ \sum_{i=1}^{M} d^2(\boldsymbol{u}, \mathbb{Q}_i). \tag{3.21}$$

This error is the squared Euclidean norm, whereas the set distance error is the L^1 *norm.*

While in the case of several nonconvex sets the set distance error can increase, Levi and Stark [110] show that when $M = 2$ the sum of set distance measures of successive iterates of algorithm (3.14) will not increase as long as the relaxation parameter $\alpha_i^{(\nu)}$ remains within certain bounds. Viewed in the context of the generalized algorithm (3.13) similar results can be shown to be related to the angle $\phi_{\boldsymbol{u}^{(\nu)}}$ between the sets relative to the iterate $\boldsymbol{u}^{(\nu)}$.

Definition 3.1.9 Let $\mathbb{Q}_1$ and $\mathbb{Q}_2$ be closed, non-empty subsets of a Hilbert space $\mathbb{X}$. Let the projections onto the respective sets be denoted by π_1 and π_2. Consider any point $\boldsymbol{u} \in \mathbb{X}$.

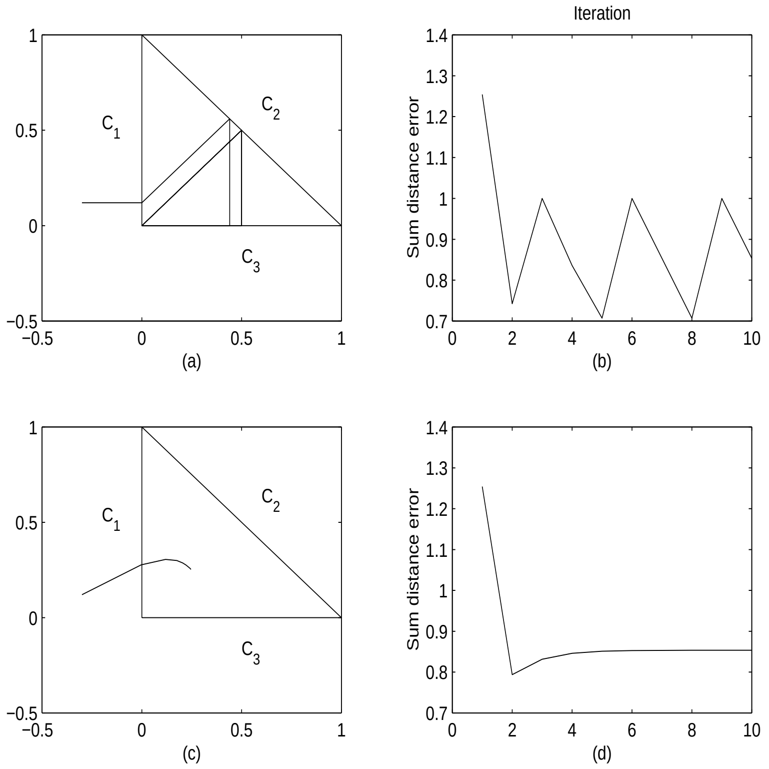


Figure 3.3: Frames (a)-(b) show the behavior of the sequential algorithm in a triangular trap. Frames (c)-(d) show the behavior of the averaged projections algorithm in the same triangular trap. The sum distance error shown in (b) and (d) for the respective algorithms is the performance measure given by equation (3.20).

Let $\boldsymbol{v}_i = \Pi_{\mathbb{Q}_i}[\boldsymbol{u}] - \boldsymbol{u}$ for $i = 1, 2$. Define the angle $\phi_{\boldsymbol{u}}$ between $\mathbb{Q}_1$ and $\mathbb{Q}_2$ relative to $\boldsymbol{u}$ by

$$\cos(\phi_{\boldsymbol{u}}) = \frac{\langle \boldsymbol{v}_1, \boldsymbol{v}_2 \rangle}{\|\boldsymbol{v}_1\| \|\boldsymbol{v}_2\|}, \quad 0 \leq \phi \leq \pi \tag{3.22}$$

The next theorem gives very conservative ranges for the relaxation parameters α_1, α_2 which ensure set distance reduction between iterates $\boldsymbol{u}^{(\nu)}$ and $\boldsymbol{u}^{(\nu+1)}$ of the general algorithm (3.13) whenever $\cos(\phi_{\boldsymbol{u}^{(\nu)}}) \geq 0$.

Theorem 3.1.10 Let $\mathbb{Q}_1$ and $\mathbb{Q}_2$ be closed, non-empty subsets of a Hilbert space $\mathbb{X}$ with intersection $\mathbb{Q}_0 = \mathbb{Q}_1 \bigcap \mathbb{Q}_2$, possibly empty. Denote the projections onto the sets $\mathbb{Q}_1$ and $\mathbb{Q}_2$ by $\Pi_{\mathbb{Q}_1}$ and $\Pi_{\mathbb{Q}_2}$ respectively. Let the set distance error $\rho[\cdot, \mathbb{Q}_0] : \mathbb{X} \to \mathbb{R}$ be given by

$$\rho[\boldsymbol{u}, \mathbb{Q}_0] \;=\; \operatorname{dist}(\boldsymbol{u}, \mathbb{Q}_1)^2 + \operatorname{dist}(\boldsymbol{u}, \mathbb{Q}_2)^2$$

where $\operatorname{dist}(\boldsymbol{u}, \mathbb{Q}_i)$ is given by (3.2). Suppose the set angle relative to the iterate $\boldsymbol{u}^{(\nu)}$ defined by (3.22) satisfies $\cos(\phi_{\boldsymbol{u}^{(\nu)}}) \geq 0$; then for all weights $\lambda_i^{(\nu)} \in [0,1],\ i = 1, 2$ satisfying $\lambda_1^{(\nu)} + \lambda_2^{(\nu)} = 1$, iterates of algorithm (3.13) satisfy

$$\rho[\boldsymbol{u}^{(\nu+1)}, \mathbb{Q}_0] \;\leq\; \rho[\boldsymbol{u}^{(\nu)}, \mathbb{Q}_0] \tag{3.23}$$

for values of the relaxation parameters $\alpha^{(\nu)}$ satisfying

$$0 \;\leq\; \alpha_1^{(\nu)} \;\leq\; 1/\lambda_1^{(\nu)}, \qquad \forall\, \alpha_2^{(\nu)} \in [0,2], \qquad \text{for } \lambda_1^{(\nu)} \in [3/4, 1] \tag{3.24}$$

$$0 \;\leq\; \alpha_1^{(\nu)} \;\leq\; 1/\lambda_1^{(\nu)}, \quad 0 \leq \alpha_2^{(\nu)} \leq \frac{1}{2(1-\lambda_1^{(\nu)})} \qquad \text{for } \lambda_1^{(\nu)} \in [1/2, 3/4] \tag{3.25}$$

$$0 \;\leq\; \alpha_2^{(\nu)} \;\leq\; 1/\lambda_2^{(\nu)}, \quad 0 \leq \alpha_1^{(\nu)} \leq \frac{1}{2(1-\lambda_2^{(\nu)})} \qquad \text{for } \lambda_1^{(\nu)} \in [1/4, 1/2] \tag{3.26}$$

$$0 \;\leq\; \alpha_2^{(\nu)} \;\leq\; 1/\lambda_2^{(\nu)}, \qquad \forall\, \alpha_1^{(\nu)} \in [0,2], \qquad \text{for } \lambda_1^{(\nu)} \in [0, 1/4]. \tag{3.27}$$

PROOF: It suffices to show that for any $\boldsymbol{u} \in \mathbb{X}[\mathbb{R}^n, \mathbb{C}]$ satisfying $\cos(\phi_{\boldsymbol{u}}) \geq 0$ and any α_i satisfying (3.24)-(3.25) that

$$\rho[T[\boldsymbol{u}], \mathbb{Q}_0] \;\leq\; \rho[\boldsymbol{u}, \mathbb{Q}_0]$$

where $T = \lambda_1 T_1 + \lambda_2 T_2$ for $T_i = (1 - \alpha_i)I + \alpha_i \Pi_{\mathbb{Q}_i}$. By the definition of the projection

$$\begin{aligned}
\rho[T[\boldsymbol{u}], \mathbb{Q}_0] \;&=\; \|\Pi_{\mathbb{Q}_1}[T[\boldsymbol{u}]] - T[\boldsymbol{u}]\|^2 + \|\Pi_{\mathbb{Q}_2}[T[\boldsymbol{u}]] - T[\boldsymbol{u}]\|^2 \\
&\leq\; \|\Pi_{\mathbb{Q}_1}[\boldsymbol{u}] - T[\boldsymbol{u}]\|^2 + \|\Pi_{\mathbb{Q}_2}[\boldsymbol{u}] - T[\boldsymbol{u}]\|^2 \\
&=\; \|(1 - \alpha_1\lambda_1)(\Pi_{\mathbb{Q}_1}[\boldsymbol{u}] - \boldsymbol{u}) - \alpha_2\lambda_2(\Pi_{\mathbb{Q}_2}[\boldsymbol{u}] - \boldsymbol{u})\|^2 \\
&\qquad + \|(1 - \alpha_2\lambda_2)(\Pi_{\mathbb{Q}_2}[\boldsymbol{u}] - \boldsymbol{u}) - \alpha_1\lambda_1(\Pi_{\mathbb{Q}_1}[\boldsymbol{u}] - \boldsymbol{u})\|^2
\end{aligned}$$

Let $v_i = \pi_i f - \boldsymbol{u},\ i = 1, 2$. Then

$$\begin{aligned}
\rho[T[\boldsymbol{u}], \mathbb{Q}_0] \;&\leq\; \|(1 - \alpha_1\lambda_1)v_1 - \alpha_2\lambda_2 v_2\|^2 + \|(1 - \alpha_2\lambda_2)v_2 - \alpha_1\lambda_1 v_1\|^2 \\
&=\; \left[(1 - \alpha_1\lambda_1)^2\|v_1\|^2 - 2(1 - \alpha_1\lambda_1)\alpha_2\lambda_2 \mathrm{Re}\{\langle v_1, v_2\rangle\} + \alpha_2^2\lambda_2^2\|v_2\|^2\right] \\
&\quad + \left[(1 - \alpha_2\lambda_2)^2\|v_2\|^2 - 2(1 - \alpha_2\lambda_2)\alpha_1\lambda_1 \mathrm{Re}\{\langle v_1, v_2\rangle\} + \alpha_1^2\lambda_1^2\|v_1\|^2\right]
\end{aligned}$$

Suppose $\|v_1\| > 0$. Let $\Gamma \equiv \frac{\|v_2\|}{\|v_1\|}$. Then $\frac{\mathrm{Re}\langle v_1, v_2\rangle}{\|v_1\|^2} = \Gamma\cos(\phi_{\boldsymbol{u}})$ and

$$\begin{aligned}\rho[T[\boldsymbol{u}], \mathbb{Q}_0] - \rho[\boldsymbol{u}, \mathbb{Q}_0] \;\leq\; & \left[(1-\alpha_1\lambda_1)^2 - 2(1-\alpha_1\lambda_1)\alpha_2\lambda_2\Gamma\cos(\phi_{\boldsymbol{u}}) + \alpha_2^2\lambda_2^2\Gamma^2\right]\|v_1\|^2 \\ + & \left[(1-\alpha_2\lambda_2)^2\Gamma^2 - 2(1-\alpha_2\lambda_2)\alpha_1\lambda_1\Gamma\cos(\phi_{\boldsymbol{u}}) + \alpha_1^2\lambda_1^2\right]\|v_1\|^2 \\ & -(1+\Gamma^2)\|v_1\|^2.\end{aligned}$$

Thus $\rho[T[\boldsymbol{u}], \mathbb{Q}_0] - \rho[\boldsymbol{u}, \mathbb{Q}_0] \leq 0$ holds whenever

$$\alpha_1\lambda_1(\alpha_1\lambda_1 - 1) + \alpha_2\lambda_2(\alpha_2\lambda_2 - 1)\Gamma^2 \leq \Gamma\cos(\phi_{\boldsymbol{u}})(\alpha_1\lambda_1 + \alpha_2\lambda_2 - 2\alpha_1\lambda_1\alpha_2\lambda_2) \qquad (3.28)$$

Suppose that $\cos(\phi_{\boldsymbol{u}}) \geq 0$. By definition $\Gamma \geq 0$, thus the condition (3.28) is clearly satisfied whenever

$$\begin{aligned}(\alpha_1\lambda_1 + \alpha_2\lambda_2 - 2\alpha_1\lambda_1\alpha_2\lambda_2) \;&\geq\; 0 && (3.29)\\ \text{and } \alpha_i \;&\leq\; 1/\lambda_i \;\; i = 1, 2. && (3.30)\end{aligned}$$

There are 4 cases to consider.

1. $\lambda_1 \in [3/4, 1]$: condition (3.30) is satisfied for all $\alpha_2 \in [0, 2]$ and for all $\alpha_1 \in [0, 1/\lambda_1]$ (recall, $\lambda_2 = 1 - \lambda_1$). Condition (3.29) is satisfied whenever $\alpha_2 \leq \frac{1}{2(1-\lambda_1)}$ which is true for all $\alpha_2 \in [0, 2]$.

2. $\lambda_1 \in [1/2, 3/4]$: condition (3.30) is satisfied for all $\alpha_1 \in [0, 1/\lambda_1]$ and $\alpha_2 \in [0, 1/(1-\lambda_1)]$. Condition (3.29), however, is satisfied whenever $\alpha_2 \leq \frac{1}{2(1-\lambda_1)}$, thus the range in (3.25).

The cases when $\lambda_1 \in [1/4, 1/2]$ and $\lambda_1 \in [0, 1/4]$ are treated similarly with the roles of α_1 and α_2 reversed.

The only case left to examine is $\|v_1\| = 0$. If we also have $\|v_2\| = 0$ then the statement of the proof is trivial. Assume then that $\|v_2\| > 0$. Repeating the above argument with the relaxation parameters and weights reversed yields the same result.

□

Remarks: With the exception of the assumption that $\|v_1\| > 0$, condition (3.28) is entirely general and may be used to formulate more refined criteria for achieving set distance reduction than the conservative estimates given in the statement of the proof. Also, the case when $\cos(\phi_{\boldsymbol{u}}) \geq 0$ is not as rare as it might seem. If $\|v_2\| = 0$, *i.e.* $\boldsymbol{u} \in \mathbb{Q}_2$, then $\cos(\phi_{\boldsymbol{u}}) = 0$. This is satisfied when the algorithm simply projects back and forth between the sets. Notice that the value 1 is always in the range of the values that the relaxation parameters may take, thus, as would be expected, simple unrelaxed alternating projections between two sets will never result in an increase in set distance error as long as the initial guess belongs to one of the sets.

3.2 Nonsmooth Analysis

Convergence results for projection methods applied to the phase retrieval problem are not possible in general due to the nonconvexity of the constraint sets. In this section we show that the nonconvexity of the constraint sets is related to the nonsmoothness of the square of the set distance error $\operatorname{dist}(\boldsymbol{u}; \mathbb{Q}_m)$ defined in Eq.(3.2). This is fundamentally different from the convex setting in a Hilbert space where the squared distance function is smooth. The nonsmoothness of the squared distance function in the nonconvex setting is a consequence of the multi-valuedness of the projection operator. In this section some insight into this relationship is given.

3.2.1 Least Squares

Consider the weighted squared set distance error for the phase retrieval problem given by the mapping $E: L^2[\mathbb{R}^2, \mathbb{R}^2] \to \mathbb{R}_+$,

$$E[\boldsymbol{u}] = \sum_{m=o}^{M} \frac{\beta_m}{2} \operatorname{dist}^2(\boldsymbol{u}; \mathbb{Q}_m) \tag{3.31}$$

where $\beta_m \geq 0$ for $m = 0, \ldots, M$ and by Corollary 3.1.7

$$\operatorname{dist}^2(\boldsymbol{u}; \mathbb{Q}_m) \equiv \inf_{\boldsymbol{w} \in \mathbb{Q}_m} \|\boldsymbol{u} - \boldsymbol{w}\|^2 = \| |\mathcal{F}_m[\boldsymbol{u}]| - \psi_m \|^2 . \tag{3.32}$$

For this least squares objective the optimization problem Pr.(2.80) becomes

$$\begin{array}{ll} \text{minimize} & E[\boldsymbol{u}] \\ \text{over} & \boldsymbol{u} \in L^2[\mathbb{R}^2, \mathbb{R}^2]. \end{array} \tag{3.33}$$

In general the optimal value for this problem is non-zero, and so classical techniques for solving the problem numerically are based on satisfying a first-order necessary condition for optimality. For smooth functions this condition simply states that the gradient takes the value zero at any local solution to the optimization problem. However, the functions $\operatorname{dist}^2(\boldsymbol{u}; \mathbb{Q}_m)$ are not differentiable. The easiest way to see this is to consider the one dimensional function $a(x) = | |x| - b|^2$ where $b > 0$. This function is not differentiable at $x = 0$ (indeed, it is not even subdifferentiably regular at x=0 [154, Def. 7.25]). It is precisely at these points that the finite dimensional projection operator $\Pi_{b\mathbb{S}}$ is multi-valued. Similarly, the functions $\operatorname{dist}^2(\boldsymbol{u}; \mathbb{Q}_m)$ are not differentiable at functions $\boldsymbol{u}$ for which there exists a set $\Omega \subset \operatorname{supp}(\psi_m)$ of positive measure on which $\boldsymbol{u} \equiv 0$. A common technique to avoid division by zero is to add a small positive quantity to the denominator of any suspect rational expression. This device is used in [71] to avoid devision by zero in the representation of the derivative of the modulus function. However, this is not a principled approach to the need for approximating the modulus function and its derivatives locally. In the next section we study an alternative approximation to the modulus function itself that posesses excellent global approximation properties.

In the nonsmooth setting the usual first-order optimality condition is replaced by a first-order variational principle of the form

$$0 \in \partial E[\boldsymbol{u}_*]. \tag{3.34}$$

where ∂ denotes a *subdifferential* operator such as those studied in [44, 45, 92, 94, 126, 129]. By Theorem 9.2 of [129], in the Hilbert space setting, the subdifferential of E is the same regardless of the choice of the subdifferential operator studied in the references given above. In this section we develop the tools necessary to prove the following property

Property 3.2.1 *Let $\psi_m : \mathbb{R}^2 \to \mathbb{R}_+$ and $\boldsymbol{u} : \mathbb{R}^2 \to \mathbb{R}^2$ satisfy Hypothesis 2.2.1; let $\Pi_{\mathbb{Q}_m} : L^2 \rightrightarrows \mathbb{Q}_m$ be defined by Eq.(3.3). Then*

$$\partial\left(dist^2(\boldsymbol{u};\mathbb{Q}_m)\right) = 2cl^*\left(\mathcal{I} - \Pi_{\mathbb{Q}_m}\right)[\boldsymbol{u}] \tag{3.35}$$

and

$$\partial E[\boldsymbol{u}] \subset \sum_{m=0}^{M} cl^* \mathcal{G}_m[\boldsymbol{u}] \tag{3.36}$$

where $\mathcal{G}_m$ is defined by Eq.(3.18) and cl^* denotes the weak-star closure.

Remark 3.2.2 *We should note that in a Hilbert-space setting $cl^* = w-cl$ where $w-cl$ denotes the weak closure.*

3.2.2 Finite dimensional nonsmooth analysis

We introduce the theory of nonsmooth analysis in stages, building up from the finite dimensional setting. We will build the theory using the elemental pieces of the squared set distance error E to fix our ideas. The crux of the problem is the pointwise modulus function

$$\kappa(\boldsymbol{u}) \equiv |\boldsymbol{u}| \tag{3.37}$$

or rather, as it turns out, the *negative* of the modulus function. By Theorem 3.1.5, the squared set distance error $\mathrm{dist}^2(\boldsymbol{u};\mathbb{Q}[b])$ of *functions* $\boldsymbol{u}$ to sets $\mathbb{Q}[b]$ defined by Eq.(3.4) is given by the *pointwise* distance defined by Eq.(3.8). In L^2 this has the following integral representation in terms of the square of the pointwise residual $r : \mathbb{R}^2 \times \mathbb{R}_+ \to \mathbb{R}$

$$\mathrm{dist}^2(\boldsymbol{u};\mathbb{Q}[b]) = \int_{\mathbb{R}^2} r^2(\boldsymbol{u}(\boldsymbol{x}); b(\boldsymbol{x}))d\boldsymbol{x}. \tag{3.38}$$

where

$$r(\boldsymbol{u}; b) = \kappa(\boldsymbol{u}) - b. \tag{3.39}$$

Similarly, define $h : L^2[\mathbb{R}^2, \mathbb{R}^2] \to \mathbb{R}$ by

$$h[\boldsymbol{u}; b] = \int_{\mathbb{R}^2} -\kappa(\boldsymbol{u}(\boldsymbol{x}))\, b(\boldsymbol{x})d\boldsymbol{x} \tag{3.40}$$

then

$$\mathrm{dist}^2(\boldsymbol{u};\mathbb{Q}[b]) = \|\boldsymbol{u}\|^2 + \|b\|^2 + 2h[\boldsymbol{u}; b]. \tag{3.41}$$

When the arguments of the functions are themselves *functions*, this is denoted as usual with square brackets.

While $\text{dist}^2(\boldsymbol{u};\mathbb{Q}[b])$ is not smooth, it is straightforward to show that it is Lipschitz continuous on bounded subsets of $L^2[\mathbb{R}^2,\mathbb{R}^2]$. A function $f:\mathbb{X}\to\mathbb{R}$ is *locally Lipschitz near* x if there exists a neighborhood $\mathbb{U}(x)\subset\mathbb{X}$ of x such that

$$|f(z)-f(y)|\leq K\|z-y\|\quad\forall\, z,y\in\mathbb{U}(x)$$

for some $K\geq 0$. For any set $\mathbb{U}\subset X$ over which f is finite-valued, f is said to be *locally Lipschitz on* $\mathbb{U}$ if it is locally Lipschitz near all $x\in\mathbb{U}$. The function is said to be *(globally) Lipschitz on* $\mathbb{U}$ if

$$|f(x)-f(y)|\leq K\|x-y\|\quad\forall\, x,y\in\mathbb{U}$$

Property 3.2.3 *If* $b\in L^2[\mathbb{R}^2,\mathbb{R}_+]$*, then the mapping* $\text{dist}^2(\cdot;\mathbb{Q}[b]):L^2[\mathbb{R}^2,\mathbb{R}^2]\to\mathbb{R}_+$ *is finite-valued and Lipschitz on any bounded subset* $\mathbb{U}\subset L^2[\mathbb{R}^2,\mathbb{R}^2]$ *with Lipschitz constant*

$$K=2M+2\|b\|_2$$

where $M=\sup_{\boldsymbol{u}\in\mathbb{U}}\|\boldsymbol{u}\|$.

PROOF: Let $\mathbb{U}$ be an L^2-bounded subset of L^2 with bound M. We work with the integral representation for the distance function given in Eq.(3.41). Let $\boldsymbol{u},\boldsymbol{w}\in L^2[\mathbb{R}^2,\mathbb{R}^2]$ be such that $\boldsymbol{u},\boldsymbol{u}+\boldsymbol{w}\in\mathbb{U}$. Then

$$\begin{aligned}
&\left|\text{dist}^2(\boldsymbol{u}+\boldsymbol{w};\mathbb{Q}[b])-\text{dist}^2(\boldsymbol{u};\mathbb{Q}[b])\right| \\
&\quad\leq\left|\|\boldsymbol{u}+\boldsymbol{w}\|^2-\|\boldsymbol{u}\|^2\right|+2\int_{\mathbb{R}^2}b(\boldsymbol{x})\left|\left[\,|\boldsymbol{u}(\boldsymbol{x})+\boldsymbol{w}(\boldsymbol{x})|-|\boldsymbol{u}(\boldsymbol{x})|\,\right]\right|\,d\boldsymbol{x} \\
&\quad\leq\left|(\|\boldsymbol{u}+\boldsymbol{w}\|+\|\boldsymbol{u}\|)(\|\boldsymbol{u}+\boldsymbol{w}\|-\|\boldsymbol{u}\|)\right|+2\int_{\mathbb{R}^2}b(\boldsymbol{x})|\boldsymbol{w}(\boldsymbol{x})|\,d\boldsymbol{x} \\
&\quad\leq 2M\|\boldsymbol{w}\|+2\|b\|\|\boldsymbol{w}\|.
\end{aligned}$$

The last inequality makes use of Hölder's Inequality. □

Remark 3.2.4 *Lipschitz continuity of the squared set distance error* E *is a straightforward consequence of Property 3.2.3 and Parseval's relation.*

We now introduce some basic definitions from nonsmooth analysis. In our discussion we allow mappings to have infinite values, thus it is convenient to define the extended reals, *i.e.* $\mathbb{R}\cup\{\infty\}$, by $\overline{\mathbb{R}}$. The *effective domain* of $f:\mathbb{R}^n\to\overline{\mathbb{R}}$, denoted $\text{dom} f\subset\mathbb{R}^n$, is the set on which f is finite. To avoid certain pathological mappings the discussion is restricted to *proper, i.e.* not everywhere infinite, *lower semi-continuous (l.s.c.)* functions. A function $f:\mathbb{X}\to\overline{\mathbb{R}}$ is *lower semi-continuous* at a point $u\in\mathbb{X}$ if

$$\liminf_{u'\to u} f(u')\geq f(u).$$

Upper semi-continuity is similar to the above definition but with the "inf" replaced by a "sup" and the inequality reversed. A function is continuous if it is both lower and upper semi-continuous.

Definition 3.2.5 (Subderivatives [154]) *For a function* $f : \mathbb{R}^m \to \overline{\mathbb{R}}$ *and a point* $\overline{u} \in R^m$ *with* $f(\overline{u})$ *finite,*

(i) the subderivative *function* $df(\overline{u}) : \mathbb{R}^m \to \overline{\mathbb{R}}$ *is defined by*

$$df(\overline{u})(\overline{w}) \equiv \liminf_{\substack{\tau \searrow 0 \\ w \to \overline{w}}} \frac{f(\overline{u} + \tau w) - f(\overline{u})}{\tau};$$

(ii) the regular subderivative *function* $\widehat{df}(\overline{u}) : \mathbb{R}^m \to \overline{\mathbb{R}}$ *is defined by*

$$\widehat{df}(\overline{u})(\overline{w}) \equiv \lim_{\delta \searrow 0} \left(\limsup_{\substack{u \xrightarrow[f]{} \overline{u} \\ \tau \searrow 0}} \left[\inf_{w \in \mathbb{B}(\overline{w}, \delta)} \frac{f(u + \tau w) - f(u)}{\tau} \right] \right).$$

In the above definition it is not assumed that f is continuous, thus the notion of f*-attentive* convergence, denoted by $\xrightarrow[f]{}$:

$$u^{(\nu)} \xrightarrow[f]{} \overline{u} \qquad \Longleftrightarrow \qquad u^{(\nu)} \to \overline{u} \ \text{ with } \ f(u^{(\nu)}) \to f(\overline{u}).$$

Definition 3.2.6 (Subgradients - finite dimensions [154]) *Consider a function* $f : \mathbb{R}^m \to \overline{\mathbb{R}}$ *and a point* $\overline{u} \in \mathbb{R}^m$ *with* $f(\overline{u})$ *finite. For* $v \in \mathbb{R}^m$ *one has*

(i) v *is a* regular subgradient *of* f *at* $\overline{u}$ *if*

$$f(u) \geq f(\overline{u}) + \langle v,\ u - \overline{u} \rangle + o(\|u - \overline{u}\|).$$

We call the set of regular subgradients v *the* regular subdifferential *of* f *at* $\overline{u}$ *and denote this set by* $\widehat{\partial} f(\overline{u})$.

(ii) v *is a* proximal subgradient *of a function* $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ *at* $\overline{u} \in$ *dom* f *if there exist* $\sigma > 0$ *and* $\delta > 0$ *such that*

$$f(u) \geq f(\overline{u}) + \langle v,\ u - \overline{u} \rangle - \frac{1}{2}\sigma |u - \overline{u}|^2 \quad \textit{when} \ |u - \overline{u}| \leq \delta. \tag{3.42}$$

(iii) v *is a* (general) subgradient *of* f *at* $\overline{u}$ *if there are sequences* $u^{(\nu)} \xrightarrow[f]{} \overline{u}$ *and* $v^{(\nu)} \in \widehat{\partial} f(u^{(\nu)})$ *with* $v^{(\nu)} \to v$. *We call the set of (general) subgradients* v *the* (general) subdifferential *of* f *at* $\overline{u}$ *denoted by* $\partial f(\overline{u})$.

(iv) v *is a* horizon, *or* singular, subgradient *of* f *at* $\overline{u}$ *if (ii) holds with the exception that rather than* $v^{(\nu)} \to v$ *one has* $\lambda^{(\nu)} v^{(\nu)} \to v$ *for some sequence* $\lambda^{(\nu)} \searrow 0$. *We call the set of horizon subgradients* v *the* horizon subdifferential *of* f *at* $\overline{u}$ *and denote this set by* $\partial^{\infty} f(\overline{u})$.

(v) v *is a* Clarke subgradient *of* f *at* $\overline{u}$ *if* f *is l.s.c. on a neighborhood of* $\overline{u}$ *and* v *satisfies*

$$\langle v,\ w\rangle \leq \widehat{d}f(\overline{u})(w) \quad \text{for all} \quad w \in \mathbb{R}^m.$$

We call the set of Clarke subgradients v *the* Clarke subdifferential *of* f *at* $\overline{u}$ *and denote this set by* $\overline{\partial} f(\overline{u})$.

(vi) v *is a* Clarke horizon subgradient *of* f *at* $\overline{u}$, *written* $v \in \overline{\partial}^{\ \infty} f(\overline{u})$, *if* f *is l.s.c. on a neighborhood of* $\overline{u}$ *and* v *satisfies*

$$\langle v,\ w\rangle \leq 0 \quad \text{for all} \quad w \in \operatorname{dom} \widehat{d}f(\overline{u}).$$

We call the set of Clarke horizon subgradients v *the* Clarke horizon subdifferential *of* f *at* $\overline{u}$ *and denote this set by* $\overline{\partial}^{\infty} f(\overline{u})$.

Remark 3.2.7 *If* f *is Lipschitz continuous, then* ∂f *is upper semicontinuous (usc), that is, has closed graph [154, Proposition 8.7].*

Remark 3.2.8 *A remark on subdifferential notation for the composition of functions is in order since this will frequently arise in the sequel. If* $g : \mathbb{X} \to \overline{\mathbb{R}}$ *is given as the composition of two functions* $f : \mathbb{Y} \to \overline{\mathbb{R}}$ *and* $h : \mathbb{X} \to \mathbb{Y}$, *i.e.* $g(x) = (f \circ h)(x) = f(h(x))$, *then we write*

$$\partial g(x) = \partial (f \circ h)(x).$$

On the other hand, we write

$$\partial f(h(x))$$

to denote the subdifferential of f *evaluated at* $h(x)$.

A Lipschitz function $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ is said to be *(subdifferentially) regular* at $\overline{u} \in \operatorname{dom} f$ with $\partial f(u) \neq \emptyset$ if

$$\partial f(\overline{u}) = \widehat{\partial} f(\overline{u}).$$

The class of subdifferentially regular functions includes, among others, all strictly differentiable functions and all convex functions [154]. The subdifferential definitions are illustrated with the following important example.

Example 3.2.9 *Let* $\kappa(\boldsymbol{u}) = |\boldsymbol{u}|$ *as defined in Eq.(3.37) and let* $b \in (0, \infty)$. *Since the function* $b\kappa(\boldsymbol{u})$ *is convex it is subdifferentially regular for all* $\boldsymbol{u}$. *It has* $\partial^{\infty}(b\kappa(\boldsymbol{u})) = \{0\}$ *for all* $\boldsymbol{u}$ *and*

$$\partial\left(b\kappa(\overline{\boldsymbol{u}})\right) = b\partial\left(\kappa(\overline{\boldsymbol{u}})\right) = \begin{cases} b\frac{\overline{\boldsymbol{u}}}{|\overline{\boldsymbol{u}}|} & \text{if } \overline{\boldsymbol{u}} \neq 0 \\ b\mathbb{B} & \text{if } \overline{\boldsymbol{u}} = 0 \end{cases} \tag{3.43}$$

where $b\mathbb{B}$ *is the ball of radius* b.

In contrast, the function $-b\kappa(\boldsymbol{u})$ *defined by Eq.(3.37) for* $b \in (0, \infty)$ *is not regular at* 0. *Nevertheless it has* $\partial^{\infty}\left(-b\kappa(\overline{\boldsymbol{u}})\right) = 0$ *for all* $\overline{\boldsymbol{u}}$ *and*

$$\partial\left(-b\kappa(\overline{\boldsymbol{u}})\right) = b\partial\left(-\kappa(\overline{\boldsymbol{u}})\right) = \begin{cases} -b\frac{\overline{\boldsymbol{u}}}{|\overline{\boldsymbol{u}}|} & \textit{if } \overline{\boldsymbol{u}} \neq 0 \\ b\mathbb{S} & \textit{if } \overline{\boldsymbol{u}} = 0 \end{cases} \tag{3.44}$$

where $b\mathbb{S}$ *is the surface of the sphere of radius* b. *The Clarke subdifferential of* $-b\kappa(\boldsymbol{u})$ *is the convex hull of the generalized subdifferential and the Clarke horizon subdifferential is zero:*

$$\overline{\partial}\left(-b\kappa(\boldsymbol{u})\right) = con\ \partial\left(-b\kappa(\boldsymbol{u})\right), \qquad \overline{\partial}^{\infty}\left(-b\kappa(\boldsymbol{u})\right) = con\ \partial^{\infty}\left(-b\kappa(\boldsymbol{u})\right) = 0.$$

PROOF: This is a modification of Exercise 8.27 of [154]. For completeness, this is proven here.

The function κ is globally Lipschitz, convex and proper with $\operatorname{dom}\kappa = \mathbb{R}^n$, so by Corollary 8.11 and Proposition 8.12 of [154] $b\kappa((\boldsymbol{u}))$ is regular with $\partial^{\infty}(b\kappa(\boldsymbol{u})) = \{0\}$ for all $\boldsymbol{u}$ and

$$\partial(b\kappa(0)) = \{\boldsymbol{v} \mid |\boldsymbol{u}|b \geq (\boldsymbol{v}, \boldsymbol{u}) \quad \forall \boldsymbol{u}\} = b\mathbb{B}.$$

For all $\boldsymbol{u} \neq 0$ the modulus function is differentiable in the usual sense, and so $\partial(b\kappa(\boldsymbol{u})) = \{\nabla(b\kappa(\boldsymbol{u}))\} = b\frac{\boldsymbol{u}}{|\boldsymbol{u}|}$.

On the other hand, by Property 8.5 of [154], $\boldsymbol{v} \in \widehat{\partial}\left(-b\kappa(0)\right)$ if an only if there is a smooth $g \leq -b\kappa$ with $\nabla g(0) = \boldsymbol{v}$. Suppose there exists such a g, then $g(0) = 0$ and on some neighborhood $\mathcal{N}$ of 0,

$$g(\tau\boldsymbol{w}) = \nabla g(t\boldsymbol{w}) \cdot \tau\boldsymbol{w} > -|\tau b\boldsymbol{w}|$$

for all $0 \leq |t| \leq |\tau|$ such that $\tau \in \mathcal{N}$. But this contradicts the definition of g.

Calculation of the generalized gradient follows from the definition: $\boldsymbol{v} \in \partial\left(-b\kappa(\overline{\boldsymbol{u}})\right)$ if there is a sequence $\boldsymbol{u}^{(\nu)} \underset{\kappa}{\rightarrow} \overline{\boldsymbol{u}}$ and $\boldsymbol{v}^{(\nu)} \in \widehat{\partial}\left(-b\kappa(\boldsymbol{u}^{(\nu)})\right)$ with $\boldsymbol{v}^{(\nu)} \rightarrow \boldsymbol{v}$. If $\overline{\boldsymbol{u}} \neq 0$ then the function is differentiable in the usual sense and all the definitions coincide. If $\overline{\boldsymbol{u}} = 0$ then the only $\boldsymbol{v}$ for which

$$\liminf_{\substack{\boldsymbol{u}^{(\nu)} \rightarrow \overline{\boldsymbol{u}} \\ \boldsymbol{u}^{(\nu)} \neq 0}} \frac{-b|\boldsymbol{u}^{(\nu)}| - \langle \boldsymbol{v}^{(\nu)}, \boldsymbol{u}^{(\nu)} \rangle}{|\boldsymbol{u}^{(\nu)}|} \geq 0$$

is $\boldsymbol{v}$ satisfying $|\boldsymbol{v}| = b$. Finally, the local boundedness of $\partial(-b\kappa)$ implies that $\partial^{\infty}\left(-b\kappa(\overline{\boldsymbol{u}})\right) = \{0\}$ by the definition of horizon subgradients.

The second statement of the lemma is a consequence of Theorem 8.49 of [154]. □

The above example immediately yields the following correspondence between finite dimensional projections $\Pi_{b\mathbb{S}}$ and the subdifferential $\partial(-b\kappa(\boldsymbol{u}))$.

Property 3.2.10 *Let* $\Pi_{b\mathbb{S}}(\boldsymbol{u})$ *be the projection defined in Eq.(3.11). For* $\boldsymbol{u} \in \mathbb{R}^2$, $b \in \mathbb{R}_+$ *and* $r^2 : \mathbb{R}^2 \rightarrow \mathbb{R}_+$ *defined in Eq.(3.39)*

$$\partial(-b\kappa(\boldsymbol{u})) = \Pi_{b\mathbb{S}}(\boldsymbol{u}), \tag{3.45}$$

and

$$\overline{\partial}(-b\kappa(\boldsymbol{u})) = con\ (\Pi_{b\mathbb{S}}(\boldsymbol{u})), \tag{3.46}$$

and

$$\partial r^2(\boldsymbol{u}; b) = 2(I - \Pi_{b\mathbb{S}}(\boldsymbol{u})), \tag{3.47}$$

where I is the finite dimensional identity operator. Moreover,

$$\overline{\partial} r^2(\boldsymbol{u}; b) = con\ (I - \Pi_{b\mathbb{S}}(\boldsymbol{u})). \tag{3.48}$$

PROOF: This follows directly from the definitions of subgradients and Example 3.2.9. □

As with the finite dimensional projection $\Pi_{\mathbb{S}b}$ and the infinite dimensional projection $\Pi_{\mathbb{Q}[b]} : L^2[\mathbb{R}^2, \mathbb{R}^2] \rightrightarrows L^2[\mathbb{R}^2, \mathbb{R}^2]$ defined in Eq.(3.7), there is a relation between the finite dimensional subdifferential $\overline{\partial} r^2(\boldsymbol{u}(\boldsymbol{x}); b(\boldsymbol{x}))$ ($\boldsymbol{x}$ fixed) and the "subdifferential" of the square distance function, $\partial(\text{dist}^2(\boldsymbol{u}; \mathbb{Q}[b]))$. In infinite dimensions there are many definitions for subdifferentials which are designed with some more exotic spaces in mind. Fortunately, in the setting of phase retrieval many of the different subdifferentials are *equivalent*. Thus we can choose the simplest object to work with for proving the desired results. Our goal in what follows is to introduce the definitions and theorems necessary to prove a general result relating the subdifferentials of an integrand to the subdifferential of the associated integral operator.

3.2.3 Integrals of multi-valued functions

We begin by reviewing some fundamental properties of integrals of multi-valued mappings defined in Def. 3.1.6. Integrals of multi-valued mappings received a great deal of attention during the 1960's and early 70's in the economics literature. The literature is vast. The principle sources which we draw upon are the works of Aumann [8–10], Richter [150], Schmeidler [161, 162]. An introductory review and bibliography can be found in Hildenbrand [89].

Many of the properties we develop are limited to non-atomic measure spaces. A subset $\mathbb{U}$ of Ω is called an *atom* in the measure space $(\Omega, \mathcal{A}, \mu)$, where $\mathcal{A}$ denotes a general $\sigma-$algebra, if $\mu(\mathbb{U}) > 0$ and if $\mathbb{V} \subset \mathbb{U}$ implies that either $\mu(\mathbb{V}) = \mu(\mathbb{U})$ or $\mu(\mathbb{V}) = 0$. A measure space $(\Omega, \mathcal{A}, \mu)$ or the measure μ on $(\Omega, \mathcal{A})$ is called *non-atomic* if $(\Omega, \mathcal{A}, \mu)$ has no atoms. The following lemma is elementary and stated without proof. For a discussion see [89].

Lemma 3.2.11 (Non-atomic measures on separable metric spaces) *A measure μ on a separable metric space $\mathbb{X}$ is non-atomic if and only if $\mu(x) = 0$ for every $x \in \mathbb{X}$.*

Lebesgue measure, for example, is non-atomic.

The key property of integrals of multi-valued mappings is that they are convex. The following theorem on convexity is fundamental to subsequent results on the correspondence between pointwise multi-valued mappings and integrals of multi-valued mappings on a measure space.

Theorem 3.2.12 (Liapunov's Theorem [111]) *Let* μ_i $(i = 1, 2, \dots)$ *be non-atomic measures on* $(\Omega, \mathcal{A})$*. Then the set*

$$\{(\mu_1(E), \dots, \mu_n(E)) \in \mathbb{R}^n \mid E \in \mathcal{A}\}$$

is a closed and convex subset in $\mathbb{R}^n$*.*

The following result is an elegant application of Liapunov's Theorem to integrals of multi-valued functions defined by Eq.(3.9).

Theorem 3.2.13 (Richter [150]) *Let* $F : \Omega \rightrightarrows \mathbb{R}^n$ *be a multi-valued mapping of the non-atomic measure space* $(\Omega, \mathcal{A}, \mu)$ *into* $\mathbb{R}^n$*. Then the integral*

$$\int F d\mu$$

is a convex set in $\mathbb{R}^n$

PROOF: An English translation of this simple proof can be found in [89]. The proof is repeated here for completeness. Let ϕ_1, $\phi_2 \in \int F$ and $0 < \lambda < 1$. Let $\mathcal{S}(F)$ denote the collection of measurable selections from F. There are integrable functions $f_1, f_2 \in \mathcal{S}(F)$ such that $\phi_i = \int f_i$. From Liapunov's Theorem 3.2.12 the set

$$\left\{ \left(\int_E f_1 d\mu, \int_E f_2 d\mu \right) \in \mathbb{R}^{2n} \mid E \in \mathcal{A} \right\}$$

is convex. Since $(0, 0)$ and (ϕ_1, ϕ_2) belong to this set, there exists a set E such that

$$(\lambda\phi_1, \lambda\phi_2) = \left(\int_E f_1 d\mu, \int_E f_2 d\mu \right).$$

Define the function $f \in \mathcal{S}(F)$ by $f(x) = f_1(x)$ if $x \in E$ and $f(x) = f_2(x)$ if $x \notin E$. Then $\int f = \lambda\phi_1 + (1 - \lambda)\phi_2$. □

Corollary 3.2.14 (Schmeidler [161]) *Let* $(\Omega, \mathcal{A}, P)$ *be a non-atomic probability measure space and let* $\mathbb{S} \subset \mathbb{R}^n$*. Let* $F(x) \equiv \mathbb{S}$ *for every* $x \in \Omega$*. Then*

$$\int F \, dP = con \, \mathbb{S}.$$

Let $\mathcal{A}$ denote a general σ-algebra on Ω. Let $\mathcal{B}^n$ denote the Borel sets of $\mathbb{R}^n$ and let $(\mathcal{A} \otimes \mathcal{B}^n)$ be the product σ-algebra generated by $\mathcal{A}$ and $\mathcal{B}^n$ [99, Chapter 11]. The multifunction $F : \Omega \rightrightarrows \mathbb{R}^n$ is said to be $\mathcal{A}$-measurable if for all open sets $\mathbb{V}$ the set

$$\{x \mid \mathbb{V} \cap F(x) \neq \emptyset\} \in \mathcal{A}.$$

The multifunction F is said to be $(\mathcal{A} \otimes \mathcal{B}^n)$*-measurable* if

$$\text{gph}\,(F) = \{(x, v) \mid v \in F(x)\} \in (\mathcal{A} \otimes \mathcal{B}^n).$$

For example, if $\mathcal{A}$ is the set of Lebesgue measurable set on $\mathbb{R}^m$, $\mathcal{M}^m$, and F is usc, that is F has closed graph, then F is $(\mathcal{M}^m \otimes \mathcal{B}^n)$-measurable.

We say that F is *integrably bounded* if there is a μ-integrable $a : \Omega \to \mathbb{R}^n_+$ such that

$$(|v_1|, \ldots, |v_n|) \leq a(x)$$

for all pairs $(x, v) \in (\Omega, \mathbb{R}^n)$ satisfying $v \in F(x)$. The multifunction F is said to be L^2-bounded if there is a function $a(x) \in L^2[\Omega, \mathcal{A}, \mu]$ such that

$$|v| \leq a(x) \ \forall v \in F(x), \ \textit{a.e.}$$

A multifunction F is *closed* if $F(x)$ is closed for each x.

To fix these ideas, let $b \in L^1[\mathbb{R}^2, \mathbb{R}_+] \cap L^2[\mathbb{R}^2, \mathbb{R}_+]$ be a nonnegative Borel measurable function satisfying

$$\int_{\mathbb{R}^n} b(\boldsymbol{x}) d\boldsymbol{x} = 1.$$

Then b is a *density* function and characterizes the probability measure P on the space $(\mathbb{R}^2, \mathcal{M}^2)$ where and $\mathcal{M}^2$ is the Lebesgue σ-algebra on $\mathbb{R}^2$ (Theorem 12.1 of [97]). The projections discussed in Section 3.1 can be interpreted as multifunctions on this measure space. Consider the multifunction $F : \mathbb{R}^2 \rightrightarrows \mathbb{R}^2$ on the probability measure space $(\mathbb{R}^2, \mathcal{M}^2, P)$

$$F(\boldsymbol{x}) = \mathbb{S} \tag{3.49}$$

where $\mathbb{S}$ is the unit sphere in $\mathbb{R}^2$. Then $\mathcal{S}(F)$ corresponds to $\Pi_{\mathbb{Q}[b]}[0]$ where $\mathcal{S}(F)$ is the collection of P-measurable selections on $(\mathbb{R}^2, \mathcal{M}^2, P)$ and $\Pi_{\mathbb{Q}[b]}[0]$ is given by Eq.(3.7). The multifunction F is closed because it is the pointwise mapping to the unit sphere and P-bounded with bound 1.

Let $L^2_m(\mathbb{R}^n, \mathcal{M}^n, P)$ denote the Hilbert space of functions mapping $\mathbb{R}^n$ to $\mathbb{R}^m$ with inner product on the probability measure $(\mathbb{R}^n, \mathcal{M}^n, P)$ given by

$$\langle f, g \rangle_P = \int_{\mathbb{R}^n} (f(\boldsymbol{x}), g(\boldsymbol{x})) b(\boldsymbol{x}) d\boldsymbol{x} \tag{3.50}$$

where $(\cdot, \cdot)$ denotes the usual finite dimensional vector inner product and $b : \mathbb{R}^n \to \mathbb{R}_+$ is a density function characterizing the probability measure P.

Proposition 3.2.15 *Let $F : \mathbb{R}^n \rightrightarrows \mathbb{R}^m$ be closed, convex-valued and L^2-bounded on $L^2_m(\mathbb{R}^n, \mathcal{M}^n, \mu)$ where μ is a complete, non-atomic measure whose set of measurable sets is $\mathcal{M}^n$. Then the set of measurable selections $\mathcal{S}(F)$ is weakly compact in $L^2_m(\mathbb{R}^n, \mathcal{M}^n, \mu)$.*

PROOF: This is a generalization of Exercise 5.14 of [45]. By [53, Theorem 1, pg. 58] we need only show that $\mathcal{S}(F)$ is weakly sequentially compact. Consider any sequence $\{f_i\} \subset \mathcal{S}(F)$. We must show that $\{f_i\}$ has a weakly convergent subsequence with limit $f_* \in \mathcal{S}(F)$. Since the sequence is L^2-bounded, reflexivity and Alaoglu's Theorem [199, Theorem 1, pg 126] imply that there exists a weakly convergent subsequence whose limit belongs to the weak closure of $\mathcal{S}(F)$. Since $\mathcal{S}(F)$ is convex by the pointwise convexity of F, the strong and weak closures of $\mathcal{S}(F)$ coincide. Hence the result follows if $\mathcal{S}(F)$ is strongly closed. Since strong convergence implies the existence of a subsequence that is almost everywhere pointwise convergent [159, Theorem 3.12], and $F(x)$ is pointwise closed, we have that $\mathcal{S}(F)$ is strongly closed. □

The next theorem, due to Hildenbrand [89], is a restatement of Theorems 3 and 4 of Aumann [8] for multifunctions on the non-atomic measure space $(\Omega, \mathcal{A}, \mu)$.

Theorem 3.2.16 (Theorem 4 and Proposition 7 of [89]) *Let F denote a multifunction from the non-atomic measure space $(\Omega, \mathcal{A}, \mu)$ to $\mathbb{R}^n$ that is $(\mathcal{A} \otimes \mathcal{B}^n)$-measurable and integrably bounded. Then*

$$\int F = \int \mathit{con}\ F.$$

Moreover, if F is closed and integrably bounded (not necessarily $(\mathcal{A} \otimes \mathcal{B}^n)$-measurable), then $\int F$ is compact.

Proposition 3.2.17 *Let $v \in \mathcal{S}(\mathit{con}\ F)$ where $F : \mathbb{R}^2 \rightrightarrows \mathbb{R}^2$ is a nonempty, closed, $(\mathcal{M}^2 \otimes \mathcal{B}^2)$-measurable, L^2-bounded multifunction on $L_2^2(\mathbb{R}^2, \mathcal{M}^2, P)$ for the probability measure P defined by the density $b : \mathbb{R}^2 \to \mathbb{R}_+$ as in Eq.(3.50). Then there exists a sequence $\{f_i\}$ of measurable selections of F which converges weakly to v.*

PROOF: This is a modification of Exercise 5.17 of [45]. Consider the box $\mathbb{I}_n = [-n, n] \times [-n, n]$ for $n = 1, 2, 3, \ldots$. Suppose each box $\mathbb{I}_n$ is partitioned into n^2 intervals of width $1/n$. Set

$$t_k^n = \frac{k}{n} - n \quad \text{for } k = 0, 1, \ldots, 2n^2,$$

and for each $t \in [-n, n]$ define

$$\overline{(t)}_n = \max\{t_k^n : t_k^n \leq t,\ k = 0, \ldots, 2n^2\} \quad \text{and} \quad \underline{(t)}_n = \min\{t_k^n : t_k^n \geq t,\ k = 0, \ldots, 2n^2\}.$$

Note that $0 < \max\{t - \overline{(t)}_n, \underline{(t)}_n - t\} \leq 1/n$ whenever $t \in [-1, n]$. By Theorem 3.2.16 there exists a selection $f_n \in F$ on $(\mathbb{R}^2, \mathcal{B}^2, P)$ corresponding to the partition of the box I_n such that

$$\int_{\mathbb{R}^2} f_n(\boldsymbol{x}) b(\boldsymbol{x}) d\boldsymbol{x} = \int_{\mathbb{R}^2} v(\boldsymbol{x}) b(\boldsymbol{x}) d\boldsymbol{x}$$

with

$$\int_{t_j^n}^{t_{j+1}^n} \int_{t_k^n}^{t_{k+1}^n} f_n(\boldsymbol{x}) b(\boldsymbol{x}) d\boldsymbol{x} = \int_{t_j^n}^{t_{j+1}^n} \int_{t_k^n}^{t_{k+1}^n} v(\boldsymbol{x}) b(\boldsymbol{x}) d\boldsymbol{x}, \quad n = 1, 2, 3, \ldots;\ j, k = 0, \ldots, 2n^2.$$

We show that the sequence f_n converges weakly to v.

Let $g \in C^\infty[\mathbb{R}^2, \mathbb{R}^2]$ and $\mathcal{X}_{\mathbb{M}}$ be the indicator of the box $\mathbb{M} = [\alpha, \beta] \times [\gamma, \eta]$. Given $\epsilon > 0$ we will show that there exists n' such that

$$|\langle g\mathcal{X}_{\mathbb{M}},\ f_n - v\rangle| \leq \epsilon$$

for all $n \geq n'$, i.e. $\langle g\mathcal{X}_{\mathbb{M}},\ f_n - v\rangle \to 0$. Let n_1 be such that $\mathbb{M} \subset \mathbb{I}_{n_1}$ for all $n \geq n_1$. Choose $n \geq n_1$. Integration by parts yields

$$\begin{aligned}
&\langle g\mathcal{X}_{\mathbb{M}},\ f_n - v\rangle = \\
&\left(g(\beta,\eta),\ \int_{\gamma}^{\eta}\int_{\alpha}^{\beta}[f_n(s,t) - v(s,t)]b(s,t)ds\,dt\right) && (3.51)\\
&-\int_{\gamma}^{\eta}\left(g_y(\beta,y),\ \int_{\gamma}^{y}\int_{\alpha}^{\beta}[f_n(s,t) - v(s,t)]b(s,t)ds\,dt\right)\,dy && (3.52)\\
&-\int_{\alpha}^{\beta}\left(g_x(x,\eta),\ \int_{\gamma}^{\eta}\int_{\alpha}^{x}[f_n(s,t) - v(s,t)]b(s,t)ds\,dt\right)\,dx && (3.53)\\
&+\int_{\gamma}^{\eta}\int_{\alpha}^{\beta}\left(g_{xy}(x,y),\ \int_{\gamma}^{y}\int_{\alpha}^{x}[f_n(s,t) - v(s,t)]b(s,t)ds\,dt\right)\,dx\,dy && (3.54)
\end{aligned}$$

Note that each of these terms contains an expression of the form

$$\begin{aligned}
\int_{\hat\gamma}^{\hat\eta}\int_{\hat\alpha}^{\hat\beta}(f_n(s,t) - v(s,t))b(s,t)ds\,dt \quad &= \quad \int_{\overline{(\hat\eta)}_n}^{\hat\eta}\int_{\hat\alpha}^{\hat\beta}(f_n(s,t) - v(s,t))b(s,t)ds\,dt \\
&+\int_{\hat\gamma}^{\underline{(\hat\gamma)}_n}\int_{\hat\alpha}^{\hat\beta}(f_n(s,t) - v(s,t))b(s,t)ds\,dt \\
&+\int_{\underline{(\hat\gamma)}_n}^{\overline{(\hat\eta)}_n}\int_{\hat\alpha}^{\underline{(\hat\alpha)}_n}(f_n(s,t) - v(s,t))b(s,t)ds\,dt \\
&+\int_{\underline{(\hat\gamma)}_n}^{\overline{(\hat\eta)}_n}\int_{\overline{(\hat\beta)}_n}^{\hat\beta}(f_n(s,t) - v(s,t))b(s,t)ds\,dt,
\end{aligned} \tag{3.55}$$

where

$$[\hat\gamma,\hat\eta]\times[\hat\alpha,\hat\beta] \subset [\gamma,\eta]\times[\alpha,\beta] \subset [-n,n]\times[-n,n].$$

In addition, for any box of the form $[\alpha',\beta']\times[\gamma',\eta']$, we have the bound

$$\begin{aligned}
\left|\int_{\gamma'}^{\eta'}\int_{\alpha'}^{\beta'}(f_n(s,t) - v(s,t))b(s,t)ds\,dt\right| \quad &\leq \quad \int_{\gamma'}^{\eta'}\int_{\alpha'}^{\beta}|f_n(s,t) - v(s,t)|b(s,t)ds\,dt \\
&\leq \quad \int_{\gamma'}^{\eta'}\int_{\alpha'}^{\beta}2|a(s,t)|b(s,t)ds\,dt \\
&= \quad 2\int_{\mathbb{R}^2}|a(\boldsymbol{x})|\mathcal{X}_{[\alpha',\beta']\times[\gamma',\eta']}(\boldsymbol{x})b(\boldsymbol{x})d\boldsymbol{x} \\
&\leq \quad 2\|a\|\int_{\mathbb{R}^2}\mathcal{X}_{[\alpha',\beta']\times[\gamma',\eta']}(\boldsymbol{x})b(\boldsymbol{x})d\boldsymbol{x} \\
&= \quad 2\|a\|\int_{[\alpha',\beta']\times[\gamma',\eta']}b(\boldsymbol{x})d\boldsymbol{x}.
\end{aligned}$$

Next note that the Lebesgue measure of each of the sets $[\overline{(\hat{\eta})}_n, \hat{\eta}] \times [\hat{\alpha}, \hat{\beta}]$, $[\hat{\gamma}, \underline{(\hat{\gamma})}_n] \times [\hat{\alpha}, \hat{\beta}]$, $[\underline{(\hat{\gamma})}_n, \hat{\eta}] \times [\hat{\alpha}, \underline{(\hat{\alpha})}_n]$, and $[\underline{(\hat{\gamma})}_n, \overline{(\hat{\eta})}_n] \times [\overline{(\hat{\beta})}_n, \hat{\beta}]$ appearing in (3.55) is bounded by

$$\frac{1}{n} \max\{(\eta - \gamma), (\beta - \alpha)\}$$

which can be made arbitrarily small. By [159, Exercise 12, page 33], for every $\bar{\epsilon} > 0$ there is a $\delta(\bar{\epsilon}) > 0$ such

$$\int_{\mathbb{E}} b(\boldsymbol{x}) d\boldsymbol{x} \leq \bar{\epsilon} \quad \text{whenever } \mathcal{M}(\mathbb{E}) \leq \delta(\bar{\epsilon}),$$

where $\mathcal{M}(\mathbb{E})$ is the Lebesgue measure of the set $\mathbb{E}$. Therefore, given $\bar{\epsilon} > 0$, we can choose n so that $\frac{1}{n} \max\{(\eta - \gamma), (\beta - \alpha)\} < \delta(\bar{\epsilon})$. By combining this with Eq.(3.55), we obtain the bound

$$\left| \int_{\hat{\gamma}}^{\hat{\eta}} \int_{\hat{\alpha}}^{\hat{\beta}} (f_n(s,t) - v(s,t)) b(s,t) ds\, dt \right| \leq 8\|a\|\sqrt{\bar{\epsilon}}. \tag{3.56}$$

If we set

$$\Gamma = \max\left\{ |g(s,t)|,\ |g_y(s,t)|,\ |g_x(s,t)|,\ |g_{xy}(s,t)|\ :\ (s,t) \in [\alpha, \beta] \times [\gamma, \eta] \right\}$$

the bounds Eq.(3.56) yield the following bound for the sum of the 4 integrands Eq.(3.51)-(3.54):

$$|\langle g\mathcal{X}_{\mathbb{M}},\ f_n - v \rangle| \leq \Gamma[1 + (\eta - \gamma) + (\beta - \alpha) + (\eta - \gamma)(\beta - \alpha)] \left[8\|a\|\sqrt{\bar{\epsilon}} \right]$$

Given any ϵ there exists an $\bar{\epsilon}$ such that the left hand side is less than ϵ, moreover, for this $\bar{\epsilon}$ there is an n' such that

$$\frac{1}{n} \max\{(\eta - \gamma), (\beta - \alpha)\} < \delta(\bar{\epsilon}) \quad \forall n \geq n'.$$

Therefore, for all $n \geq n'$ we have

$$|\langle g\mathcal{X}_{\mathbb{M}},\ f_n - v \rangle| \leq \epsilon,$$

which is what we set out to show. Since functions of the form $g\mathcal{X}_{\mathbb{M}}$, where $g \in C^{\infty}[\mathbb{R}^2, \mathbb{R}^2]$ and $\mathbb{M} \subset \mathbb{R}^2$ is a box, are dense in $L^2(\mathbb{R}^2, \mathcal{M}^2, P)$ we have that the sequence f_n converges weakly to v. □

3.2.4 Application to wavefront reconstruction

We now apply the above results to the negative modulus function $-\kappa(\boldsymbol{u})$.

Property 3.2.18 *Let $b \in L^1[\mathbb{R}^2, \mathbb{R}_+] \cap L^2[\mathbb{R}^2, \mathbb{R}_+]$ be a density function for the probability measure P on $(\mathbb{R}^2, \mathcal{M}^2)$. and let $\boldsymbol{u} \in L^2[\mathbb{R}^2, \mathbb{R}^2]$. The negative modulus function $-\kappa(\boldsymbol{u}(\boldsymbol{x}))$ has the following properties.*

(i)

$$\int \partial\left(-\kappa(0)\right) b(\boldsymbol{x}) d\boldsymbol{x} = \int -\Pi_{\mathbb{S}}(0) b(\boldsymbol{x}) d\boldsymbol{x} = con\ \mathbb{S},$$

where $\mathbb{S}$ *is the unit sphere.*

(ii) $\mathcal{S}\left(\overline{\partial}(-\kappa(\boldsymbol{u}(\cdot)))\right)$ *is weakly compact in* $L_2^2(\mathbb{R}^2, \mathcal{M}^2, P)$.

(iii)

$$\int \partial(-\kappa(\boldsymbol{u}(\boldsymbol{x})))b(\boldsymbol{x})d\boldsymbol{x} = \int -con\ \left(\Pi_{\mathbb{S}}(\boldsymbol{u}(\boldsymbol{x}))\right) b(\boldsymbol{x})d\boldsymbol{x},$$

and $\int \partial(-\kappa(\boldsymbol{u}(\boldsymbol{x})))b(\boldsymbol{x})d\boldsymbol{x}$ *is a compact subset of* $\mathbb{R}^2$.

(iv) On $L^2[\mathbb{R}^2, \mathbb{R}^2]$

$$\mathcal{S}\left(b(\cdot)\overline{\partial}(-\kappa(\boldsymbol{u}(\cdot)))\right) \subset -cl^*\Pi_{\mathbb{Q}[b]}[\boldsymbol{u}]$$

where $\mathbb{Q}[b]$ *is defined by Eq.(3.4) and* $\Pi_{\mathbb{Q}[b]}[\boldsymbol{u}]$ *by Eq.(3.3).*

PROOF: (i) This is an application of Corollary 3.2.14.
(ii) At each $\boldsymbol{x}$, $\overline{\partial}(-\kappa(\boldsymbol{u}(\boldsymbol{x})))$ is closed and convex-valued. In addition, by Example 3.2.9 every element of the set $\overline{\partial}(-\kappa(\boldsymbol{u}(\boldsymbol{x})))$ has magnitude less than or equal to 1 and so the multifunction $\overline{\partial}(-\kappa(\boldsymbol{u}(\cdot)))$ is L^2-bounded in $(\mathbb{R}^2, \mathcal{M}^2, P)$. Hence by Proposition 3.2.15 $\mathcal{S}\left(\overline{\partial}(-\kappa(\boldsymbol{u}(\boldsymbol{x})))\right)$ is weakly compact in $L_2^2(\mathbb{R}^2, \mathcal{M}^2, P)$.
(iii) We wish to apply Theorem 3.2.16, so we must show that the multifunction

$$F(\boldsymbol{x}) = [\partial(-\kappa) \circ \boldsymbol{u}](\boldsymbol{x}) = \partial(-\kappa(\boldsymbol{u}(\boldsymbol{x})))$$

is P-integrably bounded and $(\mathcal{M}^2 \otimes \mathcal{B}^2)$-measurable. By Example 3.2.9, the multifunction $F : \mathbb{R}^2 \rightrightarrows \mathbb{R}^2$ is P-integrably bounded with bound equal to 1. By Remark 3.2.7 the multifunction $\partial(-\kappa)$ has closed graph and is therefore $(\mathcal{M}^2 \otimes \mathcal{B}^2)$-measurable. By hypothesis, the function $\boldsymbol{u}$ is a Lebesgue measurable mapping from $(\mathbb{R}^2, \mathcal{M}^2)$ into $(\mathbb{R}^2, \mathcal{M}^2)$. Thus, by [89, Proposition 1.b, pg 59] the composite multifunction F defined above is $(\mathcal{M}^2 \otimes \mathcal{B}^2)$-measurable. Therefore Theorem 3.2.16 applies to give the result.
(iv) By Proposition 3.2.17 every $\boldsymbol{v}(\cdot) \in \mathcal{S}\left(b(\cdot)\overline{\partial}(-\kappa(\boldsymbol{u}(\cdot)))\right)$ is the weak limit of a sequence of functions in $\mathcal{S}\left(b(\cdot)\partial(-\kappa(\boldsymbol{u}(\cdot)))\right)$, since con $(\partial(-\kappa(\boldsymbol{u}(\cdot)))) = \overline{\partial}(-\kappa(\boldsymbol{u}(\cdot)))$ (see Example 3.2.9). If $\boldsymbol{v} \in \mathcal{S}\left(b(\cdot)\partial(-\kappa(\boldsymbol{u}(\cdot)))\right)$, then by Theorem 3.1.5 and Property 3.2.10 $-\boldsymbol{v} \in \Pi_{\mathbb{Q}[b]}[\boldsymbol{u}]$. Hence

$$\mathcal{S}\left(b(\cdot)\partial(-\kappa(\boldsymbol{u}(\cdot)))\right) \subset -\Pi_{\mathbb{Q}[b]}[\boldsymbol{u}]$$

from which the result follows. □

3.2.5 Infinite dimensional nonsmooth analysis

The next step is to relate the subdifferential of the integral of nonsmooth integrands to the integral of the subdifferential of nonsmooth integrands. This requires infinite dimensional nonsmooth analysis, for which several definitions are needed. For a complete discussion of the objects below see Ref. [44,45,92,94,126,129] and the references therein.

Definition 3.2.19 (Normal cones) *Let Ω be a nonempty subset of the Banach space $\mathbb{X}$. Denote the dual to $\mathbb{X}$ by $\mathbb{X}^*$. Let $u \in cl\Omega$, and $\epsilon \geq 0$:*

(i) *the* ϵ-normal cone *at $\overline{u} \in \Omega$, denoted $\widehat{N}^{\epsilon}_{\Omega}[u]$, is the set of* Fréchet ϵ-normals *to Ω:*

$$\widehat{N}^{\epsilon}_{\Omega}[\overline{u}] \equiv \left\{ u_* \in \mathbb{X}^* \;\middle|\; \limsup_{v \xrightarrow{\Omega} u} \frac{\langle u_*,\, v-u\rangle}{\|v-u\|} \leq \epsilon \right\}; \tag{3.57}$$

(ii) *the* normal cone *to Ω at u is defined by*

$$N_{\Omega}[u] \equiv \limsup_{\substack{v \xrightarrow{\Omega} u \\ \epsilon \downarrow 0}} \widehat{N}^{\epsilon}_{\Omega}[u]; \tag{3.58}$$

(iii) *if $\mathbb{X}$ is a real Hilbert space, and Ω is proximinal (Def. 3.1.2), then for any $u \in \mathbb{X}$, the* Proximal Normal cone *to Ω at $\overline{u}$, denoted $N^{P}_{\Omega}[\overline{u}]$, are the vectors v such that there exists a $\tau > 0$ with*

$$\overline{u} \in \Pi_{\Omega}[\overline{u} + \tau v],$$

where Π_{Ω} is defined by Eq.(3.3).

Definition 3.2.20 (Subgradients - infinite dimensions) *For any Banach space $\mathbb{X}$ and $f : \mathbb{X} \to \overline{\mathbb{R}}$,*

(i) *$u_* \in \mathbb{X}^*$ is an* analytic Fréchet ϵ-subgradient *of f at $\overline{u}$ if*

$$\liminf_{u \to \overline{u}} \frac{f[u] - f[\overline{u}] - \langle u_*,\, u-\overline{u}\rangle}{\|u-\overline{u}\|} \geq -\epsilon, \quad \epsilon \geq 0. \tag{3.59}$$

We call the set of analytic Fréchet ϵ-subgradients u_* the analytic Fréchet ϵ-subdifferential *of f at $\overline{u}$ and denote this set by $\widetilde{\partial}_{\epsilon} f[\overline{u}]$.*

(ii) *for $\mathbb{X}$ a Hilbert space and f l.s.c., $u_* \in \mathbb{X}$ is called a* proximal subgradient *of f at $\overline{u} \in dom\, f$ if*

$$(u_*, -1) \in N^{P}_{epi\, f}[\overline{u}, f[\overline{u}]].$$

We call the set of proximal subgradients u_* the proximal subdifferential *of f at $\overline{u}$ and denote this set by $\partial_P f[\overline{u}]$.*

(iii) $u_* \in \mathbb{X}^*$ *is a* (general) subgradient *of* f *at* $\overline{u} \in dom\, f$ *if*

$$(u_*, -1) \in N_{epi\, f}[\overline{u}, f[\overline{u}]]; \tag{3.60}$$

We call the set of (general) subgradients u_* *the* (general) subdifferential *of* f *at* $\overline{u}$ *and denote this set by* $\partial f[\overline{u}]$.

(iv) $u_* \in \mathbb{X}^*$ *is a horizon or* singular subgradient *of* f *at* $\overline{u} \in dom\, f$, *written* $u_* \in \partial^\infty f[\overline{u}]$, *the* singular subdifferential *of* f *at* $\overline{u}$, *if*

$$(u_*, 0) \in N_{epi\, f}[\overline{u}, f[\overline{u}]]; \tag{3.61}$$

We call the set of singular subgradients u_* *the* singular subdifferential *of* f *at* $\overline{u}$ *and denote this set by* $\partial^\infty f[\overline{u}]$.

(v) $u_* \in \mathbb{X}^*$ *is a* Dini-subgradient *of* f *at* $\overline{u} \in dom\, f$ *if*

$$\langle u_*, v\rangle \leq df[\overline{u}][v] + \epsilon\|v\| \;\forall\, v \in \mathbb{X} \tag{3.62}$$

where $df[\overline{u}](v)$ *is the subderivative defined in Def.3.2.5(i). We call the set of Dini-subgradients* u_* *the* Dini- subdifferential *of* f *at* $\overline{u}$ *and denote this set by* $\partial_\epsilon^- f[\overline{u}]$. *When* $\epsilon = 0$ *we write* $\partial^- f[\overline{u}]$ *instead of* $\partial_0^- f[\overline{u}]$*);*

(vi) for f *l.s.c. around* $\overline{u} \in dom\, f$, $u_* \in \mathbb{X}^*$ *is a* sequential A-subgradient *of* f *at* $\overline{u}$ *if there are sequences* $\epsilon^{(\nu)} \searrow 0$, $u^{(\nu)} \xrightarrow{f} \overline{u}$ *and* $u_*^{(\nu)} \in \partial_\epsilon^- f[u^{(\nu)}]$ *with* $u_*^{(\nu)} \overset{w^*}{\to} u_*$ *where* $\overset{w^*}{\to}$ *denotes weak-star convergence. We call the set of sequential A-subgradients* u_* *the* sequential A- subdifferential *of* f *at* $\overline{u}$ *and denote this set by* $\partial_A^\sigma f[\overline{u}]$.

(vii) for f *l.s.c. around* $\overline{u} \in dom\, f$, *the* A-subdifferential *of* f *at* $\overline{u} \in dom\, f$, *written* $\partial_A f[\overline{u}]$, *is defined as the* topological *limit*

$$\partial_A f[\overline{u}] \equiv \underset{u \xrightarrow{f} \overline{u},\, \epsilon\downarrow 0}{Limsup}\, \partial_\epsilon^- f[u]. \tag{3.63}$$

Even though Def.3.2.5(ii) is stated in finite dimensional settings, the definition is the same in infinite dimensions where convergence of sequences $v \to u$ is in norm.

Another subdifferential that is useful is the Clarke generalized subdifferential. This is defined below for Lipschitz mappings.

Definition 3.2.21 *For* $f : \mathbb{X} \to \overline{\mathbb{R}}$ *Lipschitz around* $\overline{u} \in \mathbb{X}$, *a real Banach space, the* Clarke generalized subdifferential *of* f *at* $\overline{u}$, *written* $\overline{\partial} f[\overline{u}]$, *satisfies*

$$\langle u_*,\, v\rangle \leq f^\circ[\overline{u}][v] \quad \forall\, v \in \mathbb{X} \tag{3.64}$$

where the generalized directional derivative is given by

$$f^\circ[\overline{u}][v] \equiv \limsup_{u\to\overline{u},\, t\downarrow 0} \frac{f[u+tv] - f[u]}{t}. \tag{3.65}$$

Theorems 3.2.24-3.2.25 and the following lemmas establish useful characterizations of subdifferentials.

Lemma 3.2.22 (Theorem 1 of [106] and Proposition 1 of [95]) *For any* $f : \mathbb{X} \to \overline{\mathbb{R}}$ *l.s.c. near* $\overline{u}$*,*

$$\partial f[\overline{u}] = \limsup_{u \xrightarrow{f} \overline{u},\ \epsilon \downarrow 0} \widetilde{\partial}_\epsilon f[u]. \tag{3.66}$$

The analytic Fréchet *zero*-subdifferential for l.s.c. mappings $f : \mathbb{X} \to \overline{\mathbb{R}}$, is equivalent to a closely related analog to the finite dimensional regular subgradient called the *Fréchet* subdifferential:

$$\widetilde{\partial}_0 f[u] = \widehat{\partial} f[u].$$

It follows from Eq.(3.66) that $\widehat{\partial} f[u] \subset \partial f[u]$. If f is l.s.c. near $\overline{u}$ and $\widehat{\partial} f[\overline{u}] = \partial f[\overline{u}]$ then f is said to be *(subdifferentially) regular* [129]. The class of subdifferentially regular functions includes, among others, all strictly differentiable functions and convex functions. It was shown in Example 3.2.9 that the negative modulus function is not regular at $u = 0$.

Lemma 3.2.23 (Mordukhovich and Shao [127]) *Let* $\mathbb{X}$ *be a Banach space and let* $f : \mathbb{X} \to \overline{\mathbb{R}}$ *be Lipschitz continuous around* $\overline{u}$*. Then* $\partial^\infty f[\overline{u}] = \{0\}$.

The next Theorem, due to Mordukhovich and Shao [129] is stated on *Asplund* spaces [128]. For our purposes, it is sufficient to note that a Hilbert space is an Asplund space.

Theorem 3.2.24 (Theorem 8.11 of [129]) *Let* $\mathbb{X}$ *be an Asplund space and let* $f : \mathbb{X} \to \overline{\mathbb{R}}$ *be Lipschitz continuous around* $\overline{u}$*. Then*

$$\overline{\partial} f[\overline{u}] = cl^* con\ \partial f[\overline{u}]. \tag{3.67}$$

Theorem 3.2.25 (Theorem 9.2 of [129]) *Let* X *be a Hilbert space. Suppose* $f : \mathbb{X} \to \overline{\mathbb{R}}$ *is Lipschitz continuous around* $\overline{u}$*, then the sets* $\partial f[\overline{u}]$ *and* $\partial^\sigma_A f[\overline{u}]$ *are weakly closed, and*

$$\partial f[\overline{u}] = \partial^\sigma_A f[\overline{u}] = \partial_A f[\overline{u}]. \tag{3.68}$$

Theorem 3.2.25 is narrower than the statement of Theorem 9.2 of [129], however it is all we need for our purposes. It should also be noted that when f is strictly differentiable then $\partial f[\boldsymbol{u}]$ coincides with the Fréchet derivative.

We are now ready to establish the connection between the projection $\Pi_{\mathbb{Q}[b]}$ defined by Eq.(3.3) and the subdifferential of $h : L^2[\mathbb{R}^2, \mathbb{R}^2] \to \mathbb{R}$ defined by Eq.(3.40) for $b \in \mathbb{U}_+$ defined in Hypothesis 2.2.1. Property 3.2.29 is a special case of Theorem 3.2.30, which is proved at the end of this section. Rather than simply applying Theorem 3.2.30, we provide another proof that motivates perturbation methods reviewed in Section 3.3.

Before proceeding we state without proof few necessary lemmas.

Lemma 3.2.26 (Lemma 1 of [93]) *Consider the measure space* $(\Omega, \mathcal{A}, \mu)$ *where* μ *is a complete* σ*-finite positive measure. For* $f : \Omega \times \mathbb{R}^m \to \overline{\mathbb{R}}$ *l.s.c. in* $u \in \mathbb{R}^m$*, and* $\mathcal{A} \otimes \mathcal{B}^m$*-measurable in* (ω, u)*, then the graph of the multi-valued map* $\Gamma : \Omega \to \mathbb{R}^m \times \mathbb{R}^m$ *defined by*

$$\Gamma(\omega)|\ \omega \to \left\{(\boldsymbol{v}, u)\ \middle|\ \boldsymbol{v} \in \partial^-_u f(\omega, u)\right\}$$

belongs to $\mathcal{A}\otimes\mathcal{B}(\mathbb{R}^m\times\mathbb{R}^m)$ *where* ∂_u^- *denotes the Dini-subdifferential with respect to* u*, and* $\mathcal{B}(\mathbb{R}^m\times\mathbb{R}^m)$ *is the collection of Borel subsets of* $\mathbb{R}^m\times\mathbb{R}^m$*.*

Lemma 3.2.27 (Aumann [9], Theorem 2) *Let* F *be a multi-valued mapping of a measure space* $(\Omega,\mathcal{A},\mu)$ *into* $\mathbb{R}^m$ *such that the graph of* F *belongs to* $\mathcal{A}\otimes\mathcal{B}^m$*. Then there exists a measurable function* $f:\Omega\to\mathbb{R}^m$ *such that* $f(\boldsymbol{x})\in F(\boldsymbol{x})$*, a.e. in* Ω*.*

Aumann's theorem is actually stated in terms of any general separable metric space $\mathbb{X}$ rather than just $\mathbb{R}^m$. For proof of this limited case see [89].

Lemma 3.2.28 ([54, 89]) *Let* $(\Omega,\mathcal{A},\mu)$ *be a complete measure space,* $\mathbb{X}$ *a complete separable metric space,* F *a multi-valued mapping from* Ω *to* $\mathbb{X}$ *with measurable (analytic) graph, and* v *a measurable function of* $\mathbb{X}$ *into* $\mathbb{R}$*. Then the function* $\sup v(F(\cdot))$ *of* Ω *into* $\mathbb{R}$*,*

$$\omega\mapsto\sup\{v(x)\mid x\in F(\omega)\},$$

is measurable, and the relation F^v *of* Ω *into* $\mathbb{X}$ *defined by*

$$\omega\mapsto\{x\in F(\omega)\mid v(x)=\sup v(F(\omega))\}$$

has a measurable (analytic) graph.

We are now ready to state the main result of this chapter.

Property 3.2.29 *Let* $b\in\mathbb{U}_+$ *be as defined in Hypothesis 2.2.1,* $\boldsymbol{u}\in L^2[\mathbb{R}^2,\mathbb{R}^2]$*,* $\Pi_{\mathbb{Q}[b]}: L^2\rightrightarrows\mathbb{Q}[b]$ *be as defined by Eq.(3.3) for* $\mathbb{Q}[b]$ *defined by Eq.(3.4), and* $h: L^2[\mathbb{R}^2,\mathbb{R}^2]\to\overline{\mathbb{R}}$ *be as defined by Eq.(3.40). Then*

$$\partial\left(h[\boldsymbol{u};b]\right)=\mathcal{S}\left(b(\cdot)\overline{\partial}\left(-\kappa(\boldsymbol{u}(\cdot))\right)\right)=cl^*(-\Pi_{\mathbb{Q}[b]}[\boldsymbol{u}]) \tag{3.69}$$

PROOF: Due to part (*iv*) of Property 3.2.18 we need only show the following two inclusions

$$\begin{aligned}\mathrm{cl}^*(-\Pi_{\mathbb{Q}[b]}[\boldsymbol{u}]) &\subset \partial h[\boldsymbol{u};b] &(3.70)\\ &\subset \mathcal{S}\left(b(\cdot)\overline{\partial}(-\kappa(\boldsymbol{u}(\cdot)))\right). &(3.71)\end{aligned}$$

We begin with inclusion (3.70).

From the proof of Property 3.2.3 it is easily seen that the mapping h is globally Lipschitz continuous with Lipschitz constant $K=\|b\|$, hence $\partial h[\boldsymbol{u};b]$ is weakly closed. Therefore, if $-\Pi_{\mathbb{Q}[b]}[\boldsymbol{u}]\subset\partial h[\boldsymbol{u};b]$ then $\mathrm{cl}^*(-\Pi_{\mathbb{Q}[b]}[\boldsymbol{u}])\subset\partial h[\boldsymbol{u};b]$. We now show that $-\Pi_{\mathbb{Q}[b]}[\boldsymbol{u}]\subset\partial h[\boldsymbol{u};b]$.

Let $\boldsymbol{v}\in-\Pi_{\mathbb{Q}[b]}[\boldsymbol{u}]$ and for all $\epsilon>0$ define

$$\tilde{\boldsymbol{u}}_\epsilon=\boldsymbol{u}\mathcal{X}_{\mathrm{supp}(\boldsymbol{u})}+\epsilon\boldsymbol{v}(1-\mathcal{X}_{\mathrm{supp}(\boldsymbol{u})}).$$

Then, by Theorem 3.1.5,

$$\|\boldsymbol{u}-\tilde{\boldsymbol{u}}_\epsilon\|=\epsilon\|\boldsymbol{v}(1-\mathcal{X}_{\mathrm{supp}(\boldsymbol{u})})\|\le\epsilon\|b\|,$$

and $\kappa(\boldsymbol{w})$ is differentiable at $\boldsymbol{w} = \tilde{\boldsymbol{u}}_\epsilon(\boldsymbol{x})$ for every $\boldsymbol{x} \in \operatorname{supp}(b)$ with

$$\boldsymbol{v}(\boldsymbol{x}) = -\nabla\kappa(\tilde{\boldsymbol{u}}_\epsilon(\boldsymbol{x}))b(\boldsymbol{x}) \quad \forall\, \epsilon > 0.$$

For every $\boldsymbol{w} \in L^2[\mathbb{R}^2, \mathbb{R}^2]$ and $\boldsymbol{x} \in \operatorname{supp}(b)$, we have

$$\frac{\kappa(\tilde{\boldsymbol{u}}_\epsilon(\boldsymbol{x}) + t\boldsymbol{w}(\boldsymbol{x})) - \kappa(\tilde{\boldsymbol{u}}_\epsilon(\boldsymbol{x}))}{t} \to (\nabla\kappa(\tilde{\boldsymbol{u}}_\epsilon(\boldsymbol{x})), \boldsymbol{w}(\boldsymbol{x})), \tag{3.72}$$

and, since κ is Lipschitz with Lipschitz constant 1,

$$\left| \frac{\kappa(\tilde{\boldsymbol{u}}_\epsilon(\boldsymbol{x}) + t\boldsymbol{w}(\boldsymbol{x})) - \kappa(\tilde{\boldsymbol{u}}_\epsilon(\boldsymbol{x}))}{t} \right| \leq |\boldsymbol{w}(\boldsymbol{x})| \quad \forall\, x \in \operatorname{supp}(b).$$

Therefore, by the Lebesgue Dominated Convergence Theorem, the function $h[\cdot; b]$ is Gâteaux differentiable at $\tilde{\boldsymbol{u}}_\epsilon$ with Gâteaux derivative $-\nabla\kappa[\tilde{\boldsymbol{u}}_\epsilon]b = \boldsymbol{v}$. Hence, since $\kappa[\cdot]$ is Lipschitz continuous, we have

$$dh[\tilde{\boldsymbol{u}}_\epsilon; b][w] = \int_{\mathbb{R}^2} (-\nabla\kappa(\tilde{\boldsymbol{u}}_\epsilon(\boldsymbol{x}))b(\boldsymbol{x}), \boldsymbol{w}(\boldsymbol{x}))dx = \langle \boldsymbol{v}, \boldsymbol{w} \rangle,$$

which implies that

$$\boldsymbol{v} \in \partial^- h[\tilde{\boldsymbol{u}}_\epsilon; b] \quad \forall\, \epsilon > 0.$$

Taking the limit as $\epsilon \downarrow 0$, we find that

$$\boldsymbol{v} \in \partial_A h[\boldsymbol{u}; b] = \partial h[\boldsymbol{u}; b].$$

Therefore, $-\Pi_{\mathbb{Q}[b]}[\boldsymbol{u}] \subset \partial h[\boldsymbol{u}; b]$.

Our proof of inclusion (3.71) is modeled after the proof of Theorem 5.18 of Clarke [45]. Since $\partial h[\boldsymbol{u}; b] \subset \overline{\partial} h[\boldsymbol{u}; b]$ it suffices to show that $\overline{\partial} h[\boldsymbol{u}; b] \subset \mathcal{S}\left(b(\cdot)\overline{\partial}(-\kappa(\boldsymbol{u}(\cdot)))\right)$. By [154, Theorem 9.13], since κ is globally Lipschitz, for all $\boldsymbol{u}$ the mapping $\partial(-\kappa(\boldsymbol{u}))$ is bounded, nonempty and compact, and one has

$$\operatorname{lip}(-\kappa(\boldsymbol{u})) = 1 < \infty$$

where $\operatorname{lip}(-\kappa(\boldsymbol{u}))$ is the *Lipschitz modulus* of $(-\kappa(\boldsymbol{u}))$ at u ([154, Def. 9.1]). Thus, by Proposition 3.2.15 the set

$$\left\{ \boldsymbol{v} \in L^2[\mathbb{R}^2, \mathbb{R}^2] \,\middle|\, \boldsymbol{v}(\boldsymbol{x}) \in \overline{\partial}(-\kappa(\boldsymbol{u}(\boldsymbol{x}))) \ a.e. \right\}$$

is convex and weakly compact in $L^2[\mathbb{R}^2, \mathbb{R}^2]$. Since P is a probability measure we also have $\operatorname{lip} h[\boldsymbol{u}; b] < \|b\|$ for all $\boldsymbol{u}$. From Theorem 3.2.24 we have that $\overline{\partial} h[\boldsymbol{u}; b]$ is weakly closed, thus $\overline{\partial} h[\boldsymbol{u}; b]$ is weakly compact. By [45, Proposition 1.5.c, pg.73] we can use support functions to establish (3.71). That is, we must show that, for any $\boldsymbol{w} \in L^2[\mathbb{R}^2, \mathbb{R}^2]$, we have

$$h^\circ[\boldsymbol{u}; b][\boldsymbol{w}] \leq \max\left\{ \langle \boldsymbol{w},\ \boldsymbol{v} \rangle_P \,\middle|\, \boldsymbol{v}(\boldsymbol{x}) \in \overline{\partial} f(\boldsymbol{u}(\boldsymbol{x})) \quad a.e. \right\}$$

where $\langle\cdot,\cdot\rangle_P$ is the inner product with respect to the measure P. Let $\{\boldsymbol{u}_i\}$ be a sequence in $L^2[\mathbb{R}^2,\mathbb{R}^2]$ converging strongly to $\boldsymbol{u}$ and $\{\tau_i\}$ be a sequence in $\mathbb{R}_+$ converging to 0 such that

$$\begin{aligned} h^\circ[\boldsymbol{u};b][\boldsymbol{w}] &= \lim_{i\to\infty}\frac{h[\boldsymbol{u}_i+\tau_i\boldsymbol{w};b]-h[\boldsymbol{u}_i;b]}{\tau_i} \\ &= \lim_{i\to\infty}\int_{\mathbb{R}^2}\frac{-\kappa(\boldsymbol{u}_i(\boldsymbol{x})+\tau_i\boldsymbol{w}(\boldsymbol{x}))+\kappa(\boldsymbol{u}_i(\boldsymbol{x}))}{\tau_i}b(\boldsymbol{x})d\boldsymbol{x}. \end{aligned}$$

Since κ is P-integrably bounded and globally Lipschitz, we may apply the following analogue to Fatou's Lemma [161] [89, Lemma 3, pg. 69] to obtain the bound

$$\begin{aligned} h^\circ[\boldsymbol{u};b][\boldsymbol{w}] &\le \int_{\mathbb{R}^2}\limsup_{i\to\infty}\frac{-\kappa(\boldsymbol{u}_i(\boldsymbol{x})+\tau_i\boldsymbol{w}(\boldsymbol{x}))+\kappa(\boldsymbol{u}_i(\boldsymbol{x}))}{\tau_i}b(\boldsymbol{x})d\boldsymbol{x} \\ &\le \int_{\mathbb{R}^2}(-\kappa)^\circ(\boldsymbol{u}(\boldsymbol{x}))(\boldsymbol{w}(\boldsymbol{x}))b(\boldsymbol{x})d\boldsymbol{x}. \end{aligned}$$

By Lemmas 3.2.27-3.2.28 and the fact that $\overline{\partial}h[\boldsymbol{u};b]$ is compact-valued, there exists a measurable selection $\boldsymbol{v}(\cdot)\in\overline{\partial}(-\kappa(\boldsymbol{u}(\cdot)))$ such that

$$(\boldsymbol{v}(\boldsymbol{x}),\ \boldsymbol{w}(\boldsymbol{x})) = (-\kappa)^\circ(\boldsymbol{u}(\boldsymbol{x}))(\boldsymbol{w}(\boldsymbol{x})).$$

Then

$$\int_{\mathbb{R}^2}(-\kappa)^\circ(\boldsymbol{u}(\boldsymbol{x}))(\boldsymbol{w}(\boldsymbol{x}))b(\boldsymbol{x})d\boldsymbol{x} = \langle\boldsymbol{v},\ \boldsymbol{w}\rangle_P$$

which establishes inclusion (3.71). This completes the proof. □

We finish this subsection with a generalization of Theorem 3.2.29. Theorem 3.2.30 establishes the equivalence of the infinite dimensional subdifferential objects in the setting relevant to phase retrieval, and establishes their relation to the finite dimensional Clarke subdifferential.

Theorem 3.2.30 *Let $\boldsymbol{u}\in L^2[\mathbb{R}^2,\mathbb{R}^2]$, let $f:\mathbb{R}^2\to\overline{\mathbb{R}}$ be globally Lipschitz continuous, and let $f(\boldsymbol{u}(\boldsymbol{x}))$ be an L^2-bounded function of $\boldsymbol{x}$ on the probability space $(\mathbb{R}^2,\mathcal{M}^2,P)$ where P is a complete, non-atomic probability measure with density $b:\mathbb{R}^2\to\mathbb{R}_+$, and $\mathcal{M}^2$ denotes the Lebesgue measurable sets on $\mathbb{R}^2$. Define the integral functional $J: L^2[\mathbb{R}^2,\mathbb{R}^2]\to\overline{\mathbb{R}}$ by*

$$J[\boldsymbol{u}] = \int_{\mathbb{R}^2} f(\boldsymbol{u}(\boldsymbol{x}))b(\boldsymbol{x})d\boldsymbol{x}.$$

Then

$$\partial J[\boldsymbol{u}] = \partial_A^\sigma J[\boldsymbol{u}] = \partial_A J[\boldsymbol{u}] \subset \left\{\boldsymbol{v}\in L^2[\mathbb{R}^2,\mathbb{R}^2]\ \middle|\ \boldsymbol{v}(\boldsymbol{x})\in\overline{\partial}f(\boldsymbol{u}(\boldsymbol{x}))\ a.e.\right\}$$

PROOF: The first two equalities follow from Theorem 3.2.25. For the last inclusion, since $\partial J[\boldsymbol{u}]\subset\overline{\partial}J[\boldsymbol{u}]$ it suffices to show the following:

$$\overline{\partial}J[\boldsymbol{u}] \quad \subset \quad \left\{\boldsymbol{v}\in L^2[\mathbb{R}^2,\mathbb{R}^2]\ \middle|\ \boldsymbol{v}(\boldsymbol{x})\in\overline{\partial}f(\boldsymbol{u}(\boldsymbol{x}))\ a.e.\right\}. \tag{3.73}$$

The proof follows exactly the proof of inclusion (3.71) with $J=h$ and $(-\kappa)=f$. □

Conjecture 3.2.1 *For the same setting as Theorem 3.2.30 the last inclusion holds with equality, that is*

$$\partial J[\boldsymbol{u}] = \partial_A^\sigma J[\boldsymbol{u}] = \partial_A J[\boldsymbol{u}] = \left\{ \boldsymbol{v} \in L^2[\mathbb{R}^2, \mathbb{R}^2] \;\middle|\; \boldsymbol{v}(\boldsymbol{x}) \in \overline{\partial} f(\boldsymbol{u}(\boldsymbol{x})) \text{ a.e.} \right\}.$$

Moreover, this implies

$$\overline{\partial} J[\boldsymbol{u}] = \partial J[\boldsymbol{u}].$$

SKETCH OF PROOF: The idea for this proof is modeled after the proof of Theorem 3 of [93]. Since $\partial_A^\sigma J[\boldsymbol{u}] \subset \overline{\partial} J[\boldsymbol{u}]$ all that needs to be shown is the following:

$$\begin{aligned} &\left\{ \boldsymbol{v} \in L^2[\mathbb{R}^2, \mathbb{R}^2] \;\middle|\; \boldsymbol{v}(\boldsymbol{x}) \in \overline{\partial} f(\boldsymbol{u}(\boldsymbol{x})) \text{ a.e.} \right\} \\ &\qquad \subset \left\{ \boldsymbol{v} = \operatorname*{w-lim}_{i\to\infty} \boldsymbol{v}_i, \ \boldsymbol{v}_i \in L^2[\mathbb{R}^2, \mathbb{R}^2] \mid \boldsymbol{v}_i \in \partial f(\boldsymbol{u}(\boldsymbol{x})) \text{ a.e.} \right\} \qquad (3.74) \\ &\qquad \subset \partial_A^\sigma J[\boldsymbol{u}]. \qquad (3.75) \end{aligned}$$

Inclusion (3.74) follows immediately from Proposition 3.2.17 since for f Lipschitz, $\overline{\partial} f(\boldsymbol{u}(\boldsymbol{x})) = \operatorname{con} \partial f(\boldsymbol{u}(\boldsymbol{x}))$ (see Theorem 8.49 of [154]).

For inclusion (3.75) suppose that $\boldsymbol{v} \in L^2[\mathbb{R}^2, \mathbb{R}^2]$, $\boldsymbol{v}(\boldsymbol{x}) \in \partial f(\boldsymbol{u}(\boldsymbol{x}))$ a.e. Let $\mathbb{Q}_j(\boldsymbol{x})$ denote the set of all $(\boldsymbol{\nu}, \boldsymbol{w}) \in \mathbb{R}^2 \times \mathbb{R}^2$ satisfying $|\boldsymbol{w} - \boldsymbol{u}(\boldsymbol{x})| < 1/j$, $|f(\boldsymbol{w}) - f(\boldsymbol{u}(\boldsymbol{x}))| < 1/j$, $|\boldsymbol{\nu} - \boldsymbol{v}(\boldsymbol{x})| < 1/j$ and $\boldsymbol{\nu} \in \partial^- f(\boldsymbol{w})$. For almost every $\boldsymbol{x}$ and for every $j > 0$ the set $\mathbb{Q}_j$ is nonempty. By Lemma 3.2.26 $\mathbb{Q}_j$ belongs to $\mathcal{M}^2 \otimes \mathcal{B}(\mathbb{R}^2 \times \mathbb{R}^2)$, and, since P is a complete measure, Lemma 3.2.27 implies the existence of a measurable selection of $\mathbb{Q}_j$ that is a pair $(\boldsymbol{v}_j(\cdot), \boldsymbol{u}_j(\cdot))$ such that $|\boldsymbol{u}_j(\boldsymbol{x}) - \boldsymbol{u}(\boldsymbol{x})| < 1/j$, $|f(\boldsymbol{u}_j(\boldsymbol{x})) - f(\boldsymbol{u}(\boldsymbol{x}))| < 1/j$, $|\boldsymbol{v}_j(\boldsymbol{x}) - \boldsymbol{v}(\boldsymbol{x})| < 1/j$, and $\boldsymbol{v}_j(\boldsymbol{x}) \in \partial^- f(\boldsymbol{u}_j(\boldsymbol{x}))$ a.e. Therefore, the sequence of measurable selections $\{\boldsymbol{u}_j\}$ converges strongly to $\boldsymbol{u}$ and $\{\boldsymbol{v}_j\}$, converges weakly to $\boldsymbol{v}$.

Since f is globally Lipschitz, so is J and for any $\boldsymbol{w} \in L_2^2(\mathbb{R}^2, \mathcal{M}^2, P)$

$$dJ[\boldsymbol{u}(\cdot)][\boldsymbol{w}(\cdot)] = \liminf_{\tau \searrow 0} \frac{J[\boldsymbol{u}(\cdot) + \tau \boldsymbol{w}(\cdot)] - J[\boldsymbol{u}(\cdot)]}{\tau}$$

Now, by Fatou's Lemma,

$$\langle \boldsymbol{v}_j(\cdot), \ \boldsymbol{w}(\cdot) \rangle_P \le dJ[\boldsymbol{u}(\cdot)][\boldsymbol{w}(\cdot)] \qquad (3.76)$$

thus $\boldsymbol{v}_j \in \partial^- J[\boldsymbol{u}_j]$ and by the definition of the sequential A-subdifferential $\boldsymbol{v} \in \partial_A^\sigma J[\boldsymbol{u}]$. By Theorem 3.2.25 $\partial_A^\sigma J[\boldsymbol{u}]$ is weakly closed the result should follow. □

3.2.6 Subdifferential Calculus

We now proceed to fundamental calculus results for subdifferentials in Section 3.2.5. We are interested in characterizing the subdifferential of sums and compositions of nonsmooth functions for application to the set distance error defined in Eq. (3.31). In general the calculus of subdifferentials is "fuzzy" in the sense that the subdifferential of sums is contained, but not equal to, the sum of subdifferentials. However, for subdifferentially regular mappings these "fuzzy" relations become equality.

Results in this direction are established under the following compactness condition.

Definition 3.2.31 *A closed set* $\Omega \subset \mathbb{X}$ *is said to be* normally compact *around* $\overline{u} \in \Omega$ *if there exist positive numbers* γ, δ, *and a compact subset* $\mathbb{Y}$ *of* $\mathbb{X}$ *such that*

$$\widehat{N}(u;\Omega) \subset K_\gamma(\mathbb{Y}) \equiv \left\{ u_* \in \mathbb{X}^* \;\middle|\; \gamma\|u_*\| \leq \max_{y\in\mathbb{Y}} |\langle u_*, y\rangle| \right\} \;\; \forall\, u \in \mathbb{B}_\delta(\overline{u}) \cap \Omega, \tag{3.77}$$

where $\mathbb{B}_\delta(\overline{u})$ *is the* δ*-ball around* $\overline{u}$*. A* function $f : \mathbb{X} \to \overline{\mathbb{R}}$ *is normally compact around* $\overline{u} \in dom\, f$ *if its epigraph is normally compact around* $(\overline{u}, f(\overline{u}))$.

The next lemma establishes that Lipschitz functions are normally compact.

Lemma 3.2.32 *For general Banach spaces* $\mathbb{X}$, *if* $f : \mathbb{X} \to \overline{\mathbb{R}}$ *is Lipschitz around a point* $\overline{u} \in dom\, f$ *then epi* f *is normally compact.*

PROOF: This is a consequence of a series of results which establish that, for general Banach spaces $\mathbb{X}$, if $\Omega \subset \mathbb{X}$ is *epi-Lipschitzian* around $\overline{u}$ in the sense of Rockafellar [153], then Ω is *compactly epi-Lipschitzian* around $\overline{u}$ [27] in the sense of Borwein and Strojwas [28], which in turn implies that Ω is normally compact around $\overline{u}$ [113]. Since f is Lipschitz around $\overline{u}$ then epi f is epi-Lipschitzian around $\overline{u}$. □

To fix these ideas, note that since $h[\overline{\boldsymbol{u}}; b]$ defined by Eq.(3.40) is globally Lipschitz for all $b \in L^2$, then epi h is closed and epi-Lipschitzian around $(\overline{\boldsymbol{u}}, h[\overline{\boldsymbol{u}}; b])$. Hence epi h is normally compact.

The next theorem due to Mordukhovich and Shao [129] contains the general sum rules for the subdifferential constructions of Def.3.2.20(ii). Again, Mordukhovich and Shao construct a general subdifferential calculus on Asplund spaces which include Hilbert spaces.

Theorem 3.2.33 (Theorem 4.1 of Mordukhovich and Shao [129]) *Let* $\mathbb{X}$ *be an Asplund space, let* $f_i : \mathbb{X} \to \overline{\mathbb{R}}$, $i = 1, 2, \ldots, n$ *be l.s.c. around* $\overline{u}$, *and let all but possibly one of these functions be normally compact around* $\overline{u}$. *Suppose also that the following qualification condition holds:*

$$[u_{i*} \in \partial^\infty f_i[\overline{u}],\; i = 1, \ldots, n \mid \sum_{i=1}^{n} u_{i*} = 0] \implies u_{1*} = \cdots = u_{n*} = 0. \tag{3.78}$$

Then one has the inclusions

$$\partial(f_1 + \cdots + f_n)[\overline{u}] \subset \partial f_1[\overline{u}] + \cdots + \partial f_n[\overline{u}], \tag{3.79}$$

$$\partial^\infty(f_1 + \cdots + f_n)[\overline{u}] \subset \partial^\infty f_1[\overline{u}] + \cdots + \partial^\infty f_n[\overline{u}], \tag{3.80}$$

Moreover, if all f_i *are subdifferentially regular at* $\overline{u}$, *then the sum* $f_1 + \cdots + f_n$ *is also subdifferentially regular at this point and equality holds in Eq.(3.79).*

Corollary 3.2.34 (Corollary 4.3 of Mordukhovich and Shao [129]) *Let* $\mathbb{X}$ *be an Asplund space and let all but possibly one of the functions* f_i *be Lipschitz continuous around* $\overline{u}$. *Then:*

(i) Eq.(3.79) holds. Moreover, if all but possibly one of the f_i are strictly differentiable around $\overline{u}$ then inclusion becomes an equality *;*

(ii) Eq.(3.80) holds with equality.

Corollary 3.2.34 can be immediately applied to the squared distance function $\operatorname{dist}^2(\boldsymbol{u}, \mathbb{Q}[b])$ defined by Eq.(3.2) for $b \in \mathbb{U}_+$ defined in Hypothesis 2.2.1. From the representation Eq.(3.41) it is clear that $\operatorname{dist}^2(\boldsymbol{u}, \mathbb{Q}[b])$ is the sum of the strictly differentiable function $g[\boldsymbol{u}] = \|\boldsymbol{u}\|^2 + \|b\|^2$ and the nonsmooth Lipschitz continuous function $2h[\overline{\boldsymbol{u}}; b]$. By Corollary 3.2.34

$$\partial(\operatorname{dist}^2(\boldsymbol{u}, \mathbb{Q}[b])) = 2\boldsymbol{u} + \partial(2h[\overline{\boldsymbol{u}}; b]).$$

From Example 3.2.9 and Property 3.2.29 we also have that $\partial(c\, h[\boldsymbol{u}; b]) = c\, \partial h[\boldsymbol{u}; b]$ for scalars $c \geq 0$ thus

$$\partial(\operatorname{dist}^2(\boldsymbol{u}, \mathbb{Q}[b])) = 2\left(\boldsymbol{u} + \operatorname{cl}^*\left(-\Pi_{\mathbb{Q}[b]}[\boldsymbol{u}]\right)\right) \tag{3.81}$$

The next theorem is a specialization of Theorem 6.7 of Mordukhovich and Shao [129] which establishes the Chain rule for subdifferential calculus.

Theorem 3.2.35 (Chain Rule) *Let $\mathbb{X}$ and $\mathbb{Y}$ be Asplund spaces, let $G : \mathbb{X} \to \mathbb{Y}$ be strictly differentiable at $\overline{u}$ with $G'(\overline{u})$ invertible, and let $f : \mathbb{Y} \to \overline{\mathbb{R}}$ be l.s.c. around $\overline{v} = G(\overline{u})$. Then*

$$\partial(f \circ G)(\overline{u}) = (G'(\overline{u}))^* \partial f(\overline{v}) \tag{3.82}$$

and

$$\partial^\infty(f \circ G)(\overline{u}) = (G'(\overline{u}))^* \partial^\infty f(\overline{v}). \tag{3.83}$$

PROOF OF PROPERTY 3.2.1: Specializing to the distance function $\operatorname{dist}^2(\boldsymbol{u}, \mathbb{Q}_m)$, for $\mathbb{Q}_m$ defined by Eq.(3.1), Corollary 3.1.7 yields

$$\begin{aligned}
\operatorname{dist}^2(\boldsymbol{u}, \mathbb{Q}_m) &= \operatorname{dist}^2(\mathcal{F}_m[\boldsymbol{u}], \mathbb{Q}[\psi_m]) \\
&= \|\mathcal{F}_m[\boldsymbol{u}]\|^2 + \|\psi_m\|^2 + 2h[\mathcal{F}_m[\boldsymbol{u}]; \psi_m].
\end{aligned}$$

By Property 3.2.29, Corollary 3.2.34 and Theorem 3.2.35

$$\begin{aligned}
\partial(\|\mathcal{F}_m[\boldsymbol{u}]\|^2 + \|\psi_m\|^2 + 2h[\mathcal{F}_m[\boldsymbol{u}]; \psi_m]) &= 2\boldsymbol{u} + 2\mathcal{F}_m^*\left[\operatorname{cl}^*\left(-\Pi_{\mathbb{Q}[\psi_m]}[\mathcal{F}_m[\boldsymbol{u}]]\right)\right] \\
&= 2\boldsymbol{u} + 2\operatorname{cl}^*\left(\mathcal{F}_m^*\left[-\Pi_{\mathbb{Q}[\psi_m]}[\mathcal{F}_m[\boldsymbol{u}]]\right]\right) \\
&= 2\boldsymbol{u} + 2\operatorname{cl}^*\left(-\Pi_{\mathbb{Q}_m}[\boldsymbol{u}]\right) \\
&= 2\operatorname{cl}^*\left(\mathcal{I} - \Pi_{\mathbb{Q}_m}\right)[\boldsymbol{u}]
\end{aligned} \tag{3.84}$$

thus proving Eq.(3.35) of Property 3.2.1. Theorem 3.2.33 and Eq.(3.35) yield Eq.(3.36). □

3.3 Optimization of Smooth Perturbations

The remainder of this chapter is dedicated to characterizing a smooth approximation of the squared set distance error. Our numerical methods are based on smooth objectives. This allows us to provide a convergence theory that is easily derived from standard results in the optimization literature. By relating the smooth approximations to the projection operators we are able to provide an interpretation of iterative transform methods in the context of the analytic methods studied in this section. One obvious solution to the problem of nonsmooth objectives is simply to square the data and the modulus. The modulus squared is a smooth function. For this reason, analytic techniques tend to favor objectives based on the modulus squared. See Ref. [57] for a very careful treatment of analytic techniques for the modulus squared. In our experiments, however, objectives based on the modulus squared, while robust, suffer from very slow rates of convergence compared to the nonsmooth or nearly nonsmooth objectives studied Section 3.2. An intuitive explanation for this is that the modulus squared smoothes out curvature information in the objective [96, 109]. Another explanation is that the singular values of the operator $|\mathcal{F}_m[\boldsymbol{u}]|^2$ are much more spread out compared to those of the operator $|\mathcal{F}_m[\boldsymbol{u}]|$; that is, the squared modulus system is more ill-conditioned than the modulus system. This results in slower convergence of methods based on linearizations of the operator $|\mathcal{F}_m[\boldsymbol{u}]|^2$. See Ref. [11,83] for a discussion. While it is difficult to work with, we have found that the modulus function outperforms the modulus squared function as an objective in optimization techniques. The principal goal of this work is to develop tools for taking advantage of these "good" aspects of the modulus, while avoiding instabilities.

Two analytic approaches are considered. The first is a direct application of smoothing methods to E which we refer to as perturbed least squares; the second is an extended least squares approach that allows us to adaptively correct for the relative variability in the diversity measurements, ψ_m.

3.3.1 Perturbed Least Squares

In order to avoid difficulties associated with nondifferentiability we now consider smooth objectives that are perturbations to the least squares objective functional E. The smooth least squares objective function we consider in this section is based on a smooth perturbation of the modulus function $\kappa(\boldsymbol{u}) = |\boldsymbol{u}|$ of the form

$$\kappa_\epsilon(\boldsymbol{u}) = \frac{|\boldsymbol{u}|^2}{(|\boldsymbol{u}|^2 + \epsilon^2)^{1/2}}. \tag{3.85}$$

This smoothing of the modulus function enjoys three key properties

$$\kappa_\epsilon(0) = 0, \quad |\kappa(\boldsymbol{u}) - \kappa_\epsilon(\boldsymbol{u})| \leq \epsilon, \quad \text{and} \quad |\nabla \kappa_\epsilon(\boldsymbol{u})| \leq 3 \quad \forall\, \boldsymbol{u}. \tag{3.86}$$

That is, κ_ϵ satisfies the following three properties: it is integrable for integrable $\boldsymbol{u}$ with $\operatorname{supp}(\kappa_\epsilon[\boldsymbol{u}]) = \operatorname{supp}(\boldsymbol{u})$, it converges *uniformly* to κ in ϵ, and it has a uniformly bounded gradient. We therefore expect κ_ϵ to be numerically stable. The corresponding perturbed squared set distance error is denoted $E_\epsilon : L^2[\mathbb{R}^2, \mathbb{R}^2] \to \mathbb{R}_+$, and is given by

$$E_\epsilon[\boldsymbol{u}] = \sum_{m=0}^{M} \frac{\beta_m}{2} \left\| \frac{|\mathcal{F}_m[\boldsymbol{u}]|^2}{\left(|\mathcal{F}_m[\boldsymbol{u}]|^2 + \epsilon^2\right)^{1/2}} - \psi_m \right\|^2 \tag{3.87}$$

where $0 < \epsilon \ll 1$. Consistent with our observations about κ_ϵ, Property 3.3.1 establishes that $E_\epsilon[\boldsymbol{u}]$ is a continuous function of ϵ for fixed $\boldsymbol{u}$. Thus we expect this perturbed objective to be numerically stable. Indeed, we have found this perturbation to perform well in practice.

Define the integral functional $J[\cdot; b, \epsilon] : L^2[\mathbb{R}^2, \mathbb{R}^2] \to \overline{\mathbb{R}}$ by

$$J[\boldsymbol{u}; b, \epsilon] = \int_{\mathbb{R}^2} r^2(\boldsymbol{u}(\boldsymbol{x}); b(\boldsymbol{x}), \epsilon) d\boldsymbol{x} \tag{3.88}$$

where

$$r(\boldsymbol{u}; b, \epsilon) = \frac{|\boldsymbol{u}|^2}{(|\boldsymbol{u}|^2 + \epsilon^2)^{1/2}} - b. \tag{3.89}$$

Equivalently, define $h[\cdot; b, \epsilon] : L^2[\mathbb{R}^2, \mathbb{R}^2] \to \mathbb{R}$ by

$$h[\boldsymbol{u}; b, \epsilon] \equiv \int_{\mathbb{R}^2} - \frac{|\boldsymbol{u}(\boldsymbol{x})|^2}{(|\boldsymbol{u}(\boldsymbol{x})|^2 + \epsilon^2)^{1/2}} b(\boldsymbol{x}) d\boldsymbol{x}. \tag{3.90}$$

then

$$J[\boldsymbol{u}; b, \epsilon] = \left\| \frac{|\boldsymbol{u}|^2}{(|\boldsymbol{u}|^2 + \epsilon^2)^{1/2}} \right\|^2 + \|b\|^2 + 2h[\boldsymbol{u}; b]. \tag{3.91}$$

For $\boldsymbol{u}$, and ψ_m, satisfying Hyp.2.2.1 and for all ϵ, $J[\boldsymbol{u}; \psi_m, \epsilon]$ is finite-valued. Our focus is on the analytic properties of J. Accordingly we rewrite E_ϵ in composition form as

$$E_\epsilon[\boldsymbol{u}] \equiv \sum_{m=0}^{M} \frac{\beta_m}{2} \left(J[\cdot; \psi_m, \epsilon] \circ \mathcal{F}_m\right)[\boldsymbol{u}]. \tag{3.92}$$

Property 3.3.1 *Let* $\boldsymbol{u} \in L^2[\mathbb{R}^2, \mathbb{R}^2]$, *and* $b \in \mathbb{U}_+$ *defined in Hyp.2.2.1. The integral functional* $J[\boldsymbol{u}; b, \epsilon]$ *defined by Eq.(3.88) is continuous with respect to* ϵ.

PROOF: Let $\boldsymbol{u} \in L^2[\mathbb{R}^2, \mathbb{R}^2]$, $b \in \mathbb{U}_+$. From Eq.(3.88)-(3.39)

$$\lim_{\epsilon \to 0} J[\boldsymbol{u}; b, \epsilon] = \lim_{\epsilon \to 0} \int_{\mathbb{R}^2} r^2(\boldsymbol{u}(\boldsymbol{x}); b(\boldsymbol{x}), \epsilon) d\boldsymbol{x}.$$

Since $\boldsymbol{u}$ and b satisfy Hyp.2.2.1, by Hölder's Inequality for all ϵ,

$$\left|r^2(\boldsymbol{u}(\boldsymbol{x}); b(\boldsymbol{x}), \epsilon)\right| \leq |\boldsymbol{u}(\boldsymbol{x})|^2 + 2b(\boldsymbol{x})|\boldsymbol{u}(\boldsymbol{x})| + b^2(\boldsymbol{x}) \in L^1.$$

For fixed $\boldsymbol{x}$, $\boldsymbol{u}(\boldsymbol{x}) \in \mathbb{R}^2$, $b(\boldsymbol{x}) \in \mathbb{R}_+$, and $r(\cdot; \cdot, \epsilon)$ is continuous in ϵ. Thus by Lebesgue's Dominated Convergence Theorem $J[\boldsymbol{u}; b, \epsilon]$ is a continuous function of ϵ with

$$\lim_{\epsilon \to 0} J[\boldsymbol{u}; b, \epsilon] = J[\boldsymbol{u}, b; 0].$$

□

Continuity of E_ϵ with respect to ϵ is a consequence of Property 3.3.1 and the fact that the transforms $\mathcal{F}_m$ are unitary linear operators. We show in Section 3.3.3 that $J[\boldsymbol{u}; b, \epsilon]$, hence E_ϵ, is Fréchet differentiable with globally Lipschitz continuous derivative. This greatly facilitates the design and analysis of algorithms for the optimization problems discussed here.

The objective E_ϵ is nonconvex in $\boldsymbol{u}$, so convergence to a global minimum cannot be guaranteed. Nevertheless, convergence of line search methods to a local extremum are easily derived from standard results in the optimization literature. This is the topic of Chapter 5

3.3.2 Extended Least Squares

The projection algorithm Alg.(3.13) allows the user to choose the relaxation parameters $\alpha_m^{(\nu)}$ and weightings $\gamma_m^{(\nu)}$ at each iteration ν. This begs the question as to what the *optimal* choice of these parameters might be. Under the change of variables Eq.(3.15) one is similarly confronted with the issue of optimally selecting the *step-lengths* $\lambda^{(\nu)}$ and weights $\beta_m^{(\nu)}$. Step-lengths are discussed in Section 5.1. In this section we consider an approach to optimal weight selection. This requires an extension of Pr.(3.33) to accommodate variable weights.

Following the work of Bell *et al* [23], define the objective

$$L_\epsilon[\boldsymbol{u}, \boldsymbol{\beta}] = \sum_{m=0}^{M} -ln(2\pi\beta_m) + \beta_m \left(J[\mathcal{F}_m[\boldsymbol{u}]; \psi_m, \epsilon] + G_m[\boldsymbol{u}]\right). \tag{3.93}$$

where $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_M)$. This objective corresponds to the negative log likelihood measure for normally distributed data errors. The weight β_m is the variance of the data set ψ_m. The functional $G_m[\boldsymbol{u}]$ is a regularization term. For the purposes of illustrating the connection between projection methods and line search methods, the regularization that is used is simply a nonnegative constant $G_m[\boldsymbol{u}] = c_m > 0$. Each data set can be matched exactly using nonparametric techniques such as projection methods. The constant reflects prior belief about the reliability of the M data sets relative to one another. Given the data ψ_m, the estimates for the true value of the vector of parameters $\boldsymbol{u} \in L^2[\mathbb{R}^2, \mathbb{R}^2]$ and the vector of variances $\boldsymbol{\beta} \in \mathbb{R}_+^M$ are obtained as the solution to the problem

$$\begin{array}{ll} \text{minimize} & L_\epsilon[\boldsymbol{u}, \boldsymbol{\beta}] \\ \text{over} & \boldsymbol{u} \in L^2[\mathbb{R}^2, \mathbb{R}^2],\ 0 < \boldsymbol{\beta} \end{array} \tag{3.94}$$

A Benders decomposition is applied to solve for the optimal vector of weights, $\boldsymbol{\beta}_*$, in terms of $\boldsymbol{u}$.

Lemma 3.3.2 *Let* $L_\epsilon : L^2[\mathbb{R}^2, \mathbb{R}^2] \times \mathbb{R}_+^{M+1} \to \mathbb{R}$ *be defined by Eq.(3.93) and let* $\boldsymbol{u} \in L^2[\mathbb{R}^2, \mathbb{R}^2]$. *Let*

$$\boldsymbol{\beta}_*[\boldsymbol{u}] \equiv (\beta_{0*}[\boldsymbol{u}], \ldots, \beta_{M*}[\boldsymbol{u}]) , \tag{3.95}$$

where

$$\beta_{m*}[\boldsymbol{u}] = \left(J[\mathcal{F}_m[\boldsymbol{u}]; \psi_m, \epsilon] + c_m\right)^{-1} \quad \textit{for } m = 0, \ldots, M. \tag{3.96}$$

If $c_m > 0$ *then* $L_\epsilon[\boldsymbol{u}, \boldsymbol{\beta}_*[\boldsymbol{u}]] \le L_\epsilon[\boldsymbol{u}, \boldsymbol{\beta}]$ *for all* $\boldsymbol{\beta} > 0$.

PROOF: This is an infinite dimensional version of Lemma 1 of Bell *et al* [23]. Their proof also holds in this setting. □

Substituting $\boldsymbol{\beta}_*[\boldsymbol{u}]$ for $\boldsymbol{\beta}$ into Eq.(3.93) yields

$$L_\epsilon[\boldsymbol{u}, \boldsymbol{\beta}_*] = \sum_{m=0}^{M} \left[-ln(2\pi) + ln(\beta_{m*} + c_m) + 1 \right].$$

Dropping the constants yields the reduced objective

$$R_\epsilon[\boldsymbol{u}] = \sum_{m=0}^{M} ln(J[\mathcal{F}_m[\boldsymbol{u}]; \psi_m, \epsilon] + c_m). \tag{3.97}$$

The corresponding optimization problem is

$$\begin{array}{ll} \text{minimize} & R_\epsilon[\boldsymbol{u}] \\ \text{over} & \boldsymbol{u} \in L^2[\mathbb{R}^2, \mathbb{R}^2]. \end{array} \tag{3.98}$$

The analytic properties of the reduced extended least squares objective $R_\epsilon[\boldsymbol{u}]$ depend on those of the underlying integral operator J. The next section is concerned with establishing the analytic properties of the integral functional J and extending these to the perturbed set distance error E_ϵ and the extended least squares objected R_ϵ.

3.3.3 Classical Analysis

In this section we establish that the perturbed least squares error E_ϵ is Fréchet differentiable and calculate its derivative.

Before examining the analytic properties of E_ϵ, some facts of classical derivatives are developed. In the following, the set $\mathbb{X}_i$ denotes a real Hilbert space. The space of bounded linear mappings from $\mathbb{X}_1$ to $\mathbb{X}_2$ is denoted by $\mathcal{L}(\mathbb{X}_1, \mathbb{X}_2)$.

Definition 3.3.3 *Let* $F : \mathbb{X}_1 \to \mathbb{X}_2$. *For* $u \in \mathbb{X}_1$, *the* Gâteaux derivative *of* F *at* u, *if it exists, is an element* $DF(u) \in \mathcal{L}\{\mathbb{X}_1, \mathbb{X}_2\}$ *satisfying*

$$\lim_{t \searrow 0} \left\| \frac{F(u + tw) - F(u)}{t} - DF(u)(w) \right\|_2 = 0 \quad \forall\, w \in \mathbb{X}_1. \tag{3.99}$$

where $DF(u)(w)$ *denotes the action of* $DF(u)$ *on the element* w.

If the above limit is uniform with respect to w *on bounded subsets of* $\mathbb{X}_1$, *then* F *is said to be* Fréchet differentiable, *with the Fréchet derivative* $F'(u)$. *Equivalently,* $F'(u)$ *satisfies*

$$\lim_{||w||_2 \to 0} \frac{\|F(u + w) - F(u) - F'(u)(w)\|_2}{\|w\|_2} = 0. \tag{3.100}$$

The Fréchet and Gâteaux derivatives obey the classical sum rule of scalar calculus. The composition of Fréchet differentiable functions is Fréchet differentiable and obeys the *Chain Rule.* Given $F : \mathbb{X}_1 \to \mathbb{X}_2$ and $G : \mathbb{X}_2 \to \mathbb{X}_3$, suppose that F is Fréchet differentiable at

$u \in \mathbb{X}_1$ and G is Fréchet differentiable at $F(x) \in \mathbb{X}_2$. The composition $G \circ F : \mathbb{X}_1 \to \mathbb{X}_3$ is Fréchet differentiable at u and is given by

$$(G \circ F)'[u] = G'[F[u]]F'[u],$$

where $G'[F[u]]F'[u] \in \mathcal{L}(\mathbb{X}_1, \mathbb{X}_3)$ is the composition of $F'[u]$ with $G'[F[u]]$.

As in Section 3.2.5 the discussion is restricted to proper l.s.c. functions. From the definition, a Fréchet differentiable function is continuous. This need not be true for a Gâteaux differentiable function. Moreover, continuity of the Gâteaux derivative is not a sufficient condition for a mapping to be Fréchet differentiable. These points are illustrated in the following examples.

Example 3.3.4

(a) *The function $f(x) = (x + \epsilon)^{-1/2}$ is analytic as a function on $\mathbb{R}_+$. In one dimension the Fréchet and Gâteaux derivatives coincide with the classical derivative, thus $f(x)$ is infinitely Fréchet differentiable on $\mathbb{R}_+$.*

(b) *(Ex.11.20 of [45]) Let $f : \mathbb{R}^2 \to \mathbb{R}$ be defined by*

$$f(x,y) = \begin{cases} \frac{y^2}{x} & \text{if } x \neq 0, \\ 0 & \text{if } x = 0 \end{cases} .$$

The Gâteaux derivative at the origin is $Df(0,0) = 0$, though this function is not continuous there.

(c) *The modulus-squared is analytic as a function on $\mathbb{R}^n$: for $\boldsymbol{x} \in \mathbb{R}^n$, $f(\boldsymbol{x}) = |\boldsymbol{x}|^2 = \|\boldsymbol{x}\|_2^2$. This is twice continuously Fréchet differentiable with higher-order derivatives identically equal to zero.*

(d) *Let*

$$h(\boldsymbol{x}) = (g \circ f)(\boldsymbol{x}) = (|\boldsymbol{x}|^2 + \epsilon^2)^{-1/2},$$

where $g(y) = (y + \epsilon)^{-1/2}$ for $y \in \mathbb{R}_+$ and $f(\boldsymbol{x}) = |\boldsymbol{x}|^2$ for $\boldsymbol{x} \in \mathbb{R}^n$. This is the composition of two analytic functions, thus is itself analytic.

For proper, lower semi-continuous scalar functions $f : \mathbb{X} \to \overline{\mathbb{R}}$, the *Mean Value Theorem* can be stated as follows. Suppose f is Gâteaux differentiable on an open neighborhood $\mathbb{U} \subset \mathbb{X}$. Then for every $u, v \in \mathbb{U}$, there is a point $w = tu + (1-t)v$, $0 < t < 1$, such that

$$f(v) - f(u) = Df(w)(v - u).$$

The proof follows from application of the Mean Value Theorem to the function $\theta : [0,1] \to \mathbb{R}$, $\theta(t) = f(u + t(v - u))$. The same result holds if f is Fréchet differentiable with the Fréchet derivative replacing the Gâteaux derivative in the mean value expression. Suppose further that the mapping $f'(\cdot) : \mathbb{U} \to \mathbb{X}$ is itself Fréchet differentiable on $\mathbb{U}$ with the Fréchet derivative at $u \in \mathbb{U}$ denoted by $D^2 f(u) \in \mathcal{L}(\mathbb{X}, \mathbb{X})$. Again, by the Mean Value Theorem, for all $u, v \in \mathbb{U}$, there is a point $w = tu + (1-t)v$, $0 < t < 1$, such that

$$f'(v) - f'(u) = f''(w)(v - u),$$

where $f''(w)(v-u)$ denotes the action of $f''(w) : \mathbb{X} \to \mathbb{X}$ on $v - u$. Thus, for each $u \in \mathbb{U}$, f admits a local second-order Taylor expansion with remainder, *i.e* there exists a ball $\mathbb{B}(\epsilon)$ of radius ϵ such that for all $\|v - u\| < \epsilon$ the following holds:

$$f(v) = f(u) + f'(u)(v-u) + \frac{1}{2} f''(w)(v-u, v-u),$$

where $w = tu + (1-t)v$ and $f''(w)(v-u, v-u)$ denotes the action of $f''(w)(v-u) : \mathbb{X} \to \overline{\mathbb{R}}$ on $v - u$.

The above results are extended to integral functionals of the form

$$J[u] = \int_{setR^n} f(x, u(x)) dx,$$

where $u \in L^2(\mathbb{R}^n, \mathbb{R}^m)$ and $f : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$.

Theorem 3.3.5 *Let* $f : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ *satisfy:*

1. $f(\cdot, u(\cdot))$ *is integrable on* $\mathbb{R}^n$ *for all* $u(\cdot) \in L^2[\mathbb{R}^n, \mathbb{R}^m]$*;*

2. *for all* $x \in \mathbb{R}^n$, $f(x, u)$ *is Gâteaux differentiable with respect to* u *as a function on* $\mathbb{R}^n \times \mathbb{R}^m$ *with Gâteaux derivative denoted by*

$$D_u f(x, u);$$

3. *There exists a* K *such that for all* x, $D_u f(x, \cdot)$ *is globally Lipschitz on* $\mathbb{R}^m$ *with Lipschitz constant* K.

Define the integral functional $J : L^2[\mathbb{R}^n, \mathbb{R}^m] \to \mathbb{R}$ *by*

$$J[u] = \int_{\mathbb{R}^n} f(x, u(x)) \, dx.$$

Then $J[u]$ *is Fréchet differentiable as a function on* $L^2[\mathbb{R}^n, \mathbb{R}^m]$ *with Fréchet derivative*

$$J'[u][w] = \int_{\mathbb{R}^n} D_u f(x, u(x))(w(x)) dx. \tag{3.101}$$

Moreover, the Fréchet derivative J' *is Lipschitz continuous on* $L^2[\mathbb{R}^n, \mathbb{R}^m]$ *with constant* K.

PROOF:

$$\begin{aligned}\left|J[u+w]-J[u]-\int_{\mathbb{R}^n} D_u f(x,u(x))(w(x))dx\right| \\ \leq \int_{\mathbb{R}^n} |f(x,u(x)+w(x))-f(x,u(x))-D_u f(x,u(x))(w(x))|\, dx.\end{aligned}$$

For fixed x

$$\begin{aligned}&|f(x,u(x)+w(x))-f(x,u(x))-D_u f(x,u(x))(w(x))| \\ &\qquad \leq \int_0^1 |D_u f(x,u(x)+\tau w(x))(w(x))-D_u f(x,u(x))(w(x))|\ d\tau \\ &\qquad \leq \int_0^1 |D_u f(x,u(x)+\tau w(x))-D_u f(x,u(x))|\ |w(x)|\ d\tau.\end{aligned}$$

Since $D_u f$ is globally Lipschitz continuous with constant K, for all u and x

$$|D_u f(x,u(x)+\tau w(x))-D_u f(x,u(x))| \leq K\tau\, |w(x)|\,,$$

thus

$$\begin{aligned}\int_0^1 |D_u f(x,u(x)+\tau w(x))(w(x))-D_u f(x,u(x))(w(x))|\, d\tau &\leq \int_0^1 K\tau |w(x)|^2 d\tau \\ &= \frac{K}{2}|w(x)|^2,\end{aligned}$$

hence

$$\begin{aligned}\left|J[u+w]-J[u]-\int_{\mathbb{R}^n} D_u f(x,u(x))(w(x))dx\right| &\leq \int_{\mathbb{R}^n} \frac{K}{2}|w(x)|^2 dx \\ &= \frac{K}{2}\|w\|^2.\end{aligned}$$

Consequently, J is Fréchet differentiable with $J'[u][w]$ given by Eq.(3.101).

Since L^2 is a Hilbert space, the kernel of the integral operator $J'[u]$ is equal to $D_u f(\cdot,u(\cdot))$. Thus if $D_u f(x,u(x))$ is globally Lipschitz with respect to u with constant K for all x then $J'[u]$ is globally Lipschitz with constant K. □

Remark 3.3.6 *Conditions 2 and 3 in Theorem 3.3.5 imply that, for all $x \in \mathbb{R}^n$, the integrand $f(x,u)$ is Fréchet differentiable with respect to u as a function on $\mathbb{R}^n \times \mathbb{R}^m$. It is not true in general that Gâteaux differentiability implies Fréchet differentiability. See [45, Ex.1.11.20] for a counter example. Moreover, it is not true in general that a Fréchet differentiable function has a globally Lipschitz continuous Fréchet derivative. We state Theorem 3.3.5 in terms of Gâteaux differentiable functions instead of Fréchet differentiable functions because it is often easier to show Gâteaux differentiability than it is to show Fréchet differentiability.*

Remark 3.3.7 *Since* $L^2[\mathbb{R}^n,\mathbb{R}^m]$ *is a Hilbert space, the derivative of* $J[u]$ *also belongs to* $L^2[\mathbb{R}^n,\mathbb{R}^m]$*. We denote this mapping by* $\nabla J[u]$*.*

To apply the above results to E_ϵ it remains to be shown that the squared residual $r^2(\boldsymbol{u}; b, \epsilon)$ defined by Eq.(3.89) is Gâteaux differentiable and globally Lipschitz with respect to $\boldsymbol{u}$ for all $\boldsymbol{x}$. Gâteaux differentiability of $r^2(\boldsymbol{u}; b, \epsilon)$ with respect to $\boldsymbol{u}$ for $\epsilon > 0$ follows from elementary vector calculus. In fact, $r^2(\boldsymbol{u}; b, \epsilon)$ is *analytic.* The derivative is denoted by $Dr^2(\boldsymbol{u}; b, \epsilon) \in \mathcal{L}(\mathbb{R}^2, \mathbb{R})$ and defined by

$$Dr^2(\boldsymbol{u}; b, \epsilon)(\boldsymbol{w}) \equiv 2r(\boldsymbol{u}; b, \epsilon) Dr(\boldsymbol{u}; b, \epsilon) \tag{3.102}$$

where

$$D_{\boldsymbol{u}} r(\boldsymbol{u}; b, \epsilon) \equiv \frac{|\boldsymbol{u}|^2 + 2\epsilon^2}{(|\boldsymbol{u}|^2 + \ \epsilon^2)^{3/2}} \boldsymbol{u}^T. \tag{3.103}$$

The next lemma shows that $D_{\boldsymbol{u}} r^2(\boldsymbol{u}; b, \epsilon)$ is globally Lipschitz continuous.

Lemma 3.3.8 *The derivative* $D_{\boldsymbol{u}} r^2$ *for* r *defined by Eq.(3.39) is globally Lipschitz continuous on* $\mathbb{R}^2$ *with global Lipschitz constant*

$$K = 16 + \frac{12}{\epsilon}|b|. \tag{3.104}$$

PROOF: The setting here is finite dimensional. The finite dimensional norm is assumed to be the 2-norm and is denoted by $|\cdot|$. From Eq.(3.102)-(3.103), for $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^2$,

$$\begin{aligned} D(r^2(\boldsymbol{u}; b, \epsilon)) &= 2 \frac{\left(|\boldsymbol{u}|^2 + 2\epsilon^2\right) r(\boldsymbol{u}; b, \epsilon)}{(|\boldsymbol{u}|^2 + \epsilon^2)^{3/2}} \boldsymbol{u} \\ &= 2 \left[\frac{|\boldsymbol{u}|^2 \boldsymbol{u}}{|\boldsymbol{u}|^2 + \epsilon^2} + \epsilon^2 \frac{|\boldsymbol{u}|^2 \boldsymbol{u}}{(|\boldsymbol{u}|^2 + \epsilon^2)^2} - \frac{b\boldsymbol{u}}{(|\boldsymbol{u}|^2 + \epsilon^2)^{1/2}} - \epsilon^2 \frac{b\boldsymbol{u}}{(|\boldsymbol{u}|^2 + \epsilon^2)^{3/2}} \right]. \end{aligned} \tag{3.105}$$

We proceed by calculating the Lipschitz constant for each of the terms in Eq.(3.105). Each of these terms takes the form

$$\frac{|\boldsymbol{u}|^p \boldsymbol{u}}{(|\boldsymbol{u}|^2 + \epsilon^2)^q}.$$

The Lipschitz constant is obtained by bounding terms of the form

$$\left| \frac{|\boldsymbol{u}|^p \boldsymbol{u}}{(|\boldsymbol{u}|^2 + \epsilon^2)^q} - \frac{|\boldsymbol{v}|^p \boldsymbol{v}}{(|\boldsymbol{v}|^2 + \epsilon^2)^q} \right| \tag{3.106}$$

Add and subtract $\frac{|\boldsymbol{v}|^p \boldsymbol{u}}{(|\boldsymbol{v}|^2+\epsilon^2)^q}$ to obtain

$$\begin{aligned} &\left| \frac{|\boldsymbol{u}|^p \boldsymbol{u}}{(|\boldsymbol{u}|^2 + \epsilon^2)^q} - \frac{|\boldsymbol{v}|^p \boldsymbol{v}}{(|\boldsymbol{v}|^2 + \epsilon^2)^q} \right| \\ &= \left| \left(\frac{|\boldsymbol{u}|^p}{(|\boldsymbol{u}|^2 + \epsilon^2)^q} - \frac{|\boldsymbol{v}|^p}{(|\boldsymbol{v}|^2 + \epsilon^2)^q} \right) \boldsymbol{u} + \frac{|\boldsymbol{v}|^p}{(|\boldsymbol{v}|^2 + \epsilon^2)^q} (\boldsymbol{u} - \boldsymbol{v}) \right| \\ &\leq \left| \frac{|\boldsymbol{u}|^p (|\boldsymbol{v}|^2 + \epsilon^2)^q - |\boldsymbol{v}|^p (|\boldsymbol{u}|^2 + \epsilon^2)^q}{(|\boldsymbol{u}|^2 + \epsilon^2)^q (|\boldsymbol{v}|^2 + \epsilon^2)^q} \boldsymbol{u} \right| + \frac{|\boldsymbol{v}|^p}{(|\boldsymbol{v}|^2 + \epsilon^2)^q} |\boldsymbol{v} - \boldsymbol{u}|. \end{aligned} \tag{3.107}$$

Unfortunately this general form must be analyzed case by case. We examine 3 different cases.

Case 1, p=2, q=1:

$$\begin{aligned}\left|\frac{|\boldsymbol{u}|^2\boldsymbol{u}}{|\boldsymbol{u}|^2+\epsilon^2}-\frac{|\boldsymbol{v}|^2\boldsymbol{v}}{|\boldsymbol{v}|^2+\epsilon^2}\right| &\le \left|\frac{|\boldsymbol{u}|^2(|\boldsymbol{v}|^2+\epsilon^2)-|\boldsymbol{v}|^2(|\boldsymbol{u}|^2+\epsilon^2)}{(|\boldsymbol{u}|^2+\epsilon^2)(|\boldsymbol{v}|^2+\epsilon^2)}\boldsymbol{u}\right|+\frac{|\boldsymbol{v}|^2}{|\boldsymbol{v}|^2+\epsilon^2}|\boldsymbol{v}-\boldsymbol{u}| \\ &\le \epsilon^2\left|\frac{|\boldsymbol{u}|^2-|\boldsymbol{v}|^2}{(|\boldsymbol{u}|^2+\epsilon^2)^{1/2}(|\boldsymbol{v}|^2+\epsilon^2)}\right|+|\boldsymbol{u}-\boldsymbol{v}|,\end{aligned}$$

where we have used the inequality

$$\left|\frac{\boldsymbol{u}}{(|\boldsymbol{u}|^2+\epsilon^2)^{1/2}}\right|\le 1 \tag{3.108}$$

Without loss of generality assume that $|\boldsymbol{v}|\le|\boldsymbol{u}|$. Then

$$|\boldsymbol{u}|^2-|\boldsymbol{v}|^2\le 2|\boldsymbol{u}|\;|\boldsymbol{v}-\boldsymbol{u}|\quad\text{for}\quad |\boldsymbol{v}|\le|\boldsymbol{u}|. \tag{3.109}$$

Using this, inequality (3.108) and

$$\frac{1}{|\boldsymbol{v}|^2+\epsilon^2}\le\frac{1}{\epsilon^2} \tag{3.110}$$

yields the bound

$$\begin{aligned}\left|\frac{|\boldsymbol{u}|^2\boldsymbol{u}}{|\boldsymbol{u}|^2+\epsilon^2}-\frac{|\boldsymbol{v}|^2\boldsymbol{v}}{|\boldsymbol{v}|^2+\epsilon^2}\right| &\le \left(\frac{2|\boldsymbol{u}|\epsilon^2}{(|\boldsymbol{u}|^2+\epsilon^2)^{1/2}(|\boldsymbol{v}|^2+\epsilon^2)}+1\right)|\boldsymbol{v}-\boldsymbol{u}| \\ &\le 3|\boldsymbol{v}-\boldsymbol{u}|.\end{aligned} \tag{3.111}$$

Case 2, p=2, q=2:
As in Case 1, assume without loss of generality that $|\boldsymbol{v}|\le|\boldsymbol{u}|$. Then the inequalities (3.108)-(3.110) yield

$$\begin{aligned}&\left|\frac{|\boldsymbol{u}|^2\boldsymbol{u}}{(|\boldsymbol{u}|^2+\epsilon^2)^2}-\frac{|\boldsymbol{v}|^2\boldsymbol{v}}{(|\boldsymbol{v}|^2+\epsilon^2)^2}\right| \\ &\le \left|\frac{|\boldsymbol{u}|^2(|\boldsymbol{v}|^2+\epsilon^2)^2-|\boldsymbol{v}|^2(|\boldsymbol{u}|^2+\epsilon^2)^2}{(|\boldsymbol{u}|^2+\epsilon^2)^2(|\boldsymbol{v}|^2+\epsilon^2)^2}\boldsymbol{u}\right|+\frac{|\boldsymbol{v}|^2}{(|\boldsymbol{v}|^2+\epsilon^2)^2}|\boldsymbol{v}-\boldsymbol{u}| \\ &= \left|\frac{(|\boldsymbol{u}|^2|\boldsymbol{v}|^2-\epsilon^4)(|\boldsymbol{v}|^2-|\boldsymbol{u}|^2)}{(|\boldsymbol{u}|^2+\epsilon^2)^2(|\boldsymbol{v}|^2+\epsilon^2)^2}\boldsymbol{u}\right|+\frac{|\boldsymbol{v}|^2}{(|\boldsymbol{v}|^2+\epsilon^2)^2}|\boldsymbol{v}-\boldsymbol{u}| \\ &\le \left(\frac{2|\boldsymbol{u}|^2\left(|\boldsymbol{u}|^2|\boldsymbol{v}|^2+\epsilon^2\right)}{(|\boldsymbol{u}|^2+\epsilon^2)^2(|\boldsymbol{v}|^2+\epsilon^2)^2}+\frac{1}{\epsilon^2}\right)|\boldsymbol{v}-\boldsymbol{u}| \\ &\le \left(\frac{2}{\epsilon^2}\frac{\left(|\boldsymbol{u}|^2|\boldsymbol{v}|^2+\epsilon^2\right)}{(|\boldsymbol{u}|^2+\epsilon^2)(|\boldsymbol{v}|^2+\epsilon^2)}+\frac{1}{\epsilon^2}\right)|\boldsymbol{v}-\boldsymbol{u}| \\ &\le \frac{5}{\epsilon^2}|\boldsymbol{v}-\boldsymbol{u}|.\end{aligned} \tag{3.112}$$

Case 3, p=0, q=n/2:

$$\begin{aligned}
&\left|\frac{\boldsymbol{u}}{(|\boldsymbol{u}|^2+\epsilon^2)^{n/2}} - \frac{\boldsymbol{v}}{(|\boldsymbol{v}|^2+\epsilon^2)^{n/2}}\right| \\
&\qquad \le \left|\frac{(|\boldsymbol{v}|^2+\epsilon^2)^{n/2} - (|\boldsymbol{u}|^2+\epsilon^2)^{n/2}}{(|\boldsymbol{u}|^2+\epsilon^2)^{n/2}(|\boldsymbol{v}|^2+\epsilon^2)^{n/2}}\boldsymbol{u}\right| + \frac{1}{(|\boldsymbol{v}|^2+\epsilon^2)^{n/2}}|\boldsymbol{u}-\boldsymbol{v}| \\
&\qquad \le \left|\frac{\left(|\boldsymbol{v}|^2+\epsilon^2\right)^{n/2} - \left(|\boldsymbol{u}|^2+\epsilon^2\right)^{n/2}}{\left(|\boldsymbol{u}|^2+\epsilon^2\right)^{(n-1)/2}\left(|\boldsymbol{v}|^2+\epsilon^2\right)^{n/2}}\right| + \frac{1}{\epsilon^n}\left|\boldsymbol{u}-\boldsymbol{v}\right|.
\end{aligned}$$

The last expression uses inequalities (3.108) and (3.110). By the Mean Value Theorem there exists a $w \in [|\boldsymbol{v}|, |\boldsymbol{u}|]$ such that

$$\left(|\boldsymbol{v}|^2+\epsilon^2\right)^{n/2} - \left(|\boldsymbol{u}|^2+\epsilon^2\right)^{n/2} = nw\left(w^2+\epsilon^2\right)^{n/2-1}(|\boldsymbol{v}|-|\boldsymbol{u}|)$$

and so

$$\begin{aligned}
\left|\left(|\boldsymbol{v}|^2+\epsilon^2\right)^{n/2} - \left(|\boldsymbol{u}|^2+\epsilon^2\right)^{n/2}\right| &\le nw\left(w^2+\epsilon^2\right)^{n/2-1}|\boldsymbol{v}-\boldsymbol{u}| \\
&\le n\left(w^2+\epsilon\right)^{(n-1)/2}|\boldsymbol{v}-\boldsymbol{u}|.
\end{aligned}$$

This yields

$$\begin{aligned}
\left|\frac{\left(|\boldsymbol{v}|^2+\epsilon^2\right)^{n/2} - \left(|\boldsymbol{u}|^2+\epsilon^2\right)^{n/2}}{\left(|\boldsymbol{u}|^2+\epsilon^2\right)^{(n-1)/2}\left(|\boldsymbol{v}|^2+\epsilon^2\right)^{n/2}}\right| &\le \left|\frac{n|\boldsymbol{v}-\boldsymbol{u}|\left(w^2+\epsilon\right)^{(n-1)/2}}{(|\boldsymbol{u}|^2+\epsilon^2)^{(n-1)/2}(|\boldsymbol{v}|^2+\epsilon^2)^{n/2}}\right| \\
&\le \frac{n}{\epsilon^n}|\boldsymbol{v}-\boldsymbol{u}|.
\end{aligned}$$

Finally,

$$\left|\frac{\boldsymbol{u}}{(|\boldsymbol{u}|^2+\epsilon^2)^{n/2}} - \frac{\boldsymbol{v}}{(|\boldsymbol{v}|^2+\epsilon^2)^{n/2}}\right| \le \frac{n+1}{\epsilon^n}|\boldsymbol{v}-\boldsymbol{u}|. \tag{3.113}$$

Cases 1-3 yield the following global bound which completes the proof

$$\left|Dr^2(\boldsymbol{u};b,\epsilon) - Dr^2(\boldsymbol{v};b,\epsilon)\right| \le \left(16 + \frac{12}{\epsilon}|b|\right)|\boldsymbol{u}-\boldsymbol{v}|.$$

□

The constant K given by in Eq.(3.104) is the pointwise Lipschitz constant for the Gâteaux derivative of the functional $r^2(\boldsymbol{u}(\boldsymbol{x}); b(\boldsymbol{x}), \epsilon)$. If $b \in L^\infty$ then for all $\boldsymbol{x}$

$$\left|D_{\boldsymbol{u}}r^2(\boldsymbol{u}(\boldsymbol{x}); b(\boldsymbol{x}),\epsilon) - D_{\boldsymbol{u}}r^2(\boldsymbol{v}(\boldsymbol{x}); b(\boldsymbol{x}),\epsilon)\right| \le \left(16 + \frac{12}{\epsilon}\|b\|_\infty\right)|\boldsymbol{u}(\boldsymbol{x}) - \boldsymbol{v}(\boldsymbol{x})|. \tag{3.114}$$

We can therefore apply Theorem 3.3.5 to the integral operator J defined by Eq.(3.88) for $\epsilon > 0$ to obtain the Fréchet derivative

$$J'[\boldsymbol{u}; b, \epsilon][\boldsymbol{w}] = \int_{\mathbb{R}^2} \left(D_{\boldsymbol{u}} r^2(\boldsymbol{u}(\boldsymbol{x}); b(\boldsymbol{x}), \epsilon),\ \boldsymbol{w}(\boldsymbol{x})\right) d\boldsymbol{x}$$

where $(\cdot,\ \cdot)$ denotes the standard finite dimensional inner product. Equations (3.39), (3.102) and (3.103) yield the gradient of J at $\boldsymbol{u}$

$$\nabla J[\boldsymbol{u}; b, \epsilon] = 2\left(\frac{|\boldsymbol{u}|^2}{(|\boldsymbol{u}|^2 + \epsilon^2)^{1/2}} - b\right) \frac{|\boldsymbol{u}|^2 + 2\epsilon^2}{(|\boldsymbol{u}|^2 +\ \epsilon^2)^{3/2}} \boldsymbol{u}. \tag{3.115}$$

By Lemma 3.3.8, Eq.(3.114), and Theorem 3.3.5, $\nabla J[\boldsymbol{u}; b, \epsilon]$ is globally Lipschitz continuous with global Lipschitz constant

$$K_{\nabla J} = \left(16 + \frac{12}{\epsilon}\|b\|_\infty\right). \tag{3.116}$$

The preceding results extend immediately to the perturbed squared set distance error $E_\epsilon[\boldsymbol{u}]$ defined by Eq.(3.92). Since $\mathcal{F}_m[\boldsymbol{u}]$ defined by Eq.(2.43) and Eq.(2.48) is a linear operator on L^2 it is Fréchet differentiable there with Fréchet derivative given by

$$\mathcal{F}'_m[\boldsymbol{u}][\boldsymbol{w}] \quad = \quad \mathcal{F}_m[\boldsymbol{w}].$$

For $\boldsymbol{u}$ and ψ_m satisfying Hypothesis 2.2.1, Theorem 3.3.5 together with the Chain Rule for Fréchet differentiable functions and Eq.(3.115) yields

$$\begin{aligned}
\left(J[\cdot; \psi_m, \epsilon] \circ \mathcal{F}_m[\boldsymbol{u}]\right)' [\boldsymbol{w}] &= J'[\mathcal{F}_m[\boldsymbol{u}]; \psi_m, \epsilon][\mathcal{F}'_m[\boldsymbol{u}][\boldsymbol{w}]] \\
&= \langle \nabla J[\mathcal{F}_m[\boldsymbol{u}]; \psi_m, \epsilon], \mathcal{F}_m[\boldsymbol{w}] \rangle \\
&= 2 \left\langle \mathcal{F}^*_m \left[r[\mathcal{F}_m[\boldsymbol{u}]; \psi_m, \epsilon] \frac{|\mathcal{F}_m[\boldsymbol{u}]|^2 + 2\epsilon^2}{(|\mathcal{F}_m[\boldsymbol{u}]|^2 +\ \epsilon^2)^{3/2}} \mathcal{F}_m[\boldsymbol{u}] \right],\ \boldsymbol{w} \right\rangle
\end{aligned} \tag{3.117}$$

for $m = 0, \ldots, M$. Thus

$$\nabla \left(J[\cdot; \psi_m, \epsilon] \circ \mathcal{F}_m[\boldsymbol{u}]\right) = 2\mathcal{F}^* \left[r[\mathcal{F}_m[\boldsymbol{u}]; \psi_m, \epsilon] \frac{|\mathcal{F}_m[\boldsymbol{u}]|^2 + 2\epsilon^2}{(|\mathcal{F}_m[\boldsymbol{u}]|^2 +\ \epsilon^2)^{3/2}} \mathcal{F}_m[\boldsymbol{u}] \right]. \tag{3.118}$$

Extending this to $E_\epsilon[\boldsymbol{u}]$ we have

$$E'[\boldsymbol{u}; \epsilon][\boldsymbol{w}] = \sum_{m=0}^{M} \frac{\beta_m}{2} \left(J[\cdot; \psi_m, \epsilon] \circ \mathcal{F}_m\right)' [\boldsymbol{u}][\boldsymbol{w}]. \tag{3.119}$$

where, by Eq.(3.119) and (3.118),

$$\nabla E_\epsilon[\boldsymbol{u}] = \sum_{m=0}^{M} \frac{\beta_m}{2} \nabla \left(J[\cdot; \psi_m, \epsilon] \circ \mathcal{F}_m\right) [\boldsymbol{u}]. \tag{3.120}$$

By Parseval's relation, Eq.(3.116) and the triangle inequality, the global Lipschitz constant $K_{\nabla E_\epsilon}$ for $\nabla E_\epsilon[\boldsymbol{u}]$ is

$$K_{\nabla E_\epsilon} = \sum_{m=0}^{M} \beta_m \left(8 + \frac{6\|\psi_m\|_\infty}{\epsilon} \right).$$

Similarly, the extended least squares objective $R_\epsilon[\boldsymbol{u}]$ defined by Eq.(3.97) is Fréchet differentiable with derivative given by

$$R'_\epsilon[\boldsymbol{u}][\boldsymbol{w}] = \sum_{m=0}^{M} \left(J[\mathcal{F}_m[\boldsymbol{u}]; \psi_m, \epsilon] + c_m\right)^{-1} \left(J[\cdot; \psi_m, \epsilon] \circ \mathcal{F}_m\right)' [\boldsymbol{u}][\boldsymbol{w}]. \tag{3.121}$$

For $\nabla \left(J[\cdot; \psi_m, \epsilon] \circ \mathcal{F}_m\right)[\boldsymbol{u}]$ given by Eq.(3.118), Eq.(3.117) and Eq.(3.96) yield

$$\begin{aligned} \nabla R_\epsilon[\boldsymbol{u}] &= \sum_{m=0}^{M} \left(\left(J[\cdot; \psi_m, \epsilon] \circ \mathcal{F}_m\right)[\boldsymbol{u}] + c_m\right)^{-1} \nabla \left(J[\cdot; \psi_m, \epsilon] \circ \mathcal{F}_m\right)[\boldsymbol{u}] \\ &= 2 \sum_{m=0}^{M} \beta_{m*}[\boldsymbol{u}] \mathcal{F}_m^* \left[r[\mathcal{F}_m[\boldsymbol{u}]; \psi_m, \epsilon] \frac{|\mathcal{F}_m[\boldsymbol{u}]|^2 + 2\epsilon^2}{\left(|\mathcal{F}_m[\boldsymbol{u}]|^2 + \epsilon^2\right)^{3/2}} \mathcal{F}_m[\boldsymbol{u}] \right]. \end{aligned} \tag{3.122}$$

Together with the fact that $ln(x + c_m)$ has a derivative bounded by $1/c_m$ on $\mathbb{R}_+$, Eq.(3.116) yields the global Lipschitz constant $K_{\nabla R_\epsilon}$ for ∇R_ϵ

$$K_{\nabla R_\epsilon} = \sum_{m=0}^{M} \frac{1}{c_m} \left(16 + \frac{12\|\psi_m\|_\infty}{\epsilon} \right).$$

The next property establishes the principle relationship between ∇E_ϵ and the operator $\mathcal{G}$ given by Eq.(3.17).

Property 3.3.9 *Let the functions* $\boldsymbol{u}$ *and* ψ_m *satisfy Hypothesis 2.2.1. At each* $\boldsymbol{u}$ *with* $E[\boldsymbol{u}] < \delta$, *there exists an* $\epsilon > 0$ *such that*

$$\|\nabla E_\epsilon[\boldsymbol{u}] \; - \; \boldsymbol{v}\| \; < \; C\delta^{1/2}, \tag{3.123}$$

for all $\boldsymbol{v} \in \; \mathcal{G}[\boldsymbol{u}]$ *where*

$$\mathcal{G} = \sum_{m=0}^{M} \beta_m \left(\mathcal{I} - \Pi_{\mathbb{Q}_m}\right)$$

and

$$C = \sqrt{2} \sum_{m=0}^{M} \beta_m^{1/2} \left(1 + \sqrt{2}\beta_m^{1/2}\right). \tag{3.124}$$

PROOF: The theorem follows from careful splitting of the norm and repeated application of Lebesgue's Dominated Convergence Theorem. Define

$$\mathbb{G}_m = \operatorname{supp}\left(\mathcal{F}_m[\boldsymbol{u}]\right), \; m = 0, 1, \ldots$$

Denote the complements of these sets by $\widetilde{\mathbb{G}}_m$. Denote the norm over the domain $\Omega \subset \mathbb{R}^2$ by

$$\|\cdot\|_\Omega \equiv \|\cdot\,\mathcal{X}_\Omega\|$$

where $\mathcal{X}_\Omega$ is the indicator function for Ω defined by Eq.(2.20). Let $\boldsymbol{v}_m \in \Pi_{\mathbb{Q}_m}[\boldsymbol{u}]$, $m = 0, 1, 2, \ldots$, and let $\boldsymbol{v} = \sum_{m=0}^{M} \beta_m(\boldsymbol{u} - \boldsymbol{v}_m)$. Then

$$\begin{aligned}
\|\nabla E_\epsilon[\boldsymbol{u}] - \boldsymbol{v}\| &\leq \sum_{m=0}^{M} \left\| \frac{\beta_m}{2} \nabla J[\mathcal{F}_m[\boldsymbol{u}]; \psi_m, \epsilon] - \beta_m(\boldsymbol{u} - \boldsymbol{v}_m) \right\| \\
&= \sum_{m=0}^{M} \beta_m \left\| \mathcal{F}_m^* \left[r\left[\mathcal{F}_m[\boldsymbol{u}]; \psi_m, \epsilon\right] \nabla r\left[\mathcal{F}_m[\boldsymbol{u}]; \psi_m, \epsilon\right] \mathcal{F}_m[\boldsymbol{u}] \right] - (\boldsymbol{u} - \boldsymbol{v}_m) \right\| \\
&= \sum_{m=0}^{M} \beta_m \left\| r\left[\mathcal{F}_m[\boldsymbol{u}]; \psi_m, \epsilon\right] \nabla r\left[\mathcal{F}_m[\boldsymbol{u}]; \psi_m, \epsilon\right] \mathcal{F}_m[\boldsymbol{u}] - \mathcal{F}_m[\boldsymbol{u} - \boldsymbol{v}_m] \right\| \\
&= \sum_{m=0}^{M} \beta_m \left\| r\left[\mathcal{F}_m[\boldsymbol{u}]; \psi_m, \epsilon\right] \nabla r\left[\mathcal{F}_m[\boldsymbol{u}]; \psi_m, \epsilon\right] \mathcal{F}_m[\boldsymbol{u}] - \mathcal{F}_m[\boldsymbol{u} - \boldsymbol{v}_m] \right\|_{\mathbb{G}_m} \\
&\qquad\qquad + \beta_m \|\mathcal{F}_m[\boldsymbol{v}_m]\|_{\widetilde{\mathbb{G}}_m} \; .
\end{aligned}$$

Now, by the definition of $\mathbb{Q}_m$ Eq.(3.1), $|\mathcal{F}_m[\boldsymbol{v}_m]| = \psi_m$. Also, on $\mathbb{G}_m[\boldsymbol{u}]$ we have $\mathcal{F}_m[\boldsymbol{v}_m] = \frac{\mathcal{F}_m[\boldsymbol{u}]}{|\mathcal{F}_m[\boldsymbol{u}]|}\psi_m$ which yields the inequality

$$\begin{aligned}
&\|\nabla E_\epsilon[\boldsymbol{u}] - \boldsymbol{v}\| \leq \\
&\sum_{m=0}^{M} \beta_m \left\| r\left[\mathcal{F}_m[\boldsymbol{u}]; \psi_m, \epsilon\right] \nabla r\left[\mathcal{F}_m[\boldsymbol{u}]; \psi_m, \epsilon\right] |\mathcal{F}_m[\boldsymbol{u}]| - (|\mathcal{F}_m[\boldsymbol{u}]| - \psi_m) \right\|_{\mathbb{G}_m} \\
&\qquad\qquad + \beta_m \|\psi_m\|_{\widetilde{\mathbb{G}}_m} \; .
\end{aligned} \tag{3.125}$$

Note that this bound is achieved *for any* $\boldsymbol{v}_m \in \Pi_{\mathbb{Q}_m}[\boldsymbol{u}]$, $m = 0, 1, 2, \ldots$

Now, by assumption $E < \delta$. This yields the following bound on the rightmost term of Eq.(3.125):

$$\begin{aligned}
& \textstyle\sum_{m=0}^{M} \frac{\beta_m}{2} \|\psi_m\|_{\widetilde{\mathbb{G}}_m}^2 && < \delta \\
\Longrightarrow \quad & \tfrac{\beta_m}{2} \|\psi_m\|_{\widetilde{\mathbb{G}}_m}^2 && < \delta \\
\Longrightarrow \quad & \|\psi_m\|_{\widetilde{\mathbb{G}}_m} && < \sqrt{\frac{2}{\beta_m}\delta} \\
\Longrightarrow \quad & \textstyle\sum_{m=0}^{M} \beta_m \|\psi_m\|_{\widetilde{\mathbb{G}}_m} && < (2\delta)^{1/2} \sum_{m=0}^{M} \beta_m^{1/2}.
\end{aligned} \tag{3.126}$$

For the remaining terms of Eq.(3.125) consider any $a \in L^2[\mathbb{R}^2, \mathbb{R}_+]$, and $b \in \mathbb{U}_+$ satisfying $\|a - b\|^2 < \delta$. Let

$$\mathbb{G} = \operatorname{supp}(a) \quad \text{and} \quad \mathbb{G}_\epsilon = \left\{ \boldsymbol{x} \;\middle|\; a(\boldsymbol{x}) > \sqrt{\epsilon} \right\}.$$

The remaining norms in Eq.(3.125) take the form

$$\left\|\left(\frac{a^2}{(a^2+\ \epsilon^2)^{1/2}}-b\right)\frac{a^3+2a\epsilon^2}{(a^2+\ \epsilon^2)^{3/2}}+(b-a)\right\|_{\mathbb{G}} \\ \leq \left\|\frac{a\epsilon^4}{(a^2+\epsilon^2)^2}\right\| \quad + \quad \left\|\left(1-\frac{a^3+2a\epsilon^2}{(a^2+\ \epsilon^2)^{3/2}}\right)b\right\|_{\mathbb{G}}. \tag{3.127}$$

Consider the first norm on the right hand side of Eq.(3.127):

$$\left\|\frac{a\epsilon^4}{(a^2+\epsilon^2)^2}\right\| \leq \left\|\frac{a\epsilon^4}{(a^2+\epsilon^2)^2}\right\|_{\mathbb{B}(\frac{1}{\sqrt{\epsilon}})} + \left\|\frac{a\epsilon^4}{(a^2+\epsilon^2)^2}\right\|_{\widetilde{\mathbb{B}}(\frac{1}{\sqrt{\epsilon}})}$$

where $\mathbb{B}(\frac{1}{\sqrt{\epsilon}})$ is the ball of radius $\frac{1}{\sqrt{\epsilon}}$. The argument of the norm over the interior of $\mathbb{B}(\frac{1}{\sqrt{\epsilon}})$ is bounded by $\frac{a\epsilon^4}{(a^2+\epsilon^2)^2} \leq \epsilon$ thus

$$\left\|\frac{a\epsilon^4}{(a^2+\epsilon^2)^2}\right\|_{\mathbb{B}(\frac{1}{\sqrt{\epsilon}})} \leq \sqrt{\pi\epsilon}$$

The norm over the complement $\widetilde{\mathbb{B}}(\frac{1}{\sqrt{\epsilon}})$ cannot be bounded by ϵ without an additional assumption that a has compact support. However since $a \in L^2$ the norm can be made arbitrarily small, *i.e* given ϵ' there is an $\epsilon_0 > 0$ such that

$$\left\|\frac{a\epsilon^4}{(a^2+\epsilon^2)^2}\right\|_{\widetilde{\mathbb{B}}(\frac{1}{\sqrt{\epsilon}})} \leq \epsilon' \quad \forall \epsilon \geq \epsilon_0.$$

Thus

$$\left\|\frac{a\epsilon^4}{(a^2+\epsilon^2)^2}\right\| \leq \sqrt{\pi\epsilon}+\epsilon' \quad \forall \epsilon \geq \epsilon_0. \tag{3.128}$$

Next consider the rightmost norm of Eq.(3.127). Rearranging terms yields

$$\frac{a^3+2a\epsilon^2}{(a^2+\ \epsilon^2)^{3/2}} = \frac{a}{(a^2+\epsilon^2)^{1/2}}\left(1+\frac{\epsilon^2}{a^2+\epsilon^2}\right).$$

From this it is clear that for all a and ϵ

$$0 \leq \frac{a^2}{a^2+\epsilon^2} \leq \frac{a^3+2a\epsilon^2}{(a^2+\ \epsilon^2)^{3/2}} \leq \left(1+\frac{\epsilon^2}{a^2+\epsilon^2}\right)^2 \leq 2.$$

Define

$$g(\alpha,\epsilon) = \left|1-\frac{\alpha^3+2\alpha\epsilon^2}{(\alpha^2+\ \epsilon^2)^{3/2}}\right|.$$

For all (α, ϵ) we have $0 \leq g(\alpha, \epsilon) \leq 1$. Indeed, for all (α, ϵ)

$$\begin{aligned} g(\alpha, \epsilon) &\leq \max\left[1 - \frac{\alpha^2}{\alpha^2 + \epsilon^2},\ \left(1 + \frac{\epsilon^2}{a^2 + \epsilon^2}\right)^2 - 1\right] \\ &= \max\left[\frac{\epsilon^2}{\alpha^2 + \epsilon^2},\ \frac{\epsilon^2}{\alpha^2 + \epsilon^2}\left(\frac{2\alpha^2 + 3\epsilon^2}{a^2 + \epsilon^2}\right)\right] \\ &\leq 5\frac{\epsilon^2}{\alpha^2 + \epsilon^2}. \end{aligned}$$

On the interval $\alpha \in [\sqrt{\epsilon}, \infty)$ we have $\frac{\epsilon^2}{\alpha^2+\epsilon^2} \leq \frac{\epsilon^2}{\epsilon+\epsilon^2} \leq \epsilon$ and

$$g(\alpha, \epsilon) \leq 5\epsilon \quad \forall\ \alpha \in [\sqrt{\epsilon}, \infty).$$

Thus, given $\epsilon' > 0$, there is an $\epsilon > 0$ such that

$$\left\|\left(1 - \frac{a^3 + 2a\epsilon^2}{(a^2 + \ \epsilon^2)^{3/2}}\right) b\right\|_{\mathbb{G}_\epsilon} \leq 5\epsilon\|b\| \leq \epsilon'. \tag{3.129}$$

On $\widetilde{\mathbb{G}}_\epsilon \cap \mathbb{G}$ from the above we have that

$$\left\|\left(1 - \frac{a^3 + 2a\epsilon^2}{(a^2 + \ \epsilon^2)^{3/2}}\right) b\right\|_{\widetilde{\mathbb{G}}_\epsilon \cap \mathbb{G}} \leq \|b\|_{\widetilde{\mathbb{G}}_\epsilon \cap \mathbb{G}}.$$

Since $\|a - b\|^2 < \delta$,

$$\|b\|_{\widetilde{\mathbb{G}}_\epsilon \cap \mathbb{G}} < \|a\|_{\widetilde{\mathbb{G}}_\epsilon \cap \mathbb{G}} + \delta^{1/2}.$$

The norm on the right converges pointwise to zero since, for fixed a,

$$\lim_{\epsilon \to 0} \|a\|_{\widetilde{\mathbb{G}}_\epsilon \cap \mathbb{G}} = \lim_{\epsilon \to 0} \|a\mathcal{X}_{\widetilde{\mathbb{G}}_\epsilon \cap \mathbb{G}}\| = 0.$$

Thus we can apply Lebesgue's Dominated Convergence Theorem to guarantee the existence of an $\epsilon' > 0$ such that for all $\epsilon \in [0, \epsilon']$

$$\|b\|_{\widetilde{\mathbb{G}}_\epsilon \cap \mathbb{G}} < \delta^{1/2}. \tag{3.130}$$

Without applying additional constraints on a the bound of Eq.(3.130) cannot be made tighter.

Letting $\epsilon' = \delta^{1/2}$ in Eq.(3.128)-(3.129) and substituting the bounds Eq.(3.126)-(3.130) into Eq.(3.125) completes the proof. □

Suppose $E[\boldsymbol{u}] < \delta$, then from Eq.(3.123) we have

$$\|\nabla E_\epsilon[\boldsymbol{u}]\|^2 - 2\langle \nabla E_\epsilon[\boldsymbol{u}],\ \boldsymbol{v}\rangle + \|\boldsymbol{v}\|^2 \leq C^2\delta$$

for every $\boldsymbol{v} \in \mathcal{G}[\boldsymbol{u}]$. Therefore, if $\|\nabla E_\epsilon[\boldsymbol{u}]\|^2 + \|\boldsymbol{v}\|^2 > C^2\delta$, then the direction $-\boldsymbol{v}$ is necessarily a direction of descent for $E_\epsilon[\boldsymbol{u}]$ for every $\boldsymbol{v} \in \mathcal{G}[\boldsymbol{u}]$. In particular, if a line search algorithm

$$\boldsymbol{u}^{(\nu+1)} = \boldsymbol{u}^{(\nu)} - \lambda^{(\nu)} \nabla E_\epsilon[\boldsymbol{u}^{(\nu)}]$$

produces a sequence with $E_\epsilon(\boldsymbol{u}^{(\nu)}) \to 0$, then the corresponding projection algorithm

$$\boldsymbol{u}^{(\nu+1)} \in \left(\mathcal{I} - \lambda^{(\nu)} \mathcal{G}^{(\nu)}\right) [\boldsymbol{u}^{(\nu)}]$$

behaves similarly. That is, the qualitative convergence behavior of the projection algorithm can be studied by examining the convergence properties of the corresponding line search algorithm for the perturbed objective. However, in the presence of noise, where the global solution to Pr.(3.33) is greater than zero, the behavior of the algorithms near the solution could differ significantly since the bound Eq.(3.123) does not guarantee that $\text{dist}\,(\nabla E_\epsilon[\boldsymbol{u}], \mathcal{G}[\boldsymbol{u}]) \to 0$.

The principle obstacle to a bound of the form Eq.(3.123) that depends only on ϵ and not on the value of $E[\boldsymbol{u}]$ is the possibility that the estimate $\boldsymbol{u}$ has a domain of positive measure over which $\boldsymbol{u}$ is near zero but the data is non-zero. This problem is consistent with the fact that ∇E_ϵ is a smooth approximation of the multi-valued projection operator. In the numerical literature for wavefront reconstruction, this difficulty is often circumvented by either implicitly or explicitly assuming that none of the estimates $\boldsymbol{u}$ have this property. If one is willing to make this assumption, then a bound of the form Eq.(3.123) that depends only on ϵ is possible.

We begin by showing in the case of one diversity image just how tight the bound is on the distance between the perturbed objective and the projection without imposing any restrictions on the estimate $\boldsymbol{u}$. The bound Eq.(3.131) is unique to the case of one diversity image. While the bound is nonzero in general, the key point is that this bound is independent of the squared set distance error.

Property 3.3.10 *Let $b \in \mathbb{U}_+$ defined in Hyp.2.2.1. Consider $h[\cdot; b, \epsilon]$ defined by Eq.(3.90). Given any $\boldsymbol{u} \in L^2[\mathbb{R}^2, \mathbb{R}^2]$,*

$$\begin{aligned} \inf_{\epsilon \to 0} \mathit{dist}\,(\nabla h[\boldsymbol{u}; b, \epsilon],\ \partial h[\boldsymbol{u}; b]) &= \inf_{\epsilon \to 0} \mathit{dist}\,\left(\nabla h[\boldsymbol{u}; b, \epsilon],\ cl^*\left(-\Pi_{\mathbb{Q}[b]}[\boldsymbol{u}]\right)\right) \\ &= \|b\|_{bdy\,\mathbb{G}} \end{aligned} \qquad (3.131)$$

where $\mathbb{G}$ is given by

$$\mathbb{G} \equiv \mathit{supp}\,(\boldsymbol{u}).$$

PROOF: This is a consequence of Properties 3.3.9 and 3.2.29. On $\mathbb{G}^\circ$, the complement of $\mathbb{G}$, for all $\boldsymbol{v}_* \in \partial h[\boldsymbol{u}; b]$

$$0 \leq \|v_*\|_{\mathbb{G}^\circ}$$

Denote the zero selection from $\text{cl}^*(-\Pi_{\mathbb{Q}[b]})$ by

$$\overline{\boldsymbol{v}}_*(\boldsymbol{x}) \equiv \begin{cases} \frac{\boldsymbol{u}(\boldsymbol{x})}{|\boldsymbol{u}(\boldsymbol{x})|} & \text{for } \boldsymbol{u}(\boldsymbol{x}) \neq 0 \\ 0 & \text{for } \boldsymbol{u}(\boldsymbol{x}) = 0. \end{cases} \qquad (3.132)$$

This selection satisfies

$$\|\overline{\boldsymbol{v}}_*\|_{\mathbb{G}^\circ} \equiv 0$$

And so the selection $\overline{\boldsymbol{v}}_*$ minimizes the inequality corresponding to Eq.(3.126). The remaining inequalities, namely Eq.(3.129) and Eq.(3.130) are independent of the choice of the selection, thus $\overline{\boldsymbol{v}}_*$ is the closest element of $\partial h[\boldsymbol{u}; b]$ to $\nabla h[\boldsymbol{u}; b, \epsilon]$. Inequality (3.129) can be made arbitrarily small on $\mathbb{G}_\epsilon = \{\boldsymbol{x} \mid |\boldsymbol{u}(\boldsymbol{x})| > \sqrt{\epsilon}\}$. The only obstacle to tighter bounds is Eq.(3.130) which yields, in this case,

$$\inf_{\epsilon \to 0} \|\nabla h[\boldsymbol{u}; b, \epsilon] \; - \; \overline{\boldsymbol{v}}_*\| \; = \|b\|_{\mathrm{bdy}\,\mathbb{G}} \,. \tag{3.133}$$

□

The bound can be interpreted as the *measure* of the boundary of the support of $\boldsymbol{u}$ for the measure P corresponding to the density $b(\boldsymbol{x})$. Thus, for the perturbation to agree with the subdifferential, the boundary of the support of $\boldsymbol{u}$ must have zero measure in the space $(\mathbb{R}^2, \mathcal{M}, P)$.

The correspondence between the direction of steepest descent for line search algorithms applied to Pr.(3.33) and the directions toward the projections is often taken for granted in the phase retrieval literature because it is implicitly or explicitly assumed that the iterates belong to the set $\mathbb{W}$ defined below:

$$\mathbb{W} \equiv \bigcap_{m=0}^{M} \mathbb{V}_m \tag{3.134}$$

where

$$\mathbb{V}_m \equiv \{\boldsymbol{v} \mid |\mathcal{F}_m[\boldsymbol{v}]| \neq 0 \text{ a.e. on } \operatorname{supp}(\psi_m)\} \,. \tag{3.135}$$

The next corollary establishes the fact that for every $\boldsymbol{u} \in \mathbb{W}$ the projection operator is single-valued and the gradient ∇E_ϵ converges pointwise to the operator $\mathcal{G}$.

Corollary 3.3.11 *Let the hypotheses of Thm. 3.3.9 hold and let* $\boldsymbol{u} \in \mathbb{V} \neq \emptyset$, *then* $\mathcal{G}[\boldsymbol{u}]$ *is single-valued. Suppose further that for each* $m = 0, 1, 2, \ldots$ *there is a* $\tilde{\psi}_m \in L^2[\mathbb{R}^2, \mathbb{R}_+]$ *such that*

$$\tilde{\psi}_m = \frac{\psi_m}{|\mathcal{F}_m[\boldsymbol{u}]|} \quad \textit{a.e.}$$

Then given any $\delta > 0$ *there exists an* $\epsilon > 0$ *such that*

$$\|\nabla E_\epsilon[\boldsymbol{u}] - \mathcal{G}[\boldsymbol{u}]\; \| \leq \delta.$$

PROOF: The single-valuedness $\mathcal{G}[\boldsymbol{u}]$ follows directly from the definition of the projections. To prove the next statement of the corollary, note that the only terms on the right-hand side of Eq.(3.125) that could not be made arbitrarily small were the terms with bounds Eq.(3.126) and Eq.(3.130). With the assumptions of the corollary these bounds are much tighter. Indeed, since the support of ψ_m is contained in the support of $\mathcal{F}_m[\boldsymbol{u}]$ the bound in Eq.(3.126) is zero since

$$\|\psi_m\|_{\tilde{\mathbb{G}}_m} = 0.$$

For the bound Eq.(3.130), define

$$\mathbb{G}_{m,\epsilon} = \left\{ \boldsymbol{\xi} \,\middle|\, |\mathcal{F}_m[\boldsymbol{u}]|(\boldsymbol{\xi}) > \sqrt{\epsilon} \right\}.$$

As usual denote the complement of this set by $\widetilde{\mathbb{G}}_{m,\epsilon}$. Since there exists a $\tilde{\psi}_m \in L^2[\mathbb{R}^2, \mathbb{R}_+]$ such that $\tilde{\psi}_m = \frac{\psi_m}{|\mathcal{F}_m[\boldsymbol{u}]|}$ a.e. then

$$\begin{aligned} \|\psi_m\|_{\widetilde{\mathbb{G}}_{m,\epsilon} \cap \mathbb{G}_m} &= \left\| \tilde{\psi}_m |\mathcal{F}_m[\boldsymbol{u}]| \right\|_{\widetilde{\mathbb{G}}_{m,\epsilon} \cap \mathbb{G}_m} \\ &\leq \left\| \tilde{\psi}_m \right\|_{\widetilde{\mathbb{G}}_{m,\epsilon} \cap \mathbb{G}_m} \|\mathcal{F}_m[\boldsymbol{u}]\|_{\widetilde{\mathbb{G}}_{m,\epsilon} \cap \mathbb{G}_m} \end{aligned}$$

As in the proof of the bound Eq.(3.130), we have that

$$\lim_{\epsilon \to 0} \|\mathcal{F}_m[\boldsymbol{u}]\|_{\widetilde{\mathbb{G}}_{m,\epsilon} \cap \mathbb{G}_m} = \lim_{\epsilon \to 0} \|\mathcal{F}_m[\boldsymbol{u}] \mathcal{X}_{\widetilde{\mathbb{G}}_{m,\epsilon} \cap \mathbb{G}_m}\| = 0.$$

Thus by Lebesgue's Dominated Convergence Theorem, given any $\delta > 0$ there exists an ϵ such that

$$\|\psi_m\|_{\widetilde{\mathbb{G}}_{m,\epsilon} \cap \mathbb{G}_m} < \delta.$$

□

For $m \geq 1$ the assumptions of Corollary 3.3.11 are extremely strong. While each of the sets $\mathbb{W}_m$ is dense in $L^2[\mathbb{R}^2, \mathbb{R}^2]$, this is not true for the intersection. Indeed, it is common that $\mathbb{W} = \emptyset$, as in the case of noisy data. Supposing $\mathbb{W} \neq \emptyset$, for $\boldsymbol{u} \in \mathbb{W}$ we define the "gradient" of the unperturbed set distance error by

$$\nabla E[\boldsymbol{u}] \equiv \lim_{\epsilon \to 0} \nabla E_\epsilon[\boldsymbol{u}].$$

Together with Lebesgue's Dominated Convergence Theorem [99, pg.133], the above corollary implies that for $\boldsymbol{u} \in \mathbb{W} \neq \emptyset$

$$\nabla E[\boldsymbol{u}] = \mathcal{G}[\boldsymbol{u}] \quad a.e.$$

Note that $\nabla E[\boldsymbol{u}]$ is not the gradient in the Fréchet sense. In Ref. [15] the authors impose assumptions that allow them to prove that this object *is* the gradient in the Fréchet sense. We noted above, however, that in most practical situations $\mathbb{W} = \emptyset$, thus the applicability of any such assumptions is negligible. Applying this theory to algorithms is also problematic. Supposing that $\mathbb{W} \neq \emptyset$, then one must find an initial point $\boldsymbol{u}_0 \in \mathbb{W}$. Once an initial admissible point is found, one must guarantee that all subsequent iterates remain in $\mathbb{W}$ as well. Algorithms that do not take this into account suffer from numerical instabilities. This issue is discussed in Chapter 6.

Chapter 4

PHASE DIVERSITY

The analysis of the wavefront reconstruction problem in the previous chapter is essential for understanding the more general problem of *simultaneous* wavefront reconstruction and deconvolution. In this problem the object φ is not a delta function, but rather an extended source

$$\boldsymbol{\mathcal{K}}[\boldsymbol{u}]\varphi = \begin{pmatrix} \kappa[\mathcal{F}_0[\boldsymbol{u}]]^2 * \varphi \\ \kappa[\mathcal{F}_1[\boldsymbol{u}]]^2 * \varphi \\ \vdots \\ \kappa[\mathcal{F}_M[\boldsymbol{u}]]^2 * \varphi \end{pmatrix} = \begin{pmatrix} \psi_0^2 \\ \vdots \\ \psi_M^2 \end{pmatrix} \tag{4.1}$$

where $\kappa[\cdot] = |\cdot|$, the pointwise modulus. This problem is clearly ill-posed. In general, for any linear equation where the linear operator as well as the input are unknown, infinitely many solutions are possible. Indeed, suppose the pair $(\mathcal{L}, \varphi)$ satisfies

$$\mathcal{L}\varphi = \psi^2. \tag{4.2}$$

Then for any pair $(\mathcal{L}_*, \varphi_*)$ satisfying

$$\mathcal{L}_*\varphi_* = -(\mathcal{L}_*\varphi + \mathcal{L}\varphi_*)$$

the pair $(\mathcal{L}_* + \mathcal{L}, \varphi + \varphi_*)$ also satisfies Eq.(4.2). In practice, computational wavefront reconstruction/deconvolution algorithms are only intended for systems that are very nearly identified [116,146], i.e. the operator $\boldsymbol{\mathcal{K}}$ is known to within a small error. In this context, the wavefront reconstruction/deconvolution problem can be viewed as one of finding an optimal filter for recovering the object φ from the image data $\boldsymbol{\psi}^{\cdot 2}$ where $\cdot 2$ indicates element-wise exponentiation of the vector $\boldsymbol{\psi}$.

For fixed $\boldsymbol{u}$, Eq.(2.44) is a system of Fredholm integral equations of the *first* kind. In physical terms, the convolution operator $\mathcal{K}_m$ smoothes the object φ, *i.e.* high frequency components, corners and edges are smoothed by integration. For example, let $\varphi = \mathcal{X}_{[0,1]} \sin(2\pi\boldsymbol{\xi} \cdot \boldsymbol{x})$. The corresponding image is given by

$$\int_{\mathbb{R}^2} \kappa^2(\boldsymbol{x} - \boldsymbol{y})\mathcal{X}_{[0,1]} \sin(2\pi\boldsymbol{\xi} \cdot \boldsymbol{y})d\boldsymbol{y} = \psi^2(\boldsymbol{x}).$$

For each $\boldsymbol{x}$ the Riemann-Lebesgue Lemma states that $\psi(\boldsymbol{x}) \to 0$ as $|\boldsymbol{\xi}| \to \infty$. The reverse process of computing φ from the image ψ^2 can therefore be expected to *amplify* high frequency components of φ. In particular, naive inversion of the convolution operator amplifies noise in the object reconstruction. All practical methods for image processing must provide for the separation of noise from the object, or in other words *filtering* the image. Ill-conditioning and the compactness of the integral operator $\mathcal{K}$ are closely related. In most applications, the convolution operator $\mathcal{K}_m$ defined by Eq.(2.46) is compact, or practically compact in the sense that its singular values decay to zero, thus inverting the normal equations for solving the least squares problem formulated below is numerically unstable.

Another source of ill-conditioning in least squares solutions to Eq.(4.1) is the form of the performance measure itself. The problem of simultaneous wavefront reconstruction and deconvolution does not admit an easy formulation in terms of κ rather than κ^2. For the phase retrieval problem the objective corresponding to the squared set distance error Eq.(3.31) is

$$E_2[\boldsymbol{u}] \equiv \sum_{m=0}^{M} \frac{\beta_m}{2} \left\| \kappa \left[\mathcal{F}_m[\boldsymbol{u}]\right]^2 - \psi_m^2 \, \right\|^2 \tag{4.3}$$

At first sight, this would seem to be an advantage since this objective is trivially Fréchet differentiable. Indeed, $\kappa[\boldsymbol{u}]^2 = |\boldsymbol{u}|^2$ has Fréchet derivative

$$[\kappa[\boldsymbol{u}]^2]'(\boldsymbol{w}) = 2(\boldsymbol{u}(\cdot), \boldsymbol{w}(\cdot)).$$

We show in Chapter 6, however, that performance of algorithms with the nonsmooth objective is far superior to performance with the smooth objective. An intuitive explanation, for this is that the modulus squared smoothes out curvature information in the objective [96, 109]. This is depicted in Figure (4.1). In addition, the singular values of the gradient of the squared-modulus kernel are more spread out. This can be seen by studying the linearized problem

$$(\kappa[\mathcal{F}_m[\boldsymbol{u}]]^2)'(\boldsymbol{w}(\cdot)) = 2(\mathcal{F}_m[\boldsymbol{u}](\cdot), \boldsymbol{w}(\cdot)) = -(\kappa[\mathcal{F}_m[\boldsymbol{u}]]^2(\cdot) - \psi_m^2(\cdot)). \tag{4.4}$$

For the phase retrieval problem the values of $\mathcal{F}_m[\boldsymbol{u}]$ vary continuously by several orders of magnitude. The linearized problem is thus highly ill-conditioned. This problem is at the heart of higher-order methods for solving the phase retrieval problem. In contrast, it was shown in the previous chapter that selections $\boldsymbol{v} \in \partial(-\kappa[\boldsymbol{u}])$ have pointwise unit magnitude. The linearized equation

$$(\boldsymbol{v}(\cdot), \boldsymbol{w}(\cdot)) = -(\kappa[\boldsymbol{u}] - b). \tag{4.5}$$

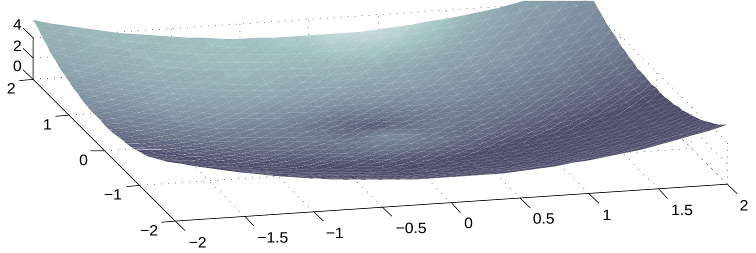
is perfectly conditioned. We expect, therefore, methods based on smooth approximations to κ such as those studied in Section 3.3 of Chapter 3 to perform better than those based on κ^2. This agrees with our observations, detailed in Chapter 6 as well as those of other researchers for similar type problems [11, 83]. While it is difficult to overcome the problem of the inefficient objective, ill-conditioning and ill-posedness is readily addressed by regularization strategies.

4.1 Least Squares Regularization

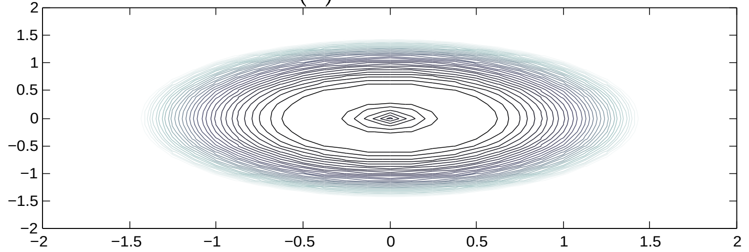
Consider the least squares performance measure for the system of operator equations given by Eq.(2.44):

$$\begin{array}{ll} \text{minimize} & \displaystyle\sum_{m=0}^{M} \frac{\beta_m}{2} \left\| \mathcal{K}_m[\boldsymbol{u}]\varphi - \psi_m^2 \right\|^2 \\ \text{over} & \boldsymbol{u} \in L^2[\mathbb{R}^2, \mathbb{R}^2], \ \ \varphi \in L^2[\mathbb{R}^2, \mathbb{R}] \end{array} \tag{4.6}$$

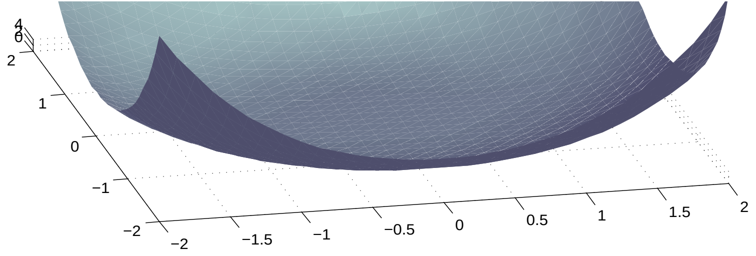
(a) Modulus kernel



(b) Contours of modulus kernel



(c) Modulus squared kernel



(d) Contours of modulus squared kernel

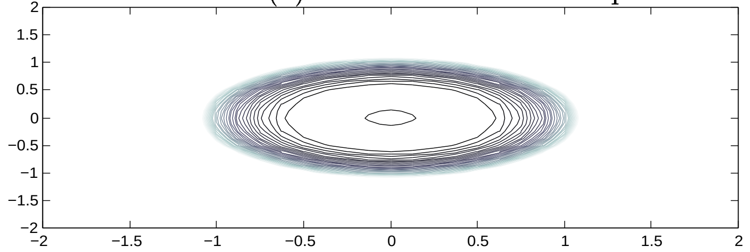


Figure 4.1: Comparison of modulus kernel to modulus-squared kernel. The contours are at the same levels for each figure and show that squared-modulus objective has less structure near zero than the modulus objective

where $\mathcal{K}_m[\boldsymbol{u}]$ is defined by Eq.(2.46). Since the kernel of the integral operator $\mathcal{K}_m[\boldsymbol{u}]$ is Fréchet differentiable it is not necessary to consider any perturbations of the form studied in the previous chapter. Tikhonov's method involves incorporating *a priori* assumptions about the size and smoothness of the solution. This is done simply by adding a penalty term to the objective

$$\begin{array}{ll} \text{minimize} & \left\|\boldsymbol{\mathcal{K}}[\boldsymbol{u}]\varphi - \boldsymbol{\psi}^{\cdot 2}\right\|^2 + \alpha^2\|\mathcal{T}\varphi - \phi\|^2 \\ \text{over} & \boldsymbol{u} \in L^2[\mathbb{R}^2, \mathbb{R}^2],\ \varphi \in L^2[\mathbb{R}^2, \mathbb{R}]. \end{array} \tag{4.7}$$

An alternative formulation is the following

$$\text{minimize} \left\| \begin{pmatrix} \mathcal{K}_0[\boldsymbol{u}] \\ \vdots \\ \mathcal{K}_M[\boldsymbol{u}] \\ \alpha\mathcal{T} \end{pmatrix} \varphi - \begin{pmatrix} \psi_0^2 \\ \vdots \\ \psi_M^2 \\ \phi \end{pmatrix} \right\|^2 \quad \text{over } \boldsymbol{u} \in L^2[\mathbb{R}^2, \mathbb{R}^2],\ \varphi \in L^2[\mathbb{R}^2, \mathbb{R}]. \tag{4.8}$$

The simplest example of regularization operators is $\mathcal{T} = \mathcal{I}$, the identity operator. For invertible $\mathcal{T} \neq \mathcal{I}$ the general regularization problem is easily transformed into a regularization problem with the identity as the regularization term. Let

$$\tilde{\boldsymbol{\mathcal{K}}} = \begin{pmatrix} \mathcal{K}_0\mathcal{T}^{-1} \\ \vdots \\ \mathcal{K}_M\mathcal{T}^{-1} \end{pmatrix}, \quad \tilde{\boldsymbol{\psi}}^{\cdot 2} = \boldsymbol{\psi}^{\cdot 2} \quad \text{and} \quad \tilde{\phi} = \phi.$$

If φ_* is a solution to Pr.(4.7) then $\tilde{\varphi}_* = \mathcal{T}\varphi_*$ is a solution to the following standard-form regularized least squares problem

$$\begin{array}{ll} \text{minimize} & \left\|\tilde{\boldsymbol{\mathcal{K}}}[\boldsymbol{u}]\tilde{\varphi} - \tilde{\boldsymbol{\psi}}^{\cdot 2}\right\|^2 + \alpha^2\|\mathcal{I}\tilde{\varphi} - \tilde{\phi}\|^2 \\ \text{over} & \boldsymbol{u} \in L^2[\mathbb{R}^2, \mathbb{R}^2],\ \tilde{\varphi} \in L^2[\mathbb{R}^2, \mathbb{R}]. \end{array} \tag{4.9}$$

The system of equations (2.44) is linear in φ and nonlinear in $\boldsymbol{u}$. This structure allows one to split the corresponding optimization problem using a Benders decomposition [24]. Benders decompositions are common techniques for splitting large-scale optimization problems such as Eq.(4.8) into smaller problems which can be solved independently of one another in sequence. The least squares performance measure admits a particularly simple way to split the problem. This was first recognized by Gonsalves [75] and later generalized by Paxman *et al* [139]. Benders decomposition involves first obtaining φ_* by optimizing over φ for fixed $\boldsymbol{u}$. Next one solves for the optimal $\boldsymbol{u}$ holding φ_* fixed. The process is repeated until the iterates exceed some tolerance. If $\operatorname{null}(\boldsymbol{\mathcal{K}}[\boldsymbol{u}]) \cap \operatorname{null}(\mathcal{T}) = \{0\}$ for fixed $\boldsymbol{u}$ then the Tikhonov solution for the object φ_* is the unique closed form solution to the optimization problem over φ_*. This is formally given as the solution to the corresponding normal equations

$$\varphi_* = \boldsymbol{\mathcal{K}}_\alpha^\sharp \begin{pmatrix} \boldsymbol{\psi}^{\cdot 2} \\ \phi \end{pmatrix} \quad \text{with} \quad \boldsymbol{\mathcal{K}}_\alpha^\sharp = \left(\boldsymbol{\mathcal{K}}^H\boldsymbol{\mathcal{K}} + \alpha^2\mathcal{T}^*\mathcal{T}\right)^{-1}\left(\boldsymbol{\mathcal{K}}^H, \mathcal{T}^*\right) \tag{4.10}$$

where $\mathcal{K}_\alpha^\sharp$ denotes the Tikhonov regularized inverse and $\mathcal{K}^H = (\mathcal{K}_0^*, \dots, \mathcal{K}_M^*)$ is the transpose of the vector of operators adjoint to $\mathcal{K}_m$.

For convolution operators, there is a very simple and efficient diagonalization process via the Fourier transform, thus one should choose a regularization $\mathcal{T}$ which shares this property. Denote the Fourier transform of the regularization operator $\mathcal{T}$ by T. By Parseval's relation and the fact that the transforms $\mathcal{F}_m$ defined by Eq.(2.43) and Eq.(2.48) are bijective unitary linear operators, the optimal solution of Pr.(4.7) is equivalent to the optimal value of the Fourier dual problem

$$\begin{array}{ll} \text{minimize} & \left\|\boldsymbol{K}[\boldsymbol{u}]\varphi^\wedge - [\boldsymbol{\psi}^{\cdot 2}]^\wedge\right\|^2 + \alpha^2\|T\varphi^\wedge - \phi^\wedge\|^2 \\ \text{over} & \boldsymbol{u} \in L^2[\mathbb{R}^2, \mathbb{R}^2],\ \varphi^\wedge \in L^2[\mathbb{R}^2, \mathbb{R}] \end{array} \tag{4.11}$$

where $\boldsymbol{K}$ is defined by Eq.(2.59). The alternative formulation yields

$$\text{minimize} \left\| \begin{pmatrix} K_0[\boldsymbol{u}] \\ \vdots \\ K_M[\boldsymbol{u}] \\ \alpha T \end{pmatrix} \varphi^\wedge - \begin{pmatrix} [\psi_0^2]^\wedge \\ \vdots \\ [\psi_M^2]^\wedge \\ \phi^\wedge \end{pmatrix} \right\|^2 \quad \text{over } \boldsymbol{u} \in L^2[\mathbb{R}^2, \mathbb{R}^2],\ \varphi^\wedge \in L^2[\mathbb{R}^2, \mathbb{R}]. \tag{4.12}$$

As above, if $\operatorname{null}(\mathcal{K}[\boldsymbol{u}]) \cap \operatorname{null}(\mathcal{T}) = \{0\}$ then $\operatorname{null}(\boldsymbol{K}[\boldsymbol{u}]) \cap \operatorname{null}(T) = \{0\}$ for fixed $\boldsymbol{u}$ and the Tikhonov solution for the transformed object $\varphi_*^\wedge$ is unique. The regularized solution is formally given as the solution to the corresponding normal equations

$$\varphi_*^\wedge = \boldsymbol{K}_\alpha^\sharp \begin{pmatrix} [\boldsymbol{\psi}^{\cdot 2}]^\wedge \\ \phi^\wedge \end{pmatrix} \quad \text{with} \quad \boldsymbol{K}_\alpha^\sharp = (\boldsymbol{K}^H\boldsymbol{K} + \alpha^2 T^*T)^{-1}\left(\boldsymbol{K}^H, T^*\right) \tag{4.13}$$

where $\boldsymbol{K}_\alpha^\sharp$ denotes the Fourier transform of the Tikhonov regularized inverse.

The dimensionality of the original simultaneous wavefront reconstruction and deconvolution problem is reduced by substituting the Tikhonov solution for the optimal object $\varphi_*^\wedge$ directly into the objective in Pr.(4.12). This is a nonlinear optimization problem in $\boldsymbol{u}$ alone.

4.2 A Statistical Perspective: the Wiener filter

The performance measure in the general optimization problem Pr.(2.81) assumes a particular underlying stochastic model which, in turn, depends on the mechanical image formation system. Chapter 2 describes the physics of the *image formation process* of an ideal continuous diffraction limited incoherent imaging system. In this section we consider the probabilistic nature of *image observation* and the correspondence of the optimal regularized least squares estimator with optimal filters for a noisy observation system.

In the discussion of the image formation process in Chapter 2, two steps are conspicuously absent in the electromagnetic wave's journey from its source to our eyes. These are *image observation* and *noise corruption.* The latter of the two processes is not by choice, and much effort is dedicated to eradicating its effects. The image observation model describes the interaction of the *electro-optical* system with the electromagnetic field as well as other "random" factors which are modeled as noise. The electro-optical system consists of imaging arrays and circuitry that translates photons into images on a computer screen that make sense to us and that we can store for future reference. The mathematical analog of this process is the *discretization* of the continuous imaging model of Eq.(4.1). Since the objective is Fréchet differentiable, the discretization does not effect the analytic limits of derivatives calculated in any direction. Thus we are free to choose any convenient discretization, or in this case, we are free to accept the discretization imposed by the image observation process. Numerical considerations of the discretization are discussed in detail in Chapter 5.

The electro-optical observation system consists of arrays of charge-coupled devices (*CCD's*) arranged in an evenly spaced Cartesian grid in the image plane of the ideal optical system depicted in Fig.(2.1). The grid elements in two dimensions are called *pixels*. From Eq.(2.43) the image ψ_m^2 for $m \geq 1$ is the modulus of the Fourier transform of the aberrated wavefront. Thus each of the pixels in the image array is a sample of the continuous Fourier transform, or, alternatively, the coefficient of the discrete Fourier series representation of the aberrated wavefront $P[\boldsymbol{u}] \exp[\sqrt{-1}\ \tilde{\theta}_m]$. We therefore discretize the problem by the Fourier series expansion. As discussed in Chapter 6 the evenly spaced arrays allows us to take advantage of the Fast Fourier Transform (FFT).

4.2.1 The least squares error metric as a log-likelihood estimator

We recycle the notation used for the continuous analysis for the following discrete analysis. Denote the vector of integers $\boldsymbol{n} = (n_1, n_2) \in \mathbb{I} \subset \mathbb{Z} \times \mathbb{Z}$. The inequality $\boldsymbol{n} \leq \boldsymbol{n}'$ indicates $n_i \leq n_i'$ $(i = 1, 2)$ Consider the discrete linear observation model with additive noise

$$\kappa[\mathcal{F}_m[\boldsymbol{u}(\boldsymbol{n})]]^2 * \varphi(\boldsymbol{n}) + \eta_m(\boldsymbol{n}) = \psi_m^2(\boldsymbol{n}) \tag{4.14}$$

where $*$ is the discrete convolution operator and $\mathcal{F}_m$ is the discrete transform corresponding to the transform defined in Eq.(2.43) and Eq.(2.48). The index $\boldsymbol{n}$ corresponds to a pixel on the imaging array. Each pixel is a CCD that essentially counts individual photons within the area element. Any miscount of a photon from the intended source is noise in the system. Miscounts are often caused by heat radiating from the optical device or other nearby objects (*black body radiation*). This is known as *thermal* noise. Thermal noise $\eta(\boldsymbol{n})$ is modeled as independent, identically distributed (iid) zero-mean Gaussian noise with variance σ_m at each pixel $\boldsymbol{n}$ [139]. The image $\psi_m^2(\boldsymbol{n})$ is thus a random variable with normal probability density

$$p[\psi_m^2(\boldsymbol{n}); \mathcal{F}_m[\boldsymbol{u}](\boldsymbol{n}), \varphi,] = \frac{1}{(2\pi\sigma_m^2)^{1/2}} \exp\left[-\frac{\mathcal{E}\left[\kappa[\mathcal{F}_m[\boldsymbol{u}]]^2 * \varphi(\boldsymbol{n}) - \psi_m^2(\boldsymbol{n})\right]^2}{2\sigma_m^2} \right] \tag{4.15}$$

where $\mathcal{E}[\cdot]$ denotes the expectation of a random variable. With the iid assumption, the probability density over the mth image is given by the following product over all pixels $\boldsymbol{n} \in \mathbb{I}$

$$d(\psi_m^2; \mathcal{F}_m[\boldsymbol{u}], \varphi) = \prod_{\boldsymbol{n} \in \mathbb{I}} \frac{1}{(2\pi\sigma_m^2)^{1/2}} \exp\left[-\frac{\mathcal{E}[\kappa[\mathcal{F}_m[\boldsymbol{u}]]^2 * \varphi(\boldsymbol{n}) - \psi_m^2(\boldsymbol{n}) -]^2}{2\sigma_m^2} \right]. \tag{4.16}$$

The distribution over all diversity images ψ_m^2 is the product of the densities of the images

$$d(\boldsymbol{\psi}^{\cdot 2};,\boldsymbol{u},\varphi) \equiv \prod_{m=0}^{M} d(\psi_m^2;\varphi,\mathcal{F}_m[\boldsymbol{u}]) \tag{4.17}$$

The maximum likelihood estimator (MLE) $(\boldsymbol{u}_*,\varphi_*)$ is the estimate that is most likely to have produced the measurement. This is found my minimizing the *log-likelihood function* given by

$$\begin{aligned} L[\boldsymbol{u},\varphi] &\equiv -\ln(d(\boldsymbol{\psi}^{\cdot 2};,\boldsymbol{u},\varphi)) \\ &= \sum_{m=0}^{M}\sum_{\boldsymbol{n}\in\mathbb{I}} \frac{\mathcal{E}\left[\kappa[\mathcal{F}_m[\boldsymbol{u}]]^2 * \varphi(\boldsymbol{n}) - \psi_m^2(\boldsymbol{n})\right]^2}{2\sigma_m^2} + \sum_{m=0}^{M} \frac{\boldsymbol{N}}{2} ln(2\pi\sigma_m^2). \end{aligned} \tag{4.18}$$

where $\boldsymbol{N}$ is the cardinality of $\mathbb{I}$. Dropping the constant $\sum_{m=0}^{M}\frac{\boldsymbol{N}}{2}ln(2\pi\sigma_m^2)$ and normalizing by the number of pixels yields the discrete analog to the continuous least squares performance measure given in Pr.(4.6)

$$\text{minimize} \quad \frac{1}{2M\boldsymbol{N}}\sum_{m=0}^{M}\frac{1}{\sigma_m^2}\mathcal{E}\left[\left\|\mathcal{K}_m[\boldsymbol{u}]\varphi - \psi_m^2\right\|_F^2\right] \tag{4.19}$$

$$\text{over} \qquad \boldsymbol{u}\in\mathbb{R}^{\boldsymbol{N}}\times\mathbb{R}^{\boldsymbol{N}}, \quad \varphi\in\mathbb{R}^{\boldsymbol{N}} \tag{4.20}$$

where $\|\cdot\|_F$ is the Frobenius norm. The formulation above yields the correspondence between the variance of the data and the weights β_m in the weighted squared set distance error E defined by Eq.(3.31):

$$\beta_m = \frac{1}{M\sigma_m^2}.$$

4.2.2 Some statistical definitions

Before we derive the Wiener filter, we define some elementary statistical functions. Throughout this discussion the random variables are assumed to be *stationary.* A random complex-valued sequence $x(\boldsymbol{n})$ is said to be *strict-sense stationary* if the joint density of any partial sequence $\{x(\boldsymbol{l}),\ 1\le\boldsymbol{l}\le\boldsymbol{k}\}$ is the same as that of the shifted sequence $\{x(\boldsymbol{l}+\boldsymbol{m}),\ 1\le\boldsymbol{l}\le\boldsymbol{k}\}$ for any vector of integers $\boldsymbol{m}$ and any length $\boldsymbol{k}$. Denote the complex conjugate of x by $\overline{x}$ and the *autocorrelation matrix*, $\mathcal{A}[\boldsymbol{n},\boldsymbol{n}']$, by

$$\mathcal{A}[\boldsymbol{n},\boldsymbol{n}'] \equiv \mathcal{E}[x(\boldsymbol{n})\overline{x}(\boldsymbol{n}')].$$

The sequence $x(\boldsymbol{n})$ is called *wide-sense* stationary if

$$\mathcal{E}[x(\boldsymbol{n})] = const$$

and

$$\mathcal{A}[\boldsymbol{n},\boldsymbol{n}'] = q(\boldsymbol{n}-\boldsymbol{n}')$$

where q is some general function of the vector of integers $\boldsymbol{n}$. For Gaussian sequences wide-sense stationarity and strict-sense stationarity are the same and we will simply state that the sequence is stationary. Define the covariance function of the stationary random sequence $x(\boldsymbol{n})$ with mean μ by

$$\mathcal{C}_x(\boldsymbol{n}) \equiv \mathcal{E}[(x(\boldsymbol{n})-\mu)(\overline{x}(0)-\overline{\mu})] = \mathcal{E}[(x(\boldsymbol{n}+\boldsymbol{n}')-\mu)(\overline{x}(\boldsymbol{n}')-\overline{\mu})] \quad \forall\, \boldsymbol{n},\ \boldsymbol{n}'. \tag{4.21}$$

Similarly, the autocorrelation function, denoted $\mathcal{A}(\boldsymbol{n})$ is defined by

$$\mathcal{A}_x(\boldsymbol{n}) \equiv \mathcal{E}[x(\boldsymbol{n})\overline{x}(0)] = \mathcal{E}[x(\boldsymbol{n}+\boldsymbol{n}')\overline{x}(\boldsymbol{n}')] \quad \forall\, \boldsymbol{n},\ \boldsymbol{n}'. \tag{4.22}$$

Using the definitions of covariance and autocorrelation functions it can be shown that the covariance and autocorrelation matrices are Hermitian and positive semidefinite. The *cross-correlation* of two jointly stationary random sequences $x(\boldsymbol{n})$ and $y(\boldsymbol{n}')$ is defined by

$$\mathcal{C}_{xy}(\boldsymbol{n}-\boldsymbol{n}') \equiv \mathcal{E}[x(\boldsymbol{n})\overline{y}(\boldsymbol{n}')]. \tag{4.23}$$

The *spectral density function*, $\mathcal{S}$, is defined as the Fourier transform of the covariance function

$$\mathcal{S}_x(\boldsymbol{\omega}) \equiv [\mathcal{C}_x]^{\wedge}(\boldsymbol{\omega}). \tag{4.24}$$

The cross-spectral density function is the Fourier transform of the cross-correlation function

$$\mathcal{S}_{xy}(\boldsymbol{\omega}) \equiv [\mathcal{C}_{xy}]^{\wedge}(\boldsymbol{\omega}).$$

4.2.3 The Wiener filter

We are now ready to derive the Wiener filter. For stationary Gaussian model, the conditional mean of φ given $\boldsymbol{\psi}^{\cdot 2}$ for fixed $\boldsymbol{u}$ is the best linear estimate of the form

$$\varphi_*(\boldsymbol{n}) = \boldsymbol{\mathcal{W}} * \boldsymbol{\psi}^{\cdot 2}(\boldsymbol{n}) \tag{4.25}$$

where the *filter impulse response* $\boldsymbol{\mathcal{W}}(\boldsymbol{n})$ is determined to minimize the mean square error of Eq.(4.19). The solution to Pr.(4.19), φ_*, satisfies the orthogonality condition

$$\mathcal{E}[(\varphi(\boldsymbol{n})-\varphi_*(\boldsymbol{n}))\boldsymbol{\psi}^{\cdot 2}(\boldsymbol{n}')] = 0 \quad \forall\, \boldsymbol{n}, \boldsymbol{n}'. \tag{4.26}$$

Equations (4.25)-(4.26) yield

$$\begin{aligned}
\mathcal{E}\left[\varphi(\boldsymbol{n})\boldsymbol{\psi}^{\cdot 2}(\boldsymbol{n}')\right] &= \mathcal{E}[\varphi_*(\boldsymbol{n})\boldsymbol{\psi}^{\cdot 2}(\boldsymbol{n}')] \\
&= \mathcal{E}\left[\boldsymbol{\mathcal{W}} * \boldsymbol{\psi}^{\cdot 2}(\boldsymbol{n})\boldsymbol{\psi}^{\cdot 2}(\boldsymbol{n}')\right] \\
&= \boldsymbol{\mathcal{W}} * \mathcal{E}\left[\boldsymbol{\psi}^{\cdot 2}(\boldsymbol{n})\boldsymbol{\psi}^{\cdot 2}(\boldsymbol{n}')\right] \\
&= \boldsymbol{\mathcal{W}} * \mathcal{C}_{\boldsymbol{\psi}^{\cdot 2}\boldsymbol{\psi}^{\cdot 2}}(\boldsymbol{n}-\boldsymbol{n}').
\end{aligned}$$

By the definition of cross-correlation Eq.(4.23), this can be written simply as

$$\mathcal{C}_{\varphi\boldsymbol{\psi}^{\cdot 2}}(\boldsymbol{n}-\boldsymbol{n}') = \boldsymbol{\mathcal{W}} * \mathcal{C}_{\boldsymbol{\psi}^{\cdot 2}\boldsymbol{\psi}^{\cdot 2}}(\boldsymbol{n}-\boldsymbol{n}'). \tag{4.27}$$

where $\mathcal{C}_{\varphi\boldsymbol{\psi}^{\cdot 2}}(\boldsymbol{n}) \equiv \left(\mathcal{C}_{\varphi\psi_1^2}(\boldsymbol{n}), \ldots, \mathcal{C}_{\varphi\psi_M^2}(\boldsymbol{n})\right)$ denotes the cross-correlation between the diversity images $\boldsymbol{\psi}^{\cdot 2}$ and the object φ, and $\mathcal{C}_{\boldsymbol{\psi}^{\cdot 2}\boldsymbol{\psi}^{\cdot 2}}(\boldsymbol{n})$ denotes the autocorrelation over all the diversity images. Together Eq.(4.25) and Eq.(4.27) are called the Wiener filter equations. Taking the Fourier transform of the far left and right sides of Eq.(4.27) yields

$$\mathcal{S}_{\varphi\boldsymbol{\psi}^{\cdot 2}}(\boldsymbol{\omega}) = \boldsymbol{W}(\boldsymbol{\omega})\mathcal{S}_{\boldsymbol{\psi}^{\cdot 2}\boldsymbol{\psi}^{\cdot 2}}(\boldsymbol{\omega})$$

where $\boldsymbol{W}$ is the Fourier transform of $\boldsymbol{\mathcal{W}}$. Thus, $\boldsymbol{W}$ is formally given by

$$\boldsymbol{W}(\boldsymbol{\omega}) = \mathcal{S}_{\varphi\boldsymbol{\psi}^{\cdot 2}}(\boldsymbol{\omega})\left[\mathcal{S}_{\boldsymbol{\psi}^{\cdot 2}\boldsymbol{\psi}^{\cdot 2}}(\boldsymbol{\omega})\right]^{-1}. \tag{4.28}$$

For the image formation model given by Eq.(4.14), assuming stationary additive noise η_m, uncorrelated with φ we have

$$\mathcal{S}_{\boldsymbol{\psi}^{\cdot 2}\boldsymbol{\psi}^{\cdot 2}}(\boldsymbol{\omega}) = \boldsymbol{K}^H\boldsymbol{K}\mathcal{S}_{\varphi\varphi}(\boldsymbol{\omega}) + \mathcal{S}_{\eta\eta}(\boldsymbol{\omega}) \quad \text{and} \quad \mathcal{S}_{\varphi\boldsymbol{\psi}^{\cdot 2}}(\boldsymbol{\omega}) = \boldsymbol{K}^H\mathcal{S}_{\varphi\varphi}(\boldsymbol{\omega}) \tag{4.29}$$

where $\boldsymbol{K}$ is the discrete Hadamard multiplication operator corresponding to $\boldsymbol{K}$ defined by Eq.(2.59). Thus we arrive at the Fourier-Wiener filter for the imaging model Eq.(4.14)

$$\boldsymbol{W}(\boldsymbol{\omega}) = \left[\boldsymbol{K}^H\boldsymbol{K}\mathcal{S}_{\varphi\varphi}(\boldsymbol{\omega}) + \mathcal{S}_{\eta\eta}(\boldsymbol{\omega})\right]^{-1}\boldsymbol{K}^H\mathcal{S}_{\varphi\varphi}(\boldsymbol{\omega}) \tag{4.30}$$

where $\boldsymbol{W}(\boldsymbol{\omega}) = (W_1, \ldots, W_M)$ and

$$W_m = \left[\boldsymbol{K}^H\boldsymbol{K}\mathcal{S}_{\varphi\varphi}(\boldsymbol{\omega}) + \mathcal{S}_{\eta\eta}(\boldsymbol{\omega})\right]^{-1}\overline{K}_m^H\mathcal{S}_{\varphi\varphi}(\boldsymbol{\omega}).$$

Comparing Eq.(4.30) with Eq.(4.13) reveals the correspondence of Tikhonov regularization to construction of the optimal Wiener filter for a linear imaging model with stationary Gaussian noise. The correspondence provides a method for choosing the regularization parameter α to satisfy maximum-likelihood criteria [5, 6]. If the noise power $\mathcal{S}_{\eta\eta}$ goes to zero, then the Wiener filter becomes the pseudoinverse of the operator $\boldsymbol{K}$. Since the convolution operator $\boldsymbol{\mathcal{K}}$ tends to smooth the object φ this is called the *blur* of the optical system. In the absence of blur, that is, when $\boldsymbol{K} = \boldsymbol{I}$ the identity, the Wiener filter is the optimal noise smoothing filter.

4.3 Regularization via Parameterization

Any numerical method will implicitly or explicitly involve a parameterization or, equivalently, discretization. Since the functionals discussed in the previous chapters are Fréchet differentiable, the discretization does not effect the analytic limits of the derivatives calculated in any direction. In Section 4.2 the physics at the *image plane* lead us to favor a pixel parameterization. We show in Section 5.3 of Chapter 5 that this parameterization lends itself easily to a multi-resolution analysis which can be exploited for numerical purposes. In many applications, however, it is more common for the physics at the *pupil plane* to determine the numerical discretization. We discuss these considerations below.

Until recently, optical design has allowed for very efficient discretizations of the apertures in terms of Zernike polynomials [26, 119, 175, 204]. The polynomials, first derived by their namesake, Friz Zernike [204], are a complete basis in L^2 whose terms are orthogonal on the circle [25, 26]. These polynomials were extended by Mahajan [119] to the annulus. More recently Swantner [175] has proposed a Gram-Schmidt procedure for constructing complete bases of polynomials orthogonal on more general regions. The Zernike polynomials enjoy an additional property that the primary optical aberrations such as piston, tilt, focus, astigmatism, coma, clover and spherical aberration, are easily represented by linear combinations of Zernike polynomials of degree 4 or less [26, pp.525-530]. Thus a wide variety of aberrations at the pupil can be represented easily by just a few basis functions.

Modern optical devices, however, no longer have apertures with simple geometric configurations (see Fig.(6.1a)). Zernike polynomials are not orthogonal on segmented pupils and non-circular apertures and they do not represent the optical aberrations common to complicated apertures as efficiently as they do for circular apertures. Even if only several hundred Zernikes are needed to characterize a wide variety of aberrations an optical device is likely to encounter, there is no known "fast Zernike transform." Projections onto the truncated series can quickly become computationally intensive. Zernike polynomials have been proposed as parameterizations for atmospheric turbulence [134], though more recent research recommends a truncated singular value decomposition (SVD) [186–188]. Both Zernike polynomials and the SVD are regularization strategies based on physical phenomena at the pupil plane. This is an issue separate from the design of numerical algorithms. It was shown in the previous section that regularization, or equivalently filtering, is closely related to the statistical properties of the optical system. It has been noted in [116] that choronographic observations, for example, have very different noise characteristics than planetary observations or atmospheric observations. These considerations are best dealt with entirely through the choice of error metric and have little to do at the outset with the parameterization. There is no question as to the importance of regularization, however this is independent of the numerical methodology. Our goal is to develop a methodology with as wide an applicability as possible. The techniques discussed in Chapter 5 have been designed with the worst case scenario in mind, that is systems with more than 500,000 variables. They certainly are not limited to such large problems and are competitive methods for smaller problems as well.

Chapter 5

NUMERICAL METHODS

In this chapter we present two basic numerical approaches for the minimization of the perturbed least squares objective, E_ϵ, and the perturbed extended least squares objective, R_ϵ. The first algorithm is a simple first-order line search method while the second is a trust region algorithm that incorporates curvature information using limited memory techniques.

5.1 Line Search

Let $F : L^2[\mathbb{R}^2, \mathbb{R}^2] \to \mathbb{R}$ be Fréchet differentiable. Given an initial estimate of the solution $\boldsymbol{u}^{(0)}$, a descent algorithm for the minimization of F generates iterates $\boldsymbol{u}^{(\nu)}$ by the rule

$$\boldsymbol{u}^{(\nu+1)} = \boldsymbol{u}^{(\nu)} + \lambda^{(\nu)} \boldsymbol{w}^{(\nu)} \tag{5.1}$$

where

$$\boldsymbol{w}^{(\nu)} \in \mathbb{D}[\boldsymbol{u}^{(\nu)}] = \left\{ \boldsymbol{w} \in L^2[\mathbb{R}^2, \mathbb{R}^2] \;\middle|\; F'[\boldsymbol{u}^{(\nu)}][\boldsymbol{w}] < 0 \right\}$$

and $\lambda^{(\nu)}$ is a well chosen step-length parameter.

There are several methods for computing a suitable step length [133]. The criteria we use is the *sufficient decrease* condition:

$$F[\boldsymbol{u}^{(\nu)} + \lambda^{(\nu)} \boldsymbol{w}^{(\nu)}] \;\leq\; F^{(\nu)} + \eta \lambda^{(\nu)} \left\langle \nabla F^{(\nu)},\ \boldsymbol{w}^{(\nu)} \right\rangle \tag{5.2}$$

where $0 < \eta < 1$ is a fixed parameter and

$$F^{(\nu)} \equiv F[\boldsymbol{u}^{(\nu)}] \qquad \text{and} \qquad \nabla F^{(\nu)} \equiv \nabla F[\boldsymbol{u}^{(\nu)}].$$

Theorem 5.1.1 *Let $F : L^2[\mathbb{R}^2, \mathbb{R}^2] \to \mathbb{R}$ be Fréchet differentiable and bounded below. Consider the following algorithm.*

Step 0: *(Initialization) Choose $\gamma \in (0,1)$, $\eta \in (0,1)$, $c \geq 1$ and $\boldsymbol{u}^{(0)} \in L^2[\mathbb{R}^2, \mathbb{R}^2]$, and set $\nu = 0$.*

Step 1: *(Search Direction) If $\mathbb{D}[\boldsymbol{u}^{(\nu)}] = \emptyset$, STOP; otherwise, choose $\boldsymbol{w}^{(\nu)} \in \mathbb{D}[\boldsymbol{u}^{(\nu)}] \cap c\mathbb{B}$ where $\mathbb{B}$ is the closed unit ball in $L^2[\mathbb{R}^2, \mathbb{R}^2]$.*

Step 2: *(Step Length) Set*

$$\begin{array}{rll} \lambda^{(\nu)} \equiv & \textit{maximize} & \gamma^s \\ & \textit{subject to} & s \in \mathbb{N} \equiv \{0, 1, 2, \dots\} \\ & \textit{with} & F[\boldsymbol{u}^{(\nu)} + \gamma^s \boldsymbol{w}^{(\nu)}] - F^{(\nu)} \leq \eta \gamma^s \left\langle \nabla F^{(\nu)},\ \boldsymbol{w}^{(\nu)} \right\rangle. \end{array}$$

Step 3: *(Update) Set* $\boldsymbol{u}^{(\nu+1)} \equiv \boldsymbol{u}^{(\nu)} + \lambda^{(\nu)}\boldsymbol{w}^{(\nu)}$ *and* $\nu = \nu + 1$. *Return to Step 1.*

If ∇F *is globally Lipschitz continuous then the sequence* $\{\boldsymbol{u}^{(\nu)}\}$ *satisfies*

$$\left\langle \nabla F^{(\nu)},\ \boldsymbol{w}^{(\nu)} \right\rangle \to 0.$$

In particular, if $\boldsymbol{w}^{(\nu)}$ *is chosen so that*

$$\boldsymbol{w}^{(\nu)} = -\frac{\tilde{c}}{\|\nabla F^{(\nu)}\|}\nabla F^{(\nu)}$$

for $0 < \tilde{c} \leq c$, *then*

$$\|\nabla F^{(\nu)}\| \to 0.$$

PROOF: The proof is by contradiction. Suppose there is a subsequence $\mathbb{K} \subset \mathbb{N}$ such that $\sup_{\mathbb{K}} \left\langle \nabla F^{(\nu)}, \boldsymbol{w}^{(\nu)} \right\rangle < \beta < 0$. Since F is bounded below, $F^{(\nu)} \searrow F_* \in \mathbb{R}$, and so $(F^{(\nu+1)} - F^{(\nu)}) \to 0$. By the choice of $\lambda^{(\nu)}$ in Step 2 we have that

$$\lambda^{(\nu)} \left\langle \nabla F^{(\nu)},\ \boldsymbol{w}^{(\nu)} \right\rangle \to 0.$$

Therefore $\lambda^{(\nu)} \underset{\mathbb{K}}{\to} 0$ and so, without loss of generality, $\lambda^{(\nu)} < 1$ for all $\nu \in \mathbb{K}$. Hence,

$$\eta\lambda^{(\nu)}\gamma^{-1}\langle \nabla F^{(\nu)},\ \boldsymbol{w}^{(\nu)}\rangle < F[\boldsymbol{u}^{(\nu)} + \lambda^{(\nu)}\gamma^{-1}\boldsymbol{w}^{(\nu)}] - F^{(\nu)} \tag{5.3}$$

for all $\nu \in \mathbb{K}$. Let K be the global Lipschitz constant for ∇F. Then

$$F[\boldsymbol{u}^{(\nu)} + \lambda^{(\nu)}\gamma^{-1}\boldsymbol{w}^{(\nu)}] - F^{(\nu)} \leq \lambda^{(\nu)}\gamma^{-1}\left[\langle \nabla F^{(\nu)},\ \boldsymbol{w}^{(\nu)}\rangle + K(\lambda^{(\nu)}\gamma^{-1}\|\boldsymbol{w}^{(\nu)}\|)\right]. \tag{5.4}$$

Together, Eq.(5.3)-(5.4) yield

$$0 < (1-\eta)\beta + K(\lambda^{(\nu)}\gamma^{-1}\|\boldsymbol{w}^{(\nu)}\|).$$

Taking limits over $\nu \in \mathbb{K}$,

$$\lambda^{(\nu)}\gamma^{-1}\|\boldsymbol{w}^{(\nu)}\| \to 0 \qquad \Longrightarrow \qquad K(\lambda^{(\nu)}\gamma^{-1}\|\boldsymbol{w}^{(\nu)}\|) \to 0$$

which yields the contradiction $0 < (1-\eta)\beta < 0$.

We next show convergence of the norm of the gradient to zero for

$$\boldsymbol{w}^{(\nu)} = -\frac{\tilde{c}}{\|\nabla F^{(\nu)}\|}\nabla F^{(\nu)}$$

where $0 < \tilde{c} \leq c$. This is a direction of descent lying within $c\mathbb{B}$. Thus, for this choice of $\boldsymbol{w}^{(\nu)}$,

$$\left\langle \nabla F^{(\nu)},\ \boldsymbol{w}^{(\nu)} \right\rangle = -\tilde{c}\|\nabla F^{(\nu)}\| \to 0.$$

$\square$

5.2 Acceleration Techniques: Limited Memory BFGS with Trust Regions

The line search method discussed in the previous section is a first-order method and cannot be expect to converge quickly. Indeed, in experiments discussed in Chapter 6 we observe very poor rates of convergence for this method. In the remainder of this chapter we study techniques for accelerating convergence and lowering CPU time. These techniques depend to a large degree on the size of the problem. For problems with a hundred unknowns and a twice differentiable objective, it is feasible to use Newton's method to achieve near quadratic convergence close to a local solution. For large problems, however, this is not a reasonable strategy. For example, if the objective $F[\boldsymbol{u}]$ for the wavefront reconstruction problems Pr.(3.33) and Pr.(3.94) is discretized into a pixel basis for a 512×512 image, the number of unknowns is 2^{10}. The Hessian corresponding to a system of 2^{10} unknowns, assuming it exists, is a dense $2^{10} \times 2^{10}$ matrix. Explicit representation of this matrix is not practical. For this reason, many researchers discretize the problem into polynomial expansions that are truncated to achieve low dimension optimization problems. While this is sometimes physically justified, it is not always the case that one can find an efficient parameterization that is consistent with the physics of image observation discussed in Chapter 4. The methodology we study here does not rely on truncated parameterizations to achieve accelerated convergence of algorithms and efficient use of CPU time.

Limited memory methods provide an efficient way to use approximate Hessian information without explicitly forming the matrix. These methods are derived from matrix secant methods that approximate curvature information of the objective function from preceding steps and gradients. In this section we derive a compact representation of the Broyden-Fletcher-Goldfarb (BFGS) matrix secant update and it's compact representation for limited memory implementations. Limited memory methods are made robust with the introduction of trust regions. For a thorough treatment of matrix secant and trust region methods see Ref. [55].

5.2.1 Matrix secants and the BFGS update

Denote the discretized unknown functions $\boldsymbol{u}$ by the same variable with the 2 dimensions stacked into one column vector, *i.e* $\boldsymbol{u} \in \mathbb{R}^n$ for some integer n. Matrix secant iterates are generated by

$$u^{(\nu+1)} = u^{(\nu)} - \left(M^{(\nu)}\right)^{-1} \nabla F(u^{(\nu)}) \tag{5.5}$$

where $M^{(\nu)} \in \mathbb{R}^{n\times n}$ is an approximation to $\nabla^2 F^{(\nu)}$ satisfying the matrix secant equation:

$$M^{(\nu)}(\boldsymbol{u}^{(\nu-1)} - \boldsymbol{u}^{(\nu)}) = \nabla F^{(\nu-1)} - \nabla F^{(\nu)}. \tag{5.6}$$

Equation (5.6) is a system of n equations in n^2 unknowns, thus infinitely many solutions are possible. Common choices for the secant approximation $M^{(\nu)}$ are Broyden's update, the symmetric rank one (SR1) update, and the BFGS update. Limited memory techniques for BFGS matrices are reviewed here, however similar methods for alternative updates are possible.

The BFGS update to the true Hessian is given by

$$M^{(\nu)} = M^{(\nu-1)} + \frac{\boldsymbol{y}^{(\nu)}\boldsymbol{y}^{(\nu)^T}}{\boldsymbol{y}^{(\nu)^T}\boldsymbol{s}^{(\nu)}} - \frac{M^{(\nu-1)}\boldsymbol{s}^{(\nu)}\boldsymbol{s}^{(\nu)^T}M^{(\nu-1)}}{\boldsymbol{s}^{(\nu)^T}M^{(\nu-1)}\boldsymbol{s}^{(\nu)}}, \quad \nu = 1, 2, \ldots \tag{5.7}$$

where

$$\boldsymbol{y}^{(\nu)} \equiv \nabla F^{(\nu+1)} - \nabla F^{(\nu)}, \qquad \boldsymbol{s}^{(\nu)} \equiv \boldsymbol{u}^{(\nu+1)} - \boldsymbol{u}^{(\nu)}. \tag{5.8}$$

The BFGS approximation is symmetric and positive definite as long as $\boldsymbol{s}^{(\nu)^T}\boldsymbol{y}^{(\nu)} > 0$ and $M^{(\nu-1)}$ is symmetric and positive definite. Additionally, the BFGS update has a explicit recursive formula for the inverse

$$[M^{(\nu)}]^{-1} = V^{(\nu-1)^T}[M^{(\nu-1)}]^{-1}V^{(\nu-1)} + z^{(\nu-1)}s^{(\nu-1)}s^{(\nu-1)^T} \tag{5.9}$$

where

$$V^{(\nu)} \equiv I - z^{(\nu)}y^{(\nu)}s^{(\nu)^T}, \qquad z^{(\nu)} \equiv (y^{(\nu)^T}s^{(\nu)})^{-1}.$$

Now consider the recursion

$$\begin{aligned}
&[M^{(\nu)}]^{-1} = (V^{(\nu-1)^T} \ldots V^{(\nu-m)^T})\left[M^{(0,\nu)}\right]^{-1}(V^{(\nu-m)} \ldots V^{(\nu-1)}) \\
&+z^{(\nu-m)}(V^{(\nu-1)^T} \ldots V^{(\nu-m+1)^T})s^{(\nu-m)}s^{(\nu-m)^T}(V^{(\nu-m+1)} \ldots V^{(\nu-1)}) \\
&+z^{(\nu-m+1)}(V^{(\nu-1)^T} \ldots V^{(\nu-m+2)^T})s^{(\nu-m+1)}s^{(\nu-m+1)^T}(V^{(\nu-m+2)} \ldots V^{(\nu-1)}) \\
&+\ldots \\
&+z^{(\nu-1)}s^{(\nu-1)}s^{(\nu-1)^T}
\end{aligned} \tag{5.10}$$

Here m is the number of $\{y^{(i)}, s^{(i)}\}$ pairs that are stored, $i = \nu - m, \ldots, \nu - 1$, and $\left[M^{(0,\nu)}\right]^{-1}$ is a generating matrix specific to the νth iterate. The $[M^{(\nu)}]^{-1}$ defined by (5.10) is equivalent to the inverse BFGS matrix generated by updating the initial matrix $M^{(0,\nu)}$ m times according to the recursion (5.9). When $m = \nu$ and the generating matrix $M^{(0,\nu)} = M_0$ for all ν then the formula (5.10) is identical to (5.9). When the number of updates is zero, *i.e.* $m = 0$, and $\left[M^{(0,\nu)}\right]^{-1} = M_0^{-1} = I$ for all iterates ν the recursion (5.7) simply yields the identity for the Hessian approximation at every iteration. This corresponds to steepest descent in the sequence defined by (5.5). Limited memory methods occupy a middle ground between the two cases, $m = 0$ and $m = \nu$, with a generating matrix $M^{(0,\nu)}$ that changes at each iteration. Limited memory is a convenient way to make use of curvature information for high dimensional problems, however two-step quadratic convergence rates cannot reasonably be expected for $\nu \gg m$.

The above recursions are helpful for illustration but are not used in practice. Compact representations of the BFGS approximation provide for efficient implementations of BFGS matrix secants. Moreover, as with conjugate gradients, the product $[M^{(\nu)}]^{-1}\nabla f(u^{(\nu)})$ is computed without actually forming the matrix $[M^{(\nu)}]^{-1}$. Before proceeding we introduce some notation.

- $S^{(\nu)} \equiv [s^{(\nu-m)}, \ldots, s^{(\nu-1)}] \in \mathbb{R}^{n\times m}$;

- $Y^{(\nu)} \equiv [y^{(\nu-m)}, \ldots, y^{(\nu-1)}] \in \mathbb{R}^{n\times m}$;
- ${S^{(\nu)}}^T Y^{(\nu)} = L^{(\nu)} + D^{(\nu)} + R^{(\nu)} \in \mathbb{R}^{m\times m}$ where $L^{(\nu)}$, $R^{(\nu)}$, and $D^{(\nu)}$ are lower triangular, upper triangular and diagonal matrices respectively;
- $\overline{R}^{(\nu)} \equiv D^{(\nu)} + R^{(\nu)}$;
- $\Psi^{(\nu)} \equiv [M_0 S^{(\nu)} \quad Y^{(\nu)}] \in \mathbb{R}^{n\times 2m}$;
- $\widetilde{\Psi}^{(\nu)} \equiv [S^{(\nu)} \quad M_0^{-1} Y^{(\nu)}] \in \mathbb{R}^{n\times 2m}$;
- $\Gamma^{(\nu)} \equiv \begin{bmatrix} {S^{(\nu)}}^T M_0 S^{(\nu)} & L^{(\nu)} \\ {L^{(\nu)}}^T & -D^{(\nu)} \end{bmatrix} \in \mathbb{R}^{2m\times 2m}$;
- $\widetilde{\Gamma}^{(\nu)} \equiv \begin{bmatrix} {\overline{R}^{(\nu)}}^{-T} \left(D^{(\nu)} + {Y^{(\nu)}}^T M_0^{-1} Y^{(\nu)} \right) {\overline{R}^{(\nu)}}^{-1} & -\overline{R}^{(\nu)}{}^{-T} \\ -{\overline{R}^{(\nu)}}^{-1} & 0 \end{bmatrix} \in \mathbb{R}^{2m\times 2m}$.

Let $M^{(\nu)}$ be the BFGS approximation at the νth iterate with the symmetric positive definite generating matrix M_0. Let the ν pairs $\{y^{(i)}, s^{(i)}\}_{i=1}^{\nu-1}$ satisfy ${s^{(i)}}^T y^{(i)} > 0$. Byrd *et al* [36] show that for $m = \nu$ in the above definitions,

$$M^{(\nu)} = M_0 - \Psi^{(\nu)} {\Gamma^{(\nu)}}^{-1} {\Psi^{(\nu)}}^T, \tag{5.11}$$

and

$$[M^{(\nu)}]^{-1} = M_0^{-1} - \widetilde{\Psi}^{(\nu)} \widetilde{\Gamma}^{(\nu)} [\widetilde{\Psi}^{(\nu)}]^T. \tag{5.12}$$

In (5.12), assuming M_0^{-1} is given, the only inverse to be computed is that of the $\nu \times \nu$ upper triangular matrix $\overline{R}^{(\nu)}$. This is easily accomplished with back substitution.

5.2.2 Limited Memory BFGS (L-BFGS)

Limited memory techniques amount to generating *at each iteration* the BFGS matrix from the m most recent of the pairs $\{y^{(i)}, s^{(i)}\}_{i=\nu-m}^{\nu-1}$ and the generating matrix $M^{(0,\nu)}$. Typically $m \in [5, 10]$. The choice of $M^{(0,\nu)}$ that is often used is $M^{(0,\nu)} = \mu^{(\nu)} I$ where I is the identity matrix and $\mu^{(\nu)}$ is some scaling (see [166]). With this generating matrix the only inverse one need compute in (5.12) is that of the $m \times m$ upper triangular matrix $\overline{R}^{(\nu)}$. This is easily accomplished with back substitution. The quasi-Newton iteration then yields

$$u^{(\nu+1)} = u^{(\nu)} - \frac{1}{\mu^{(\nu)}} \nabla F(u^{(\nu)}) - \widetilde{\Psi}^{(\nu)} \widetilde{\Gamma}^{(\nu)} \left([\widetilde{\Psi}^{(\nu)}]^T \nabla F(u^{(\nu)}) \right). \tag{5.13}$$

The complexity of this operation is $O(mn)$ while the complexity of computing $\widetilde{\Gamma}^{(\nu)}$ is $O(m^3)$.

5.2.3 Trust Region L-BFGS

Acceptance of the step to the next iterate depends on the accuracy of the quadratic approximation

$$\widetilde{F}^{(\nu+1)} = F^{(\nu)} + \nabla F^{(\nu)^T} \cdot \boldsymbol{s}^{(\nu)} + \frac{1}{2}\boldsymbol{s}^{(\nu)^T} M^{(\nu)} \boldsymbol{s}^{(\nu)} \tag{5.14}$$

against the true function value $F^{(\nu+1)}$. A measurement of this accuracy is given by the ratio of the actual change in the function value between iterates $\boldsymbol{u}^{(\nu)}$ and $\boldsymbol{u}^{(\nu+1)}$ and the predicted change,

$$\rho(\boldsymbol{s}^{(\nu)}) = \frac{\text{actual change}^{(\nu)}}{\text{predicted change}^{(\nu)}} = \frac{F^{(\nu)} - F^{(\nu+1)}}{-\nabla F^{(\nu)^T} \cdot \boldsymbol{s}^{(\nu)} - \frac{1}{2}\boldsymbol{s}^{(\nu)^T} M^{(\nu)} \boldsymbol{s}^{(\nu)}}. \tag{5.15}$$

If the ratio is below some tolerance $\tilde{\eta}$ then the step is restricted. A line search strategy such as the one given in Theorem 5.1.1 can be employed to find an acceptable step size, however this often requires several function evaluations. In applications such as nonparametric phase retrieval, function and gradient evaluations are the most expensive part of each iteration, thus we consider alternative strategies for finding acceptable steps. We have found in practice that a single application of a *trust region* strategy is usually all that is required to find a step that satisfies Eq.(5.2). A trust region is a ball around the current iterate $\boldsymbol{u}^{(\nu)}$ within which the quadratic approximation is reliable.

The trust region subproblem with trust region radius $\Delta^{(\nu)}$ is given by

$$\begin{array}{lll} TR(\Delta^{(\nu)}) & \textit{minimize} & \nabla F^{(\nu)^T} \boldsymbol{s} + \frac{1}{2}\boldsymbol{s}^T M^{(\nu)} \boldsymbol{s}. \\ & \|\boldsymbol{s}\| \leq \Delta^{(\nu)} & \end{array}$$

The Lagrangian of $TR(\Delta^{(\nu)})$ yields the following unconstrained, *implicit* trust region subproblem

$$\begin{array}{lll} TR'(\omega^{(\nu)}) & \textit{minimize} & \nabla F(u^{(\nu)})^T s + \frac{1}{2} s^T \left(M^{(\nu)} + \omega^{(\nu)} I\right) s. \\ & s \in \mathbb{R}^n & \end{array}$$

A solution, $s_*(\omega^{(\nu)})$, to $TR'(\omega^{(\nu)})$ corresponds to a solution to $TR(\Delta^{(\nu)})$ with $\Delta^{(\nu)} = \|s_*(\omega^{(\nu)})\|$. The larger $\omega^{(\nu)}$ the smaller the trust region radius $\Delta^{(\nu)}$.

In [33] Burke and Wiegmann derive a compact representation of the inverse of the matrix $\omega^{(\nu)} I + M^{(\nu)}$ for solving the trust region subproblem that can be computed with the same computational complexity as the computation of $M^{(\nu)}$. To see how this is done, let $\tau^{(\nu)} = \omega^{(\nu)} + \mu^{(\nu)}$. Recall that the generating matrix for the νth iterate is given by $M^{(0,\nu)} = \mu^{(\nu)} I$. For $\Gamma^{(\nu)}$ invertible,

$$M^{(\nu)} + \omega^{(\nu)} I = \tau^{(\nu)} I - \Psi^{(\nu)} [\Gamma^{(\nu)}]^{-1} \Psi^{(\nu)^T}. \tag{5.16}$$

If $M^{(\nu)} + \omega^{(\nu)} I$ is invertible (*i.e.* as long as $D^{(\nu)}$ is positive definite), the Sherman-Morrison-Woodbury formula yields

$$\left[\tau^{(\nu)} I - \Psi^{(\nu)} [\Gamma^{(\nu)}]^{-1} \Psi^{(\nu)^T}\right]^{-1} = (\tau^{(\nu)})^{-1} \left[I + \Psi^{(\nu)} \left(\tau^{(\nu)} \Gamma^{(\nu)} - \Psi^{(\nu)^T} \Psi^{(\nu)}\right)^{-1} \Psi^{(\nu)^T}\right] \tag{5.17}$$

where

$$\begin{aligned}&\tau^{(\nu)}\Gamma^{(\nu)} - {\Psi^{(\nu)}}^T\Psi^{(\nu)}\\ &= \left[\begin{array}{ll} \mu^{(\nu)}\omega^{(\nu)}{S^{(\nu)}}^T S^{(\nu)} & \omega^{(\nu)}L^{(\nu)} - \mu^{(\nu)}\overline{R}^{(\nu)} \\ \omega^{(\nu)}{L^{(\nu)}}^T - \mu^{(\nu)}{\overline{R}^{(\nu)}}^T & -\left({Y^{(\nu)}}^T Y^{(\nu)} + (\mu^{(\nu)} + \omega^{(\nu)})D^{(\nu)}\right)\end{array}\right] \in \mathbb{R}^{2m\times 2m}.\end{aligned} \tag{5.18}$$

The inverse of (5.18) can be computed efficiently using the Cholesky factorization. The factorization involves the related matrix

$$\begin{aligned}&\left[\begin{array}{ll} -\left({Y^{(\nu)}}^T Y^{(\nu)} + (\mu^{(\nu)} + \omega^{(\nu)})D^{(\nu)}\right) & \omega^{(\nu)}{L^{(\nu)}}^T - \mu^{(\nu)}{\overline{R}^{(\nu)}}^T \\ \omega^{(\nu)}L^{(\nu)} - \mu^{(\nu)}\overline{R}^{(\nu)} & \mu^{(\nu)}\omega^{(\nu)}{S^{(\nu)}}^T S^{(\nu)}\end{array}\right] = \\ &\left[\begin{array}{ll} K^{(\nu)} & 0 \\ -(\omega^{(\nu)}L^{(\nu)} - \mu^{(\nu)}\overline{R}^{(\nu)})[K^{(\nu)}]^{-T} & B^{(\nu)}\end{array}\right]\left[\begin{array}{ll} -{K^{(\nu)}}^T & [K^{(\nu)}]^{-1}(\omega^{(\nu)}L^{(\nu)} - \mu^{(\nu)}\overline{R}^{(\nu)})^T \\ 0 & {B^{(\nu)}}^T\end{array}\right].\end{aligned} \tag{5.19}$$

Here $K^{(\nu)}$ is the lower triangular Cholesky factor of ${Y^{(\nu)}}^T Y^{(\nu)} + (\mu^{(\nu)} + \omega^{(\nu)})D^{(\nu)}$ and $B^{(\nu)}$ is the lower triangular Cholesky factor satisfying

$$\begin{aligned}&B^{(\nu)}{B^{(\nu)}}^T = \omega^{(\nu)}\mu^{(\nu)}{S^{(\nu)}}^T S^{(\nu)}\\ &+ \left(\omega^{(\nu)}L^{(\nu)} - \mu^{(\nu)}\overline{R}^{(\nu)}\right)\left({Y^{(\nu)}}^T Y^{(\nu)} + (\mu^{(\nu)} + \omega^{(\nu)})D^{(\nu)}\right)\left(\omega^{(\nu)}L^{(\nu)} - \mu^{(\nu)}\overline{R}^{(\nu)}\right)^T.\end{aligned}$$

It is straight forward to show that $K^{(\nu)}$ and $B^{(\nu)}$ exist and are nonsingular. The following lemma is an extension of a proof given by Byrd *et al* for the computation of $\Gamma^{(\nu)}$ [36].

Lemma 5.2.1 *If* ${y^{(i)}}^T s^{(i)} > 0$ *for all* $i = \nu - m, \ldots, \nu - 1$ *and* $\omega^{(\nu)}$, $\mu^{(\nu)} > 0$, *then* ${Y^{(\nu)}}^T Y^{(\nu)} + (\mu^{(\nu)} + \omega^{(\nu)})D^{(\nu)}$ *and*

$$\begin{aligned}&\omega^{(\nu)}\mu^{(\nu)}{S^{(\nu)}}^T S^{(\nu)}\\ &+ \left(\omega^{(\nu)}L^{(\nu)} - \mu^{(\nu)}\overline{R}^{(\nu)}\right)\left({Y^{(\nu)}}^T Y^{(\nu)} + (\mu^{(\nu)} + \omega^{(\nu)})D^{(\nu)}\right)\left(\omega^{(\nu)}L^{(\nu)} - \mu^{(\nu)}\overline{R}^{(\nu)}\right)^T\end{aligned}$$

are positive definite.

PROOF: By definition, ${y^{(i)}}^T s^{(i)} > 0$ for all $i = \nu - m, \ldots, \nu - 1$ implies that $D^{(\nu)}$ is positive definite, hence ${Y^{(\nu)}}^T Y^{(\nu)} + (\mu^{(\nu)} + \omega^{(\nu)})D^{(\nu)}$ is also positive definite as long as $\mu^{(\nu)}, \omega^{(\nu)} > 0$, and

$$\left(\omega^{(\nu)}L^{(\nu)} - \mu^{(\nu)}\overline{R}^{(\nu)}\right)\left({Y^{(\nu)}}^T Y^{(\nu)} + (\mu^{(\nu)} + \omega^{(\nu)})D^{(\nu)}\right)\left(\omega^{(\nu)}L^{(\nu)} - \mu^{(\nu)}\overline{R}^{(\nu)}\right)^T$$

is positive semidefinite. Now suppose that for some v

$$\begin{aligned}&v^T\left(\omega^{(\nu)}L^{(\nu)} - \mu^{(\nu)}\overline{R}^{(\nu)}\right)\left({Y^{(\nu)}}^T Y^{(\nu)} + (\mu^{(\nu)} + \omega^{(\nu)})D^{(\nu)}\right)\left(\omega^{(\nu)}L^{(\nu)} - \mu^{(\nu)}\overline{R}^{(\nu)}\right)^T v\\ &= 0\end{aligned} \tag{5.20}$$

Then $\overline{R}^{(\nu)^T} v = L^{(\nu)^T} v = 0$. Recall that $\overline{R}^{(\nu)} = D^{(\nu)} + R^{(\nu)}$, thus the only v satisfying (5.20) is $v = 0$.

□

To avoid notational clutter, we drop the index (ν) from the iterates. Solving $TR(\Delta)$, assuming that the solution lies on the boundary of the trust region, can be recast as solving the system of equations

$$(M + \omega I)s + \nabla F = 0 \tag{5.21}$$

$$\|s\|^2 - \Delta^2 = 0. \tag{5.22}$$

Moré and Sorenson [130] propose an efficient way of solving this system by using Newton's method to find the zeros of $\phi(\omega)$ where

$$\phi(\omega) = \frac{1}{\Delta} - \frac{1}{\|s(\omega)\|} \tag{5.23}$$

and $s(\omega) = -(M + \omega I)^{-1}\nabla F$. This form of ϕ was first proposed by Reinsch [149]. Newton's iteration for solving $\phi(\omega) = 0$ yields

$$\omega^{(j+1)} = \omega^{(j)} - \frac{\phi(\omega^{(j)})}{\phi'(\omega^{(j)})} \tag{5.24}$$

where

$$\phi'(\omega) = -\frac{\nabla F^T [M + \omega I]^{-3} \nabla F}{\|s(\omega)\|^3}. \tag{5.25}$$

Burke [32] has derived the formula for the general nth inverse of matrices in Sherman-Morrison-Woodbury form which yields an explicit formula for $[M + \omega I]^{-3}$. Since $[M + \omega I]^{-1}$ is used in other computations, it is more efficient to compute

$$(M + \omega I)^{-3} = (M + \omega I)^{-2}(M + \omega I)^{-1}$$

with

$$(M + \omega I)^{-2} = \frac{1}{\tau^2}\left[I + \Psi\left(\tau\Gamma - \Psi^T\Psi\right)^{-1}\Psi^T + \tau\Psi\left(\tau\Gamma - \Psi^T\Psi\right)^{-1}\Gamma\left(\tau\Gamma - \Psi^T\Psi\right)^{-1}\Psi^T\right]. \tag{5.26}$$

Setting $v_0 = \Psi^T\nabla F$, $v_1 = (\tau\Gamma - \Psi^T\Psi)^{-1}v_0$, and $v_2 = (\tau\Gamma - \Psi^T\Psi)^{-1}\Gamma v_1$ iteration (5.24) can be written as

$$\omega^{(j+1)} = \omega^{(j)} - \frac{\sigma}{\delta}\left[\frac{\sqrt{\sigma}}{\Delta} - \tau\right] \tag{5.27}$$

where

$$\sigma = \tau^2\|s(\omega)\|^2 = v_1^T\Psi^T\Psi v_1 + 2v_0^T v_1 + \|\nabla F\|^2 \quad \text{and}$$

$$\delta = \tau^3\left[\nabla F\right]^T(\omega I + M)^{-3}\nabla F = \sigma + \tau\left[v_1^T\Psi^T\Psi v_2 + v_0^T v_2\right].$$

The step calculated by the trust region subproblem is given by

$$s = \frac{-1}{\tau}(\nabla F + \Psi v_1) \tag{5.28}$$

with $\|s\| = \frac{\sigma}{\tau^2}$. The change in the function predicted by the quadratic model (5.14) is given by

$$\text{predicted change } = [\nabla F]^T s + \frac{1}{2} s^T M s = \frac{-1}{2\tau}([\nabla F]^T[\nabla F] + v_0^T v_1 + \frac{\omega}{\tau}\sigma). \tag{5.29}$$

5.2.4 Algorithms and Implementation

A crucial parameter in matrix secant methods is the scaling $\mu^{(\nu)}$. There are many definitions for the optimal $\mu^{(\nu)}$ [138]. One such scaling suggested by Shanno and Phua [166] is

$$\mu^{(\nu)} = \frac{\boldsymbol{y}^{(\nu-1)^T} \boldsymbol{y}^{(\nu-1)}}{\boldsymbol{s}^{(\nu-1)^T} \boldsymbol{y}^{(\nu-1)}}. \tag{5.30}$$

As noted in [33], for the proper scaling the trust region is required only a small fraction of the time. The scaling has the effect of implicitly imposing a trust region. Since computations with trust regions are much more expensive than unconstrained L-BFGS, it is reasonable to default to unconstrained L-BFGS and only invoke the trust region when the objective value does not behave as predicted. The trust region is invoked only if the ratio $\rho(s^{(\nu)})$ given by Eq.(5.15) falls below a given tolerance, indicating that the quadratic model (5.14) is not reliable. It has also been noted in [33] that when a step does not give sufficient decrease in the objective value, or even causes an *increase* in the objective, it is still worthwhile to keep that step direction and use it to update the L-BFGS matrix, even though the step is not taken. This is because bad steps still contain curvature information, albeit information about curvature in the *wrong* direction. In our experiments the trust region was rarely restricted more than once before an acceptable step was found. The strategy of keeping even bad steps does not promise much savings for this problem, so have not included this in our implementations.

Algorithm 5.2.2 (Limited Memory BFGS with Trust Regions) *:*

Step 0: *(Initialization): Choose* $\tilde{\eta} > 0$, $\zeta > 0$, $\overline{m} \in \{1, 2, \ldots, n\}$, *and* $\boldsymbol{u}^{(0)} \in \mathbb{R}^n$, *and set* $\nu = m = 0$. *Compute* $\nabla F^{(0)}$, $F^{(0)}$ *and* $\|\nabla F^{(0)}\|$.

Step 1: *(L-BFGS step) If m=0 compute* $\boldsymbol{u}^{(\nu+1)}$ *by some line search algorithm (*e.g. *the algorithm in Theorem 5.1.1)); otherwise compute* $\boldsymbol{s}^{(\nu)} = -\left(M^{(\nu)}\right)^{-1} \nabla F^{(\nu)}$ *where* $M^{(\nu)}$ *is the L-BFGS update [36],* $\boldsymbol{u}^{(\nu+1)} = \boldsymbol{u}^{(\nu)} + \boldsymbol{s}^{(\nu)}$, $F^{(\nu+1)}$, *and the predicted change Eq.(5.14).*

Step 2: *(Trust Region) If the step* $\boldsymbol{s}^{(\nu)}$ *violates the appropriate criteria (*e.g. $\rho(\boldsymbol{s}^{(\nu)}) < \tilde{\eta}$ *for* ρ *given by Eq.(5.15)) reduce the trust region* $\Delta^{(\nu)}$, *solve the trust region subproblem for* $\boldsymbol{s}^{(\nu)}$ *[33], and compute* $\boldsymbol{u}^{(\nu+1)} = \boldsymbol{u}^{(\nu)} + \boldsymbol{s}^{(\nu)}$, $F^{(\nu+1)}$, *and the predicted change Eq.(5.14). Repeat Step 2.*

Step 3: *(Update) Compute* $\nabla F^{(\nu+1)}$, $\|\nabla F^{(\nu+1)}\|$, $\boldsymbol{y}^{(\nu)}$ *from Eq.(5.8), and* ${\boldsymbol{s}^{(\nu)}}^T\boldsymbol{y}^{(\nu)}$. *Discard the vector pair* $\{\boldsymbol{s}^{(\nu-m)}, \boldsymbol{y}^{(\nu-m)}\}$ *from storage. If* ${\boldsymbol{s}^{(\nu)}}^T\boldsymbol{y}^{(\nu)} \leq \zeta$ *set* $m = \max\{m-1, 0\}$, $\Delta^{(\nu+1)} = \infty$, $\mu^{(\nu+1)} = \mu^{(\nu)}$, *and* $M^{(\nu+1)} = M^{(\nu)}$ *(*i.e. *shrink the memory and don't update); otherwise set* $\mu^{(\nu+1)} = \frac{{\boldsymbol{y}^{(\nu)}}^T\boldsymbol{y}^{(\nu)}}{{\boldsymbol{s}^{(\nu)}}^T\boldsymbol{y}^{(\nu)}}$, $\Delta^{(\nu+1)} = \infty$, $m = \min\{m+1, \overline{m}\}$, *add the vector pair* $\{\boldsymbol{s}^{(\nu)}, \boldsymbol{y}^{(\nu)}\}$ *to storage, and update* $M^{(\nu+1)}$ *[36]. Set* $\nu = \nu + 1$ *and return to Step 1.*

Remark 5.2.3 *With a slight modification Alg.5.2.2 can be used as a backtracking line search algorithm where* $\overline{m} = 1$ *and* $M^{(\nu)} = \mu^{(\nu)} I$ *for all* ν.

Algorithm 5.2.4 (Explicit Trust-Region Updating) *Given* $u^{(\nu)}$, $\nabla F(u^{(\nu)})$, $\mu^{(\nu)}$, $\tilde{\eta}$, *and* $0 < \beta_\Delta < 1$.

Step 0: *Calculate* $L^{(\nu)}$, $\Gamma^{(\nu)}$, $\Psi^{(\nu)}$ *and* ${\Psi^{(\nu)}}^T\Psi^{(\nu)}$. *Let* $\Delta^{(\nu)} = \|s^{(\nu-1)}\|$.

Step 1: *Let* s_+ *solve the trust region subproblem TR(*$\Delta^{(\nu)}$*) (see algorithm 5.2.5).*

Step 2: *If the ratio* $\rho(s_+) < \tilde{\eta}$ *set* $\Delta^{(\nu)} = \beta_\Delta \|s_+\|$ *and return to Step 1.; otherwise set* $s^{(\nu)} = s_+$, *and calculate* $u^{(\nu+1)} = u^{(\nu)} + s_+$, $\nabla F(u^{(\nu+1)})$, $y^{(\nu)}$, ${S^{(\nu-1)}}^T\nabla F(u^{(\nu+1)})$, ${Y^{(\nu-1)}}^T\nabla F(u^{(\nu+1)})$, *and the scalar* $\|\nabla F(u^{(\nu+1)})\|_2^2$

Step 3: *Return* $u^{(\nu+1)}$, $s^{(\nu)}$, $\nabla F(u^{(\nu+1)})$, $y^{(\nu)}$, ${S^{(\nu-1)}}^T\nabla F(u^{(\nu+1)})$, ${Y^{(\nu-1)}}^T\nabla F(u^{(\nu+1)})$, *and the scalar* $\|\nabla F(u^{(\nu+1)})\|_2^2$ *to the calling algorithm, and end.*

Algorithm 5.2.5 (Trust Region Subproblem)

Step 0: *(Initialization) Given matrices* $\Gamma^{(\nu)}$, $\Psi^{(\nu)}$, ${\Psi^{(\nu)}}^T\Psi^{(\nu)}$, $\mu^{(\nu)}$, ω, *and* $\Delta^{(\nu)}$. *Let* $v_0 = {\Psi^{(\nu)}}^T\nabla F(u^{(\nu)})$, *and let* ϵ *be the stopping tolerance:*

Step 1: *Set* $\tau = \omega + \mu^{(\nu)}$.

Step 2: *Set* $v_1 = (\tau\Gamma^{(\nu)} - {\Psi^{(\nu)}}^T\Psi^{(\nu)})^{-1}v_0$.

Step 3: *Set* $v_2 = (\tau\Gamma^{(\nu)} - {\Psi^{(\nu)}}^T\Psi^{(\nu)})^{-1}\Gamma^{(\nu)}v_1$.

Step 4: *Set* $\sigma = \tau^2\|s(\omega)\|^2 = v_1^T{\Psi^{(\nu)}}^T\Psi^{(\nu)}v_1 + 2v_0^Tv_1 + \|\nabla F(u^{(\nu)})\|^2$.

Step 5: *If* $|\sqrt{\sigma} - \tau\Delta^{(\nu)}| \leq \sqrt{\sigma}\Delta^{(\nu)}\|\nabla F(u^{(\nu)})\|\epsilon$, *goto Step 9.*

Step 6: *Set* $\delta = \sigma + \tau[v_1^T{\Psi^{(\nu)}}^T\Psi^{(\nu)}v_2 + v_0^Tv_2]$.

Step 7: *Set* $\omega_+ = \omega + \frac{\sigma}{\delta}\left[\frac{\sigma}{\Delta^{(\nu)}} - \tau\right]$.

Step 8: *If* $\omega_+ \geq 0$, *set* $\omega = \omega_+$; *otherwise, set* $\omega = .2\omega$. *Return to Step 1.*

Step 9: *Set* $s_+ = \frac{-1}{\tau}(\nabla F(u^{(\nu)}) + \Psi^{(\nu)} v_1)$

Step 10: *Calculate* $F(u^{(\nu)} + s_+)$, *the actual change, predicted change via (5.29), and* $\rho(s^{(\nu)})$ *via (5.15). If* $\rho(s_+) < \tilde{\eta}$ *adjust the trust region* $\Delta^{(\nu)}$ *and goto Step 0; otherwise, set* $s^{(\nu)} = s_+$.

Step 11: *Return* $s^{(\nu)}$ *to the calling algorithm and end.*

Algorithm 5.2.6 (Implicit Trust-Region Updating) *Given* $u^{(\nu)}$, $\nabla F(u^{(\nu)})$, $\mu^{(\nu)}$, $\tilde{\eta}$, *and* $0 < \beta_\omega$.

Step 0: *Calculate* $L^{(\nu)}$, $\Gamma^{(\nu)}$, $\Psi^{(\nu)}$ *and* ${\Psi^{(\nu)}}^T \Psi^{(\nu)}$. *Let* $\omega = \|\nabla F(u^{(\nu)})\|/(2\|s^{(\nu-1)}\|)$.

Step 1: *Set* $\tau = \omega + \mu^{(\nu)}$. *Let* s_+ *solve the trust region subproblem TR'(*ω*) via (5.28). Compute* $F(u^{(\nu)} + s_+)$, *the actual change, the predicted change via (5.29), and* $\rho(s_+)$ *via (5.15).*

Step 2: *If the ratio* $\rho(s_+) < \tilde{\eta}$ *set* $\omega = \omega + \beta_\omega(\omega + \mu^{(\nu)})$ *and return to 1.; otherwise set* $s^{(\nu)} = s_+$, *and calculate* $u^{(\nu+1)} = u^{(\nu)} + s_+$, $\nabla F(u^{(\nu+1)})$, $y^{(\nu)}$, ${S^{(\nu-1)}}^T \nabla F(u^{(\nu+1)})$, ${Y^{(\nu-1)}}^T \nabla F(u^{(\nu+1)})$, *and the scalar* $\|\nabla F(u^{(\nu+1)})\|_2^2$.

Step 3: *Return* $u^{(\nu+1)}$, $s^{(\nu)}$, , $\nabla F(u^{(\nu+1)})$, $y^{(\nu)}$, ${S^{(\nu-1)}}^T \nabla F(u^{(\nu+1)})$, ${Y^{(\nu-1)}}^T \nabla F(u^{(\nu+1)})$, *and the scalar* $\|\nabla F(u^{(\nu+1)})\|_2^2$ *to the calling algorithm, and end.*

5.3 Multi-resolution Analysis

In this section we discuss a further advantage of the pixel basis to other parameterizations, that is multi-resolution analysis. These methods are elementary and have been implemented without discussion in [115]. They have received more attention in a recent article by Ohneda [135]. For the wavefront reconstruction problem discussed in Chapter 3, multi-resolution techniques are the natural thing to do and are easily implemented. We motivate these methods with a discussion of filtering, and show the corresponding interpretation as a multi-resolution analysis.

Since image noise is often nonsmooth, it shows up as high frequency components of the Fourier transform of the noisy image. A common technique for separating out noise is to truncate the Fourier transform of the images. This is sometimes called *windowing* the Fourier transform of the image. For the imaging model given by Eq.(2.42), the image ψ_m is the magnitude of the Fourier transform of the aberrated generalized pupil function $\mathcal{R}[\boldsymbol{u}] \exp[\sqrt{-1}\,\tilde{\theta}_m]$ (see Eq.(2.43)). To filter noise from the wavefront estimate $\boldsymbol{u}$ one simply truncates the observed image to eliminate high frequency components of the estimate $\boldsymbol{u}$. Let $\mathcal{X}_n$ denote the indicator function for the $n \times n$ box of pixels centered at zero. For a discretized image ψ_m centered at zero we have the following system of equations for the filtered image

$$\mathcal{X}_{\tilde{n}} \odot |\mathcal{F}_m[\boldsymbol{u}]| = \mathcal{X}_{\tilde{n}} \odot \psi_m, \quad m = 1, \ldots, M. \tag{5.31}$$

where $\odot$ represents the discrete Hadamard matrix product and $\mathcal{F}_m$ $(m = 1, \ldots, M)$ are the discrete counterparts of the continuous operators defined in Eq. (2.43). Note that the filtering is *not* applied to the physical domain equation $(m = 0)$ given by Eq.(2.41) and Eq.(2.48). This has to do with the relation between filtering in the Fourier domain and blurring in the physical domain. This discussed in more detail below.

The multiresolution approach relies on our ability to write the left hand side of Eq.(5.31) as a localized average of nearby pixels of $\mathcal{R}[\boldsymbol{u}] \exp[\sqrt{-1}\, \tilde{\theta}_m]$, that is a low resolution version of the original function. To do this note that the pointwise modulus of functions in $L^2[\mathbb{R}^2, \mathbb{R}^2]$ is equivalent to the pointwise modulus of functions in $L^2[\mathbb{R}^2, \mathbb{C}]$. Thus

$$|\mathcal{F}_m[\boldsymbol{u}]| = |\mathcal{R}[\mathcal{F}_m[\boldsymbol{u}]]|$$

where $\mathcal{R}$ is the isomorphism between $L^2[\mathbb{R}^2, \mathbb{R}^2]$ and $L^2[\mathbb{R}^2, \mathbb{C}]$ (Eq.(2.76)). Since the Hadamard product commutes with the pointwise modulus function we may write the filtered function on the righthand side of Eq.(5.31) as

$$\mathcal{X}_{\tilde{n}} \odot |\mathcal{F}_m[\boldsymbol{u}]| = |\mathcal{X}_{\tilde{n}} \odot \mathcal{R}[\mathcal{F}_m[\boldsymbol{u}]]| \, .$$

Recalling Eq.(2.43) for $m \geq 1$, by the Discrete Convolution Theorem we have

$$|\mathcal{X}_{\tilde{n}} \odot \mathcal{R}[\mathcal{F}_m[\boldsymbol{u}]]| = \left| \left[\mathcal{X}_{\tilde{n}}^{\vee} * \mathcal{R}[\boldsymbol{u}] \exp[\sqrt{-1}\, \tilde{\theta}_m] \right]^{\wedge} \right| \qquad m = 1, \ldots, M.$$

Here $\wedge$ and $\vee$ indicate the discrete Fourier transform and it's inverse respectively.

For $x \in \mathbb{R}$, the Fourier transform of the window function is the sinc function defined by

$$\operatorname{sinc}(x) \equiv \frac{sin(\pi x)}{\pi x} = \mathcal{X}_{[-\tilde{x}, \tilde{x}]}^{\wedge}$$

where $\mathcal{X}_{[-\tilde{x}, \tilde{x}]}$ is the indicator function of the interval centered at zero of length $2\tilde{x}$. In n-dimensions, the Fourier transform of the window function is just the vector of sinc functions of each of the components separately. For $\boldsymbol{x} \in \mathbb{R}^n$

$$\operatorname{sinc}(\boldsymbol{x}) \equiv \left(\frac{sin(\pi x_1)}{\pi x_1}, \ldots, \frac{sin(\pi x_n)}{\pi x_n} \right).$$

For a review of these objects see [76, 98]. Convolution against a sinc function, $\mathcal{X}_{\tilde{n}}^{\vee}$, can be approximated by a localized discrete linear operator, $\mathcal{A}_n[\cdot]$, that averages blocks of adjacent pixels. For the moment we leave the definition of $\mathcal{A}_n[\cdot]$ ambiguous - many different averaging operators are possible. For $m \geq 1$ the convolution on the right hand side of Eq.(5.31) can therefore be approximated by

$$\mathcal{X}_{\tilde{n}}^{\vee} * \mathcal{R}[\boldsymbol{u}] \exp[\sqrt{-1}\, \tilde{\theta}_m] \approx \mathcal{A}_{\tilde{n}} \left[\mathcal{R}[\boldsymbol{u}] \exp[\sqrt{-1}\, \tilde{\theta}_m] \right]. \tag{5.32}$$

This yields the following approximation of Eq.(5.31)

$$\left|\left[\mathcal{A}_{\tilde{n}}\left[\mathcal{R}[\boldsymbol{u}]\exp[\sqrt{-1}\,\tilde{\theta}_m]\right]\right]^{\wedge}\right| \approx \mathcal{X}_{\tilde{n}} \odot \psi_m, \quad m = 1,\ldots,M. \tag{5.33}$$

The filtering operation applied to the images $\psi_m, \quad m = 1,\ldots,M$ cannot be directly applied the the physical domain constraint represented by the "image" ψ_0. The analog in the physical domain is an averaging operation. To see this consider the (discrete) Fourier dual of Eq.(2.49)

$$|\boldsymbol{u}|^{\wedge} = \psi_0^{\wedge}.$$

Now, apply the filter $\mathcal{X}_{\tilde{n}}$

$$\mathcal{X}_{\tilde{n}} \odot |\boldsymbol{u}|^{\wedge} = \mathcal{X}_{\tilde{n}} \odot \psi_0^{\wedge}.$$

Again, by the Discrete Convolution Theorem the Fourier dual of the filtering operation, *i.e.* the filtering operation in the physical domain, is given by

$$\mathcal{X}_{\tilde{n}}^{\vee} * |\boldsymbol{u}| = \mathcal{X}_{\tilde{n}}^{\vee} * \psi_0.$$

We approximate the right hand side of the above equation by $\mathcal{X}_{\tilde{n}}^{\vee} * |\boldsymbol{u}| \approx |\mathcal{A}_{\tilde{n}}\left[\mathcal{R}[\boldsymbol{u}]\right]|$ where $\mathcal{A}_{\tilde{n}}$ is the averaging operator discussed above. This yields the approximate physical domain relation corresponding to filtering in the Fourier domain

$$|\mathcal{A}_{\tilde{n}}[\mathcal{R}[\boldsymbol{u}]]| \approx \mathcal{X}_{\tilde{n}}^{\vee} * \psi_0. \tag{5.34}$$

Equations (5.33) and (5.34) constitute a low resolution imaging system. The averaging operator $\mathcal{A}_{\tilde{n}}$ blurs information in adjacent pixels of the wavefront estimate $\boldsymbol{u}$, smoothing out edges as well as noise. It is not necessary, therefore, to maintain a high pixelization for the wavefront estimate $\boldsymbol{u}$ since fine detail is lost by averaging. In Eq.(5.33) only the center $\tilde{n}$ pixels of the image are kept in the calculation. Our implementations rely on the Fast Fourier Transform Algorithm (FFT) to calculate the discrete Fourier transforms. These require square arrays with dimensions that are powers of 2. Our computations take advantage of the lower resolution image by using a pixelization of $\boldsymbol{u}$ that is consistent with the size of the window $\mathcal{X}_{\tilde{n}}$. This dramatically reduces the dimensionality of the optimization problem and thus computation time. It cannot be expected that the solution to the low resolution problem will be as good as the high resolution, however, we use the low resolution solutions as a bootstrap to higher resolution estimates. Ideally, all of the hard work is done at low resolution and relatively few iterations are necessary to achieve a solution at the highest resolution. This is indeed what we achieve (see Fig.(6.9)).

Chapter 6

NUMERICAL RESULTS

6.1 Phase retrieval

This chapter details the results of numerical experiments comparing the average performance of line search and Limited Memory BFGS (L-BFGS) methods with projection methods of similar type for noiseless and noisy data for the phase retrieval problem.

The aperture of the pupil consists of seven, meter-class panels shown in Fig.(6.2). This design is one of several configurations being studied at NASA's Goddard Space Flight Center for use on the Next Generation Space Telescope, Hubble's replacement. To recover the phase three diversity images are used, two out of focus and one in focus image. From this example the advantage of choosing a pixel basis over some parameterization (for example, Zernike polynomials [25, 26, 119, 204]) is apparent. Most obvious is the irregular shape of the pupil and the phase jumps across the separate panels which make it difficult to find an orthogonal parameterization [175]. Another advantage of the pixel basis is that it allows for the most accurate representation of the domain without introducing any regularization implicit in less precise parameterizations. Results for noisy data are shown in Fig.(6.3). Using a pixel basis the methods recover artifacts such as Gibbs phenomenon associated with the filtering of the data. Issues surrounding filtering and regularization of the data are independent of the numerical method and depend on the types of observations being made [116].

Two projection algorithms are compared to line search and L-BFGS algorithms for the least squares and extended least squares objectives Eq.(3.87) and Eq.(3.98). The first projection algorithm is evenly averaged ($\gamma_m^{(\nu)} = 1/4$ for all ν and $m = 0, \ldots, 3$) and unrelaxed ($\alpha_m = 1$ for all ν and $m = 0, \ldots, 3$) (Alg.(3.13)). This algorithm is denoted *AP* for Averaged Projections. The second projection algorithm is an unrelaxed implementation of Alg.(3.14) denoted by *SP* for Sequential Projections. In this implementation the pupil domain projection is computed at every second iterate. This is consistent with higher-end implementations which, with optimal parallelization, would compute the pupil projection more often because it is less computationally expensive than the image domain projections. The projection algorithms are compared to line search algorithms for the evenly weighted least squares measure E_ϵ (*LS*) and the extended least squares reduced objective R_ϵ (*ELS*). An additional comparison is made to an L-BFGS trust region algorithm applied to the reduced objective R_ϵ (*L-BFGS*). See Alg.5.2.2 and Remark 5.2.3. The value of the constants in R_ϵ are taken to be $c_m = 1$ for $m = 0, \ldots, 3$. For the limited memory implementation, a memory length of 4 was chosen.

The formulation of the projections in Eq.(3.12) is numerically unstable. There are several sources of this instability, the most elementary being the possibility of division by zero. In order to achieve a reasonable comparison of computational complexity to line search methods applied to E_ϵ or R_ϵ, the projections are calculated naively as prescribed by

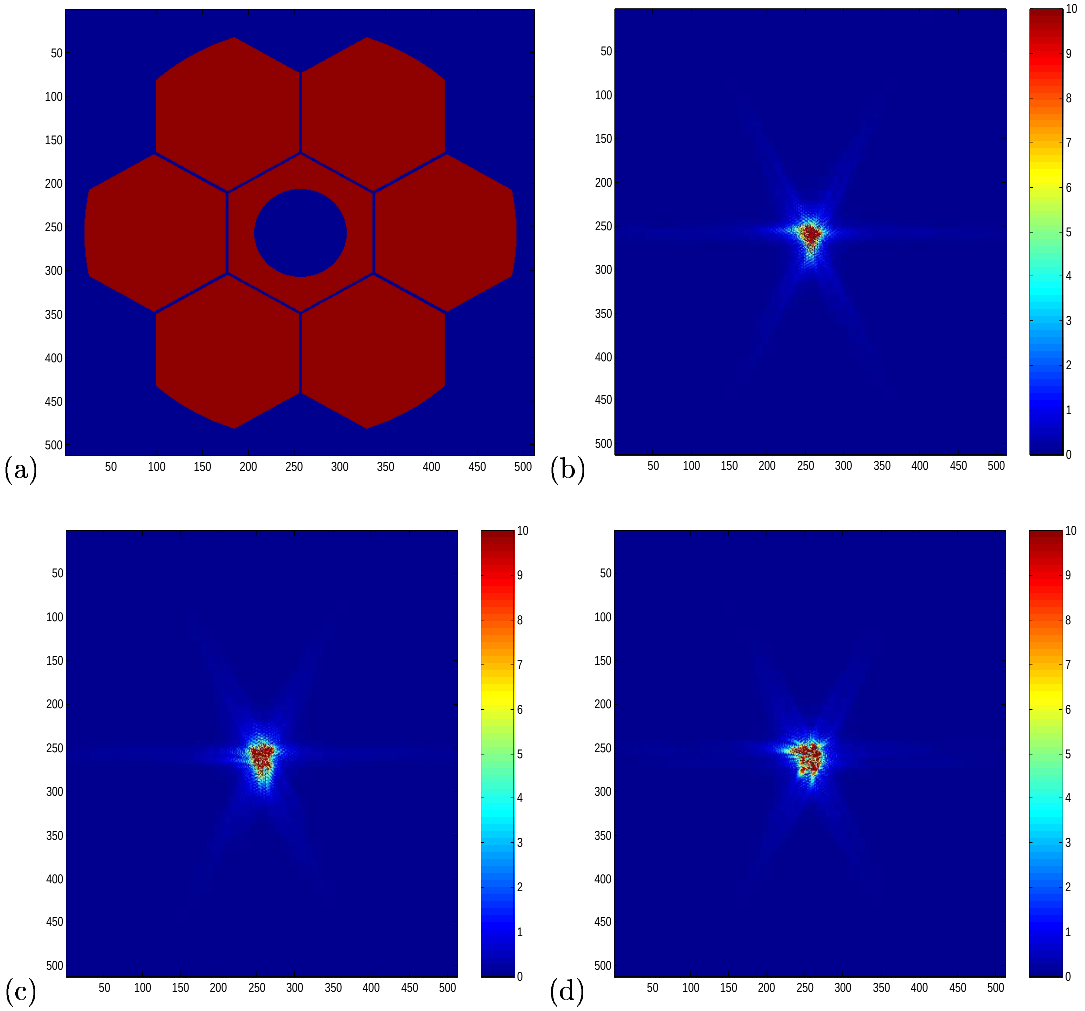


Figure 6.1: Aperture (a) and noiseless image data (b)-(d) for a segmented pupil on a 512 by 512 grid. The 3 diversity images are the optical system's response to a point source at focus, and plus/minus defocus respectively.

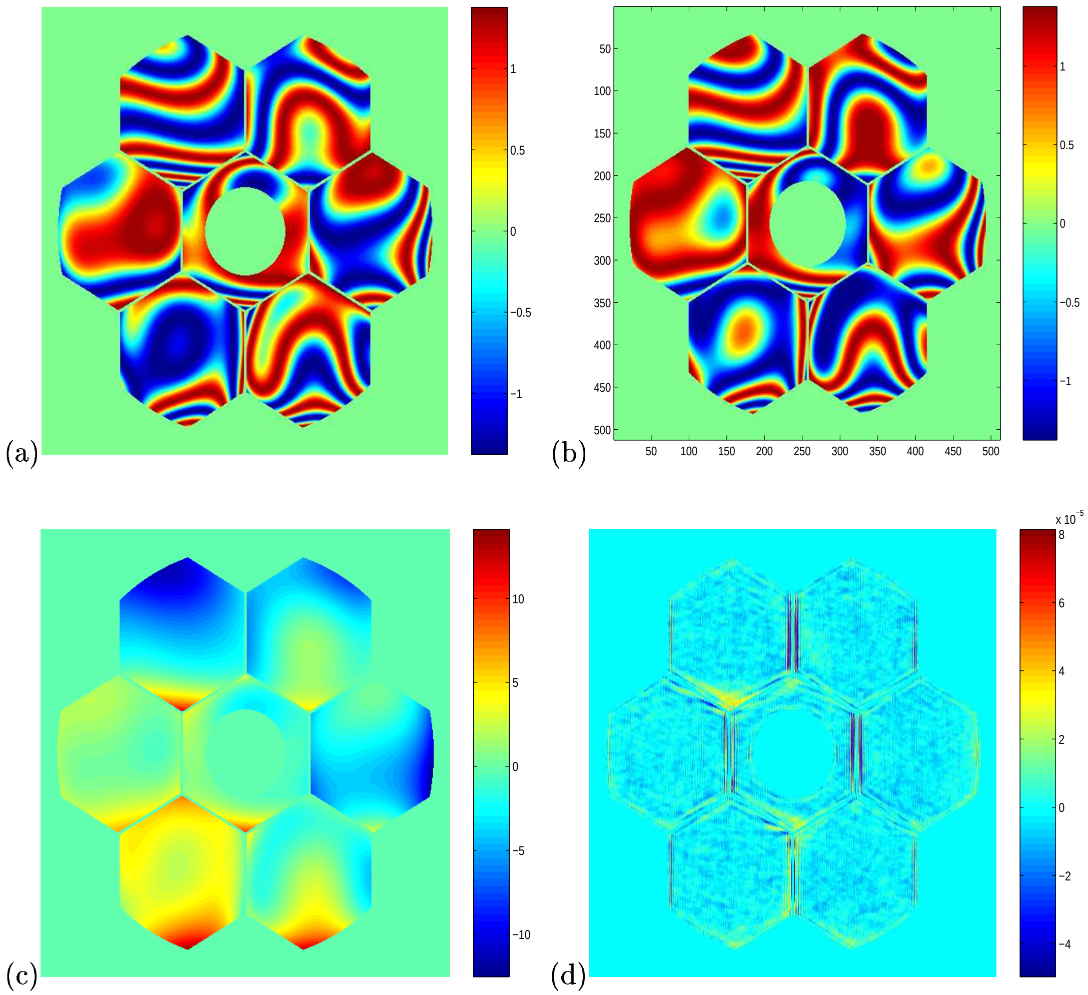


Figure 6.2: Real and imaginary parts, (a) and (b) respectively, of an aberrated wavefront for the segmented pupil recovered from 3 noiseless diversity point source images on a 512 by 512 grid. The wavefront phase is unwrapped (c) and compared to the true phase. The wavefront error (d) is in units of wavelength.

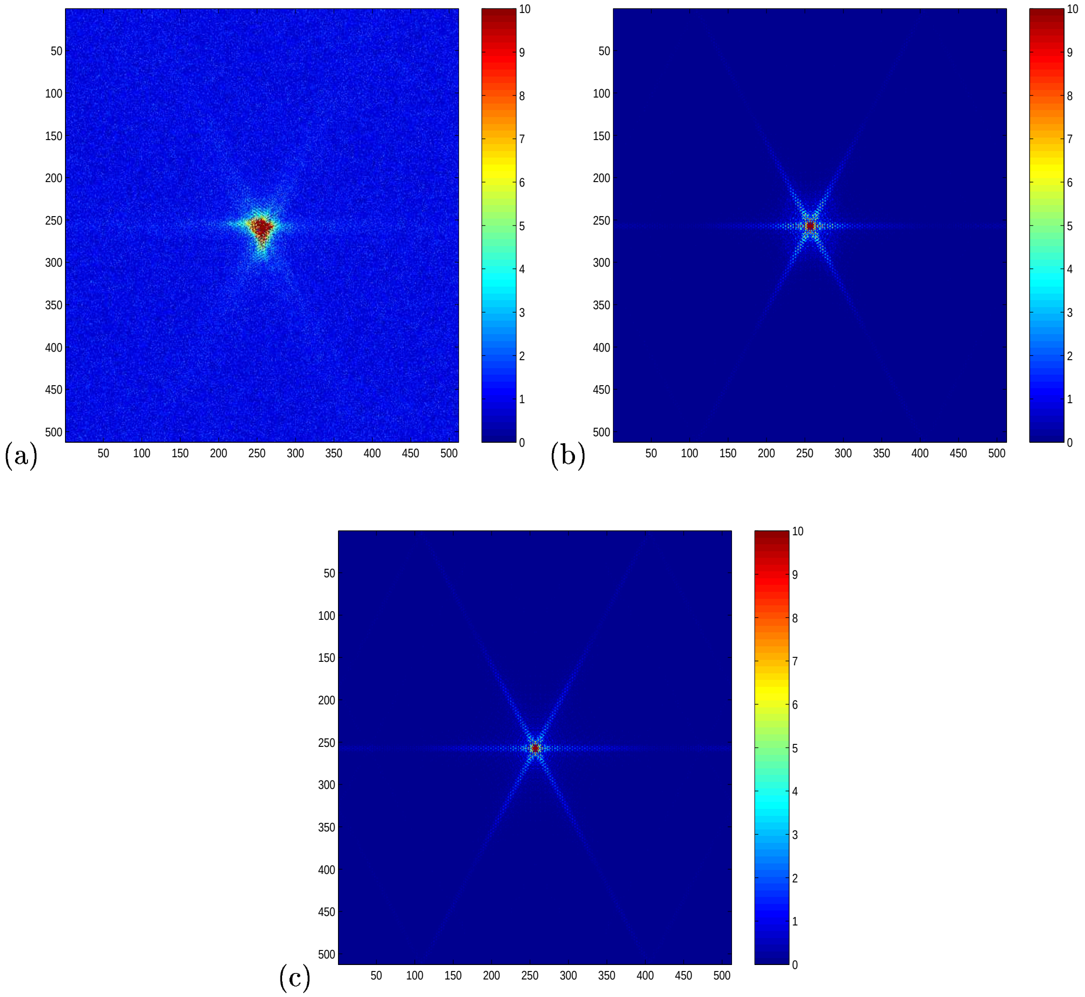


Figure 6.3: Noisy point-spread function (a) for a segmented pupil on a 512 by 512 grid. The recovered point-spread function (b) was first filtered with a Fourier window filter before processing by the wavefront reconstruction algorithm. Frame (c) shows the true, unaberrated point-spread function.

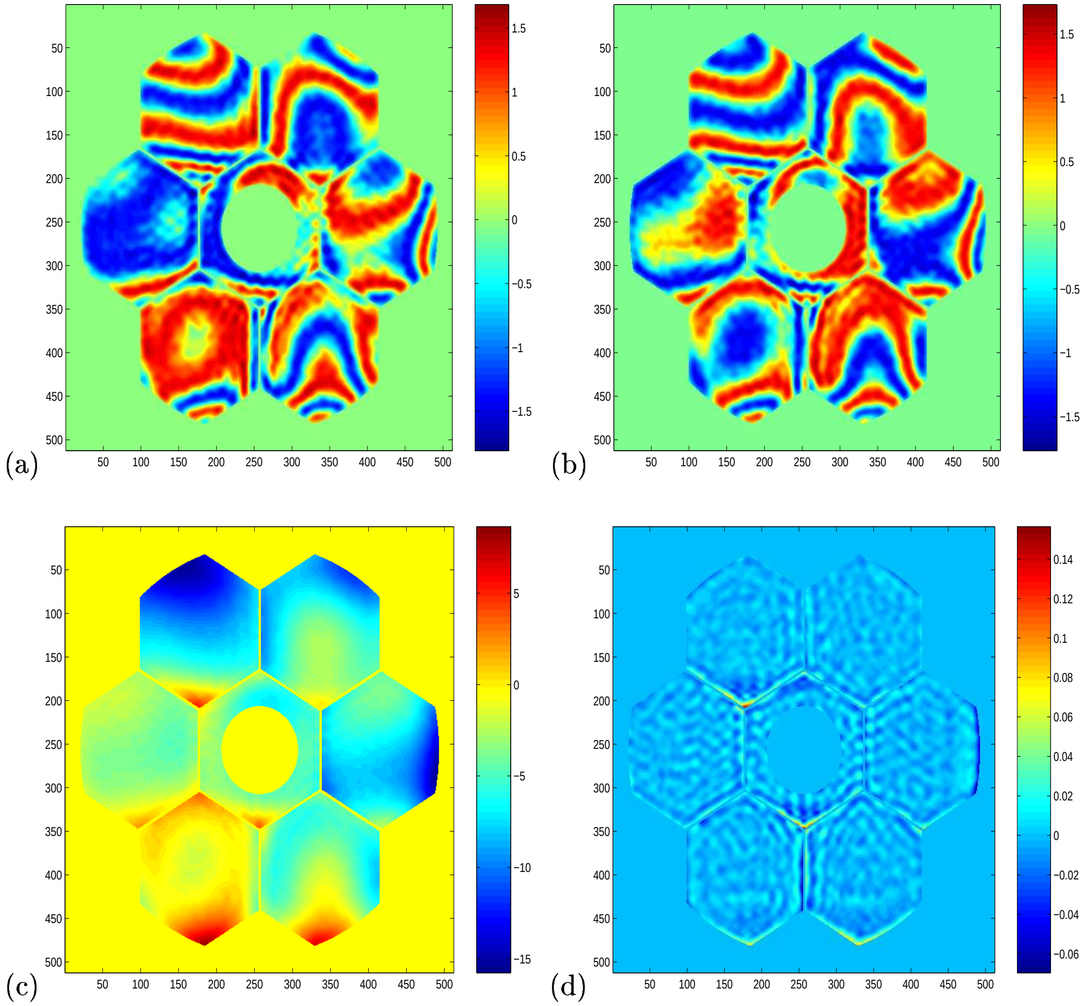


Figure 6.4: The real and imaginary parts, (a) and (b) respectively, of the aberrated wavefront for the segmented pupil recovered from 3 filtered noisy diversity point source images on a 512 by 512 grid. The wavefront phase (c) is unwrapped and compared to the true phase. The wavefront error (d) is in units of wavelength. The ridges in the wavefront error is due to Gibbs phenomenon associated with the noise filter.

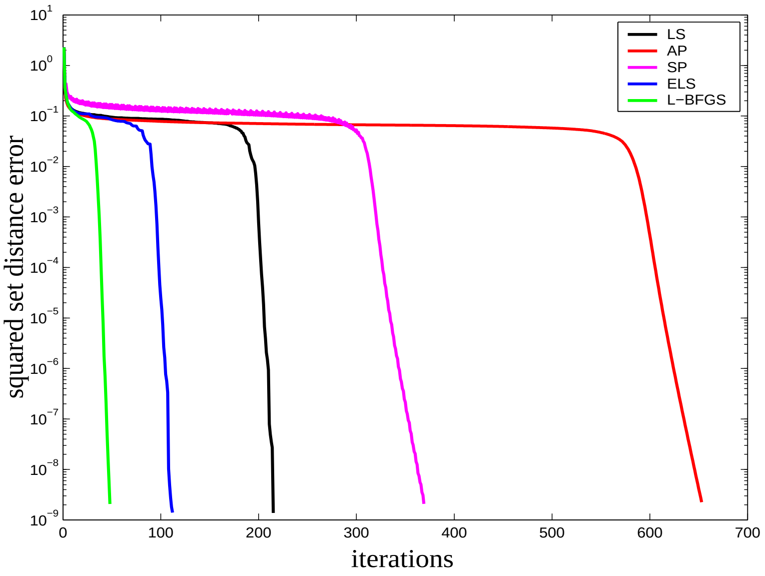


Figure 6.5: Comparison of algorithms applied to the test problem shown in Fig.(6.1). The algorithms have different objectives, so we compare the behavior of the squared set distance error defined by Eq.(3.31) for each.

Eq.(3.5). We observe that about 6% of the projection runs are exited due to divide by zero errors. A second source of instability arises when $\Pi_{\mathbb{Q}_m}[\boldsymbol{u}]$ is multi-valued. This is easily remedied by taking a selection $\pi[\boldsymbol{u}; \psi_m, \theta]$ given by Eq.(3.5). While it is unlikely that an iterate will be exactly zero, how one interprets machine zero in this context is an important consideration for numerical stability. In a neighborhood of zero corresponding to machine precision, the phase and amplitude of the estimated wavefront at a grid point $\boldsymbol{u}(\boldsymbol{x}_j)$ are not reliable. If at the same point the data $\psi_m(\boldsymbol{x}_j)$ is relatively large, then, even though the projection $\Pi_{\mathbb{Q}_m}[\boldsymbol{u}]$ is single-valued, the error will be amplified. About 6% of our trials with projection algorithms resulted in little or no progress from the initial guess. Since the norm of the gradient of a slightly perturbed E in these regions was found to be well away from zero, we attribute this outcome to the instability due to phase error amplification. Non-convergence due to divide by zero errors and possible phase error amplification were discounted from the averages computed in Tbl.(6.1). That is, approximately 12% of the runs for which the projection algorithm fails are not included in Tbl.(6.1). On the other hand, all of the runs for the analytic algorithms converge and are included in the table.

The behavior of the squared set distance error for a sample run for each of the algorithms is illustrated in Fig.(6.5). Each of the algorithms behaves qualitatively the same, as would be expected. Each spends the majority of time in a flat region where little progress is made, until a neighborhood of a solution is found and error reduction in all cases is rapid. In the flat region the gradient and curvature of the objective are very small. This region corresponds to what is described in projection methods as a 'tunnel'. The notoriously slow convergence of projection methods is easily understood in terms of the notoriously slow convergence of first-order methods. The limited memory implementation does much better in the flat region, though it too is slowed considerably.

The behavior of the algorithms varies considerably depending on the initialization, hence the average performance of the algorithms over 30 random initial guesses is tabulated in Tbl.(6.1). The initial guesses all have unit magnitude in the pupil domain with random phase uniformly distributed on $[0, 2\pi]$. In Tbl.(6.1) average cpu times, along with standard deviations of the experiments, are compared using the LS algorithm as a baseline - the results for the other algorithms are normalized by the LS performance given at the far left of the table. The standard deviations reflect the robustness of the algorithm and consistency of performance. With the exception of the SP algorithm, on average each algorithm requires the same number of function evaluations per iteration. The limiting calculation for this application is the Fourier transform which is accomplished with the FFT algorithm. Each squared set distance error evaluation requires one FFT per diversity image. Each gradient or projection calculation requires 2 FFTs per diversity image. The SP algorithm requires at most 3 fewer FFTs per iteration than the line search or AP algorithms since only one projection is calculated at each iteration. Hence the per iteration cost of the SP algorithm is .6 times that of the other algorithms. For L-BFGS and LS implementations, when the trust region is invoked or when backtracking is required to generate the proper step size additional function evaluations are needed. When the trust region is restricted, usually only one restriction is necessary when the scaling Eq.(5.30) is used. For backtracking, usually three backtracking steps are required. The added computational cost for implementing limited memory methods is not noticeable in cpu time. The average time per iteration for L-BFGS methods is 1.047 seconds for a 512×512 image using a parallel cluster of 16 processors, compared to 1.017 seconds for line search methods. There is, however, a considerable difference in the memory requirements depending on how many previous steps are stored.

The performance of the algorithms on apodized (*i.e.* filtered) noisy data shown in Fig.(6.3) is very similar in character to the noiseless experiments. Since the methods use a pixel basis, all of the algorithms attempt to match the data exactly, including the noise. Filtering for data analysis is treated as a separate issue from filtering for numerical efficiency or stability. While it has been noted that other noise models are more appropriate [140], the noise in these experiments is additive and normally distributed, consistent with the least squares performance measure. The squared set distance error $E = 0.050$ is the outer edge of the neighborhood of the solution, *i.e.* the "knee" in the error reduction shown in Fig.(6.5). Once inside this neighborhood, error reduction is rapid in all cases. With the exception of the SP algorithm, error reduction flattens out at $E = 0.0138$. In every trial the SP algorithm fails to reduce the error below $E = 0.02$. In practice, however, this difference between the SP "solution" and that of the other algorithms does not result in noticeable differences in the eyeball norm for the phase estimate.

Table 6.1: Relative cpu time of projection and analytic algorithms averaged over 30 random trials. The baseline is the LS algorithm. Outliers were not included in the totals for algorithms with an asterisk.

	No Noise			Noise					
	$E \leq 20e^{-9}$			$E \leq 0.05$			$E \leq 0.0138$		
	mean	low	high	mean	low	high	mean	low	high
LS	248	99	970	161	68	483	222	159	518
AP*	2.29	99	1680	2.7	126	1765	2.3	162	1808
SP*	.96	72	591	1.19	35	746	-	-	-
ELS	.66	74	365	.77	35	258	.84	76	304
L-BFGS	.29	41	196	.44	37	159	.47	72	182

6.2 Multi-resolution Techniques

In Fig.(6.6.a) a series of windowing operations is depicted for three 512×512 diversity images. First, the center 32×32 pixels of each diversity image are kept, and the remaining pixels are set to zero, that is for $m = 1, 2, 3$, we set

$$\widetilde{\psi_m} = \mathcal{X}_{32} \odot \psi_m.$$

The corresponding pupil domain operation is to smooth the pupil by convolution with the sinc function. This is achieved by setting

$$\widetilde{\psi_0} = \left[\mathcal{X}_{32} \odot \psi_0^{\wedge}\right]^{\vee}.$$

The resulting pupil domain constraint is depicted in Fig.(6.7.b). For $m = 1, 2, 3$, the dimension reduction of the images $\widetilde{\psi_m}$ is straight forward. One simply ignores the zero pixels outside of the window. In the pupil domain the reduction of dimension is achieved by assigning single values to blocks of 16×16 adjacent pixels. In our implementations the value that is assigned is the average of the 16^2 pixels. The corresponding wavefront reconstruction problem is $1/16$ the original problem size. The solution to Problem (3.98) corresponding to this resolution is depicted in Fig.(6.8).

The next step is to use the solution depicted in Fig.(6.8) as an initial guess for the next resolution, which in this example is 128×128 pixels. To do this, one simply divides the pixels of the low resolution solution into 16 sub-pixels. the image and pupil domain data are treated the same as with the 32×32 case. The solution to the 128×128 problem is then used as the initial guess for the full resolution problem. In Fig.(6.9) the squared set distance error versus iteration for a multi resolution implementation of the trust region L-BFGS algorithm is shown. Notice that the flat regions typical of these problems are encountered at low resolution. The higher resolution runs are started in a neighborhood

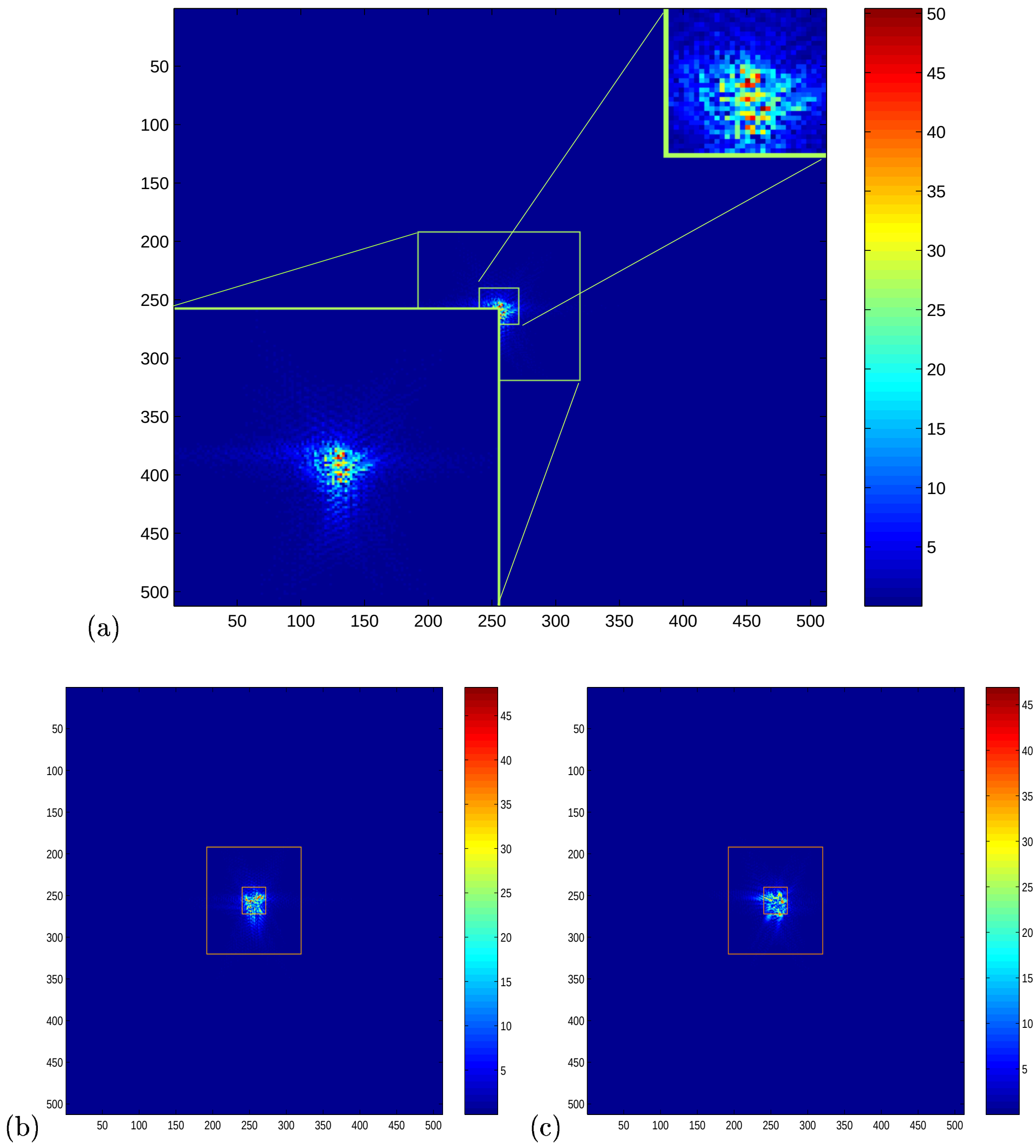


Figure 6.6: Multi-resolution image data. Three levels of multi-resolution windowing operations are depitced for each diversity image. Frame (a) shows close-ups of each of the three resolution levels. The center 32 × 32 pixels of each diversity image, together with the corresponding low-resolution pupil constraint Fig.(6.7.b) are used to generate the approximate solution shown in Fig.(6.8.a)-(6.8.b). This solution is used to initialize the same problem with the center 128 × 128 pixels and the correpsonding low-resolution pupil constraint (Fig.(6.7c)) as data. The solution to this problem, shown in Fig.(6.8.c)-(6.8.d), is used to initialize the full resolution problem. The progress of the set distance error versus iteration of the multi-resolution implementation is shown in Fig.(6.9).

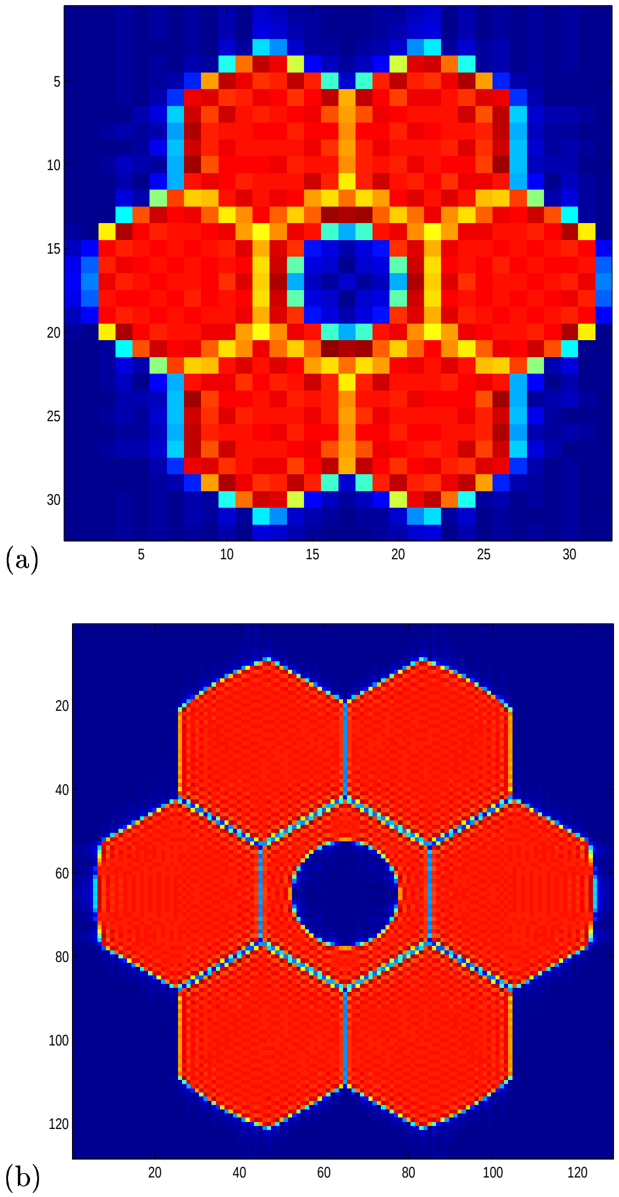
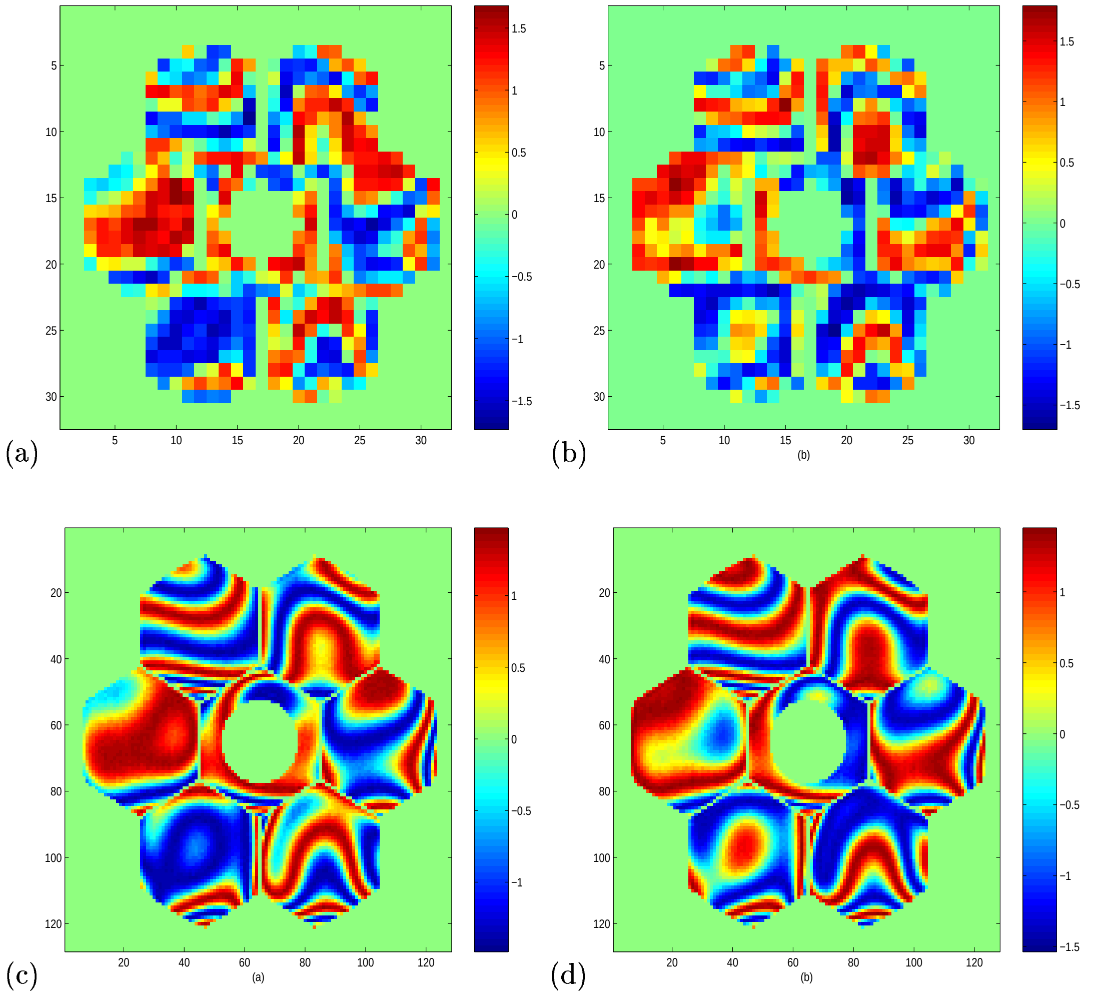


Figure 6.7: Multi-Resolution pupil domain constraints. The lowest resolution pupil constraint (a) corresponds to the 32 × 32 image data shown in Fig.(6.6). The medium resolution pupil constraint (c) corresponds to the 128 × 128 image data.



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

Figure 6.8: Aberrated wavefront for the segmented pupil recovered from 3 diversity point source images on successively finer grids. The real and imaginary parts of the low resolution wavefront, (a) and (b) respectively, is generated from a truncation of the image data to the center 32 by 32 pixels. This solution is used as a first guess for the next resolution, 128 by 128. The real and imaginary parts of the 128 by 128 pixel resolution, (c) and (d) respectively, is used as a first guess for the full resolution problem shown in Fig.(6.2).

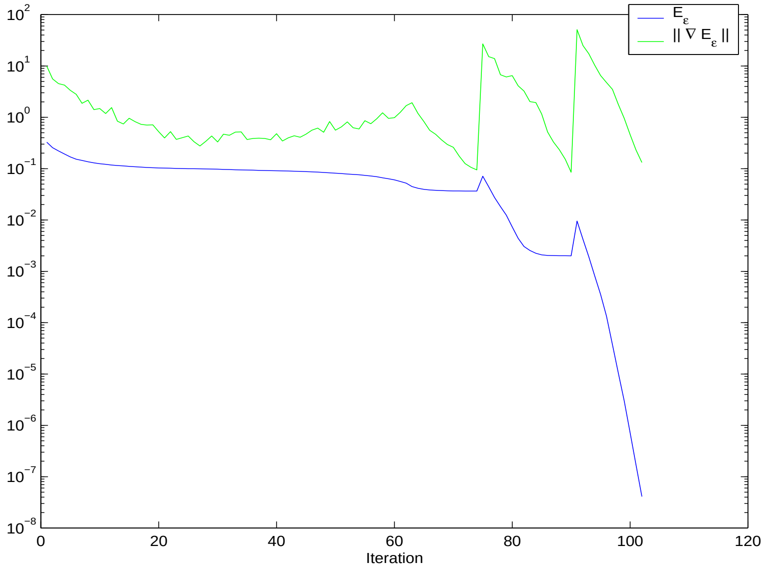


Figure 6.9: Squared set distance error and corresponding norm of the gradient versus iteration for a multi-resolution implementation of the trust region L-BFGS algorithm Alg.5.2.2. The flat region of the iterations is handled at low resolutions. Only when the estimate is in the neighborhood of a solution does the algorithm switch to higher resolution calculations.

of the solution and very few iterations are required for convergence. All of the hard work is accomplished cheaply at low resolutions. Starting from an initial phase guess of zero, in Matlab the multi-resolution implementations resulted in a factor of 17 speed up in cpu time over the full resolution run.

6.3 Smooth Objectives

Figure 6.10 shows the performance of the modulus squared objective E_2 compared to the modulus objective E_ϵ at different resolutions. Note that the two algorithms behave qualitatively the same at low resolutions, Fig.(6.10.a)-(6.10.b), while at higher resolutions, Fig.(6.10.c)-(6.10.d), the algorithms behave dramatically differently. The slowed convergence at the end of the high-resolution tests for E_2 is due to the ill-conditioning associated with the objective. The ill-conditioning is not seen at lower resolutions since the truncation of the Fourier coefficients involved in generating the lower resolution solution amounts to a truncation of the small singular-values that lead to ill-conditioning. Reducing the resolution of the data essentially regularizes the inverse problem through truncation.

6.4 Simultaneous Deconvolution and Wavefront Reconstruction

The ill-conditioning associated with the kernel of the integral operator Eq.(4.3) is a serious limitation for the simultaneous deconvolution and wavefront reconstruction problem. However, at low resolutions with moderate wavefront aberrations on the order of one wavelength, the technique is stable and efficient. Approximately 1000 iterations were required to reconstruct the image shown in Fig.(6.12) from the data shown in Figure 6.11 shows the simultaneous reconstruction and deconvolution of a noiseless image of a simulated “spiral galaxy”.

The method is robust in the presence of noise. Figure 6.13 shows the data from images corrupted by severe noise. In this example the root mean squared signal to noise ratio for the defocused images is approximately 2.6. For the in-focus image the root mean squared signal to noise ratio is approximately 13. The recovered object shown in Fig.(6.14.a) underestimates the peak values of the true image. Our regularization parameter α in the standard Tikhonov regularized object estimate Eq.(4.10) was chosen to be the average of the spectral density of the noise given by Eq.(4.24). The value used in our experiments was $\mathcal{S}_{\eta\eta}(\omega) \approx \alpha = 2$ for all frequencies ω. It is clear from this that such a severe regularization will result in an underestimate of the peak values of the object. This is the smoothing effect of the Wiener filter. The recovered wavefront shown in Fig.(6.14.c) differs from the true by a tilt, that is, a planar phase error shown in Fig(6.14.e).

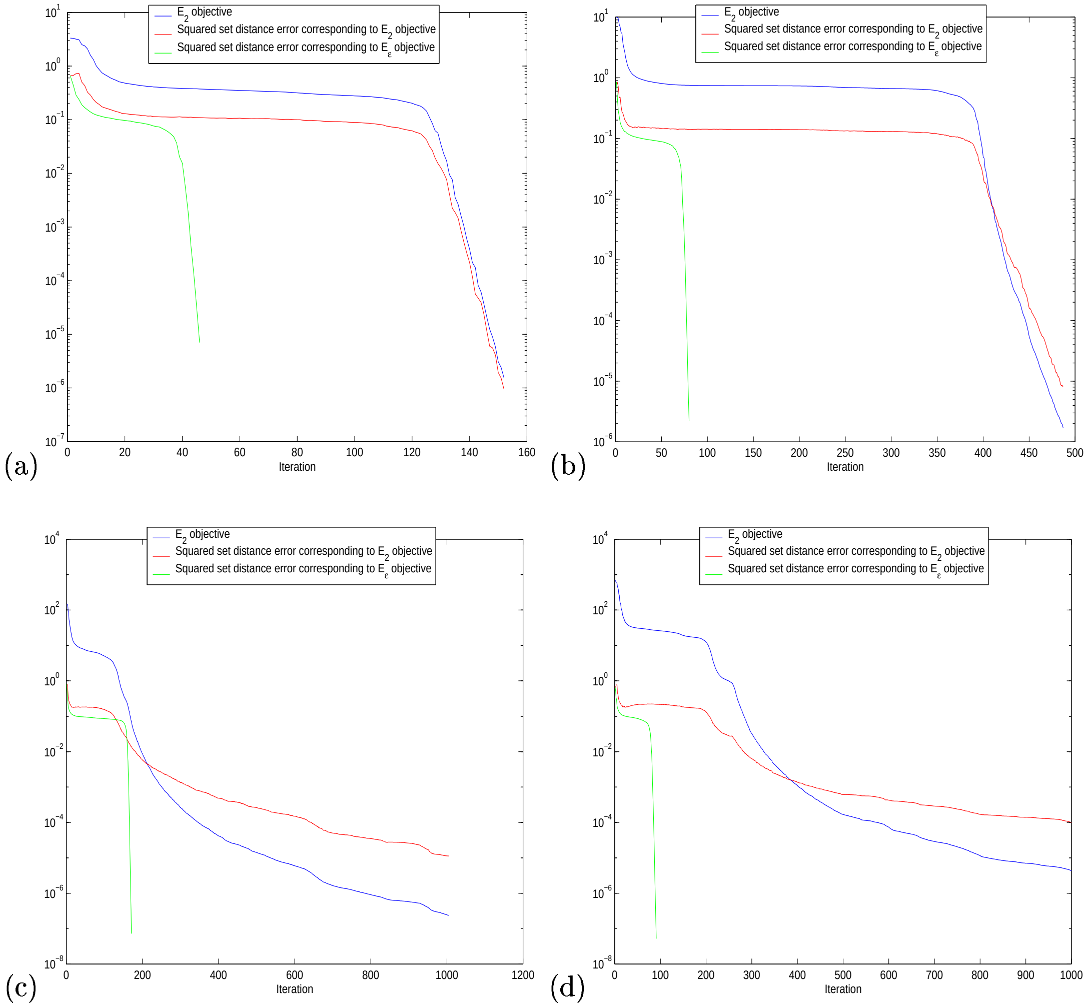


Figure 6.10: Comparison of performance of a numerical optimization algorithm for solving Pr.(2.80) with $\varphi = \delta$ and the objective E_2 defined by Eq.(4.3) versus the objective E_ϵ defined by Eq.(3.87). The behavior of the squared set distance error E is calculated and plotted for the algorithm with both objectives for comparison. Frames (a)-(d) are at 32×32, 64×64, 256×256, and 512×512 resolution respectively.

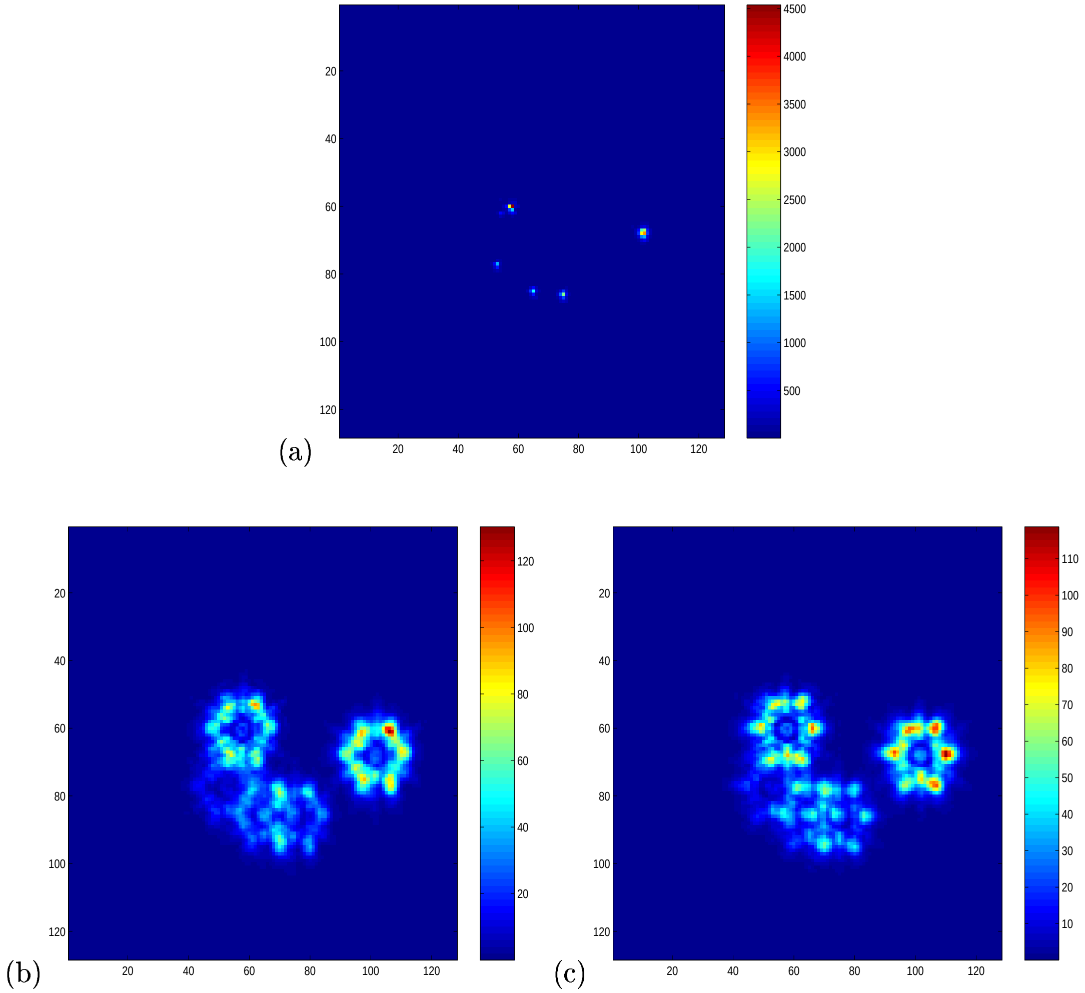


Figure 6.11: Noiseless image data. Three images of the same object taken at 3 different focus settings.

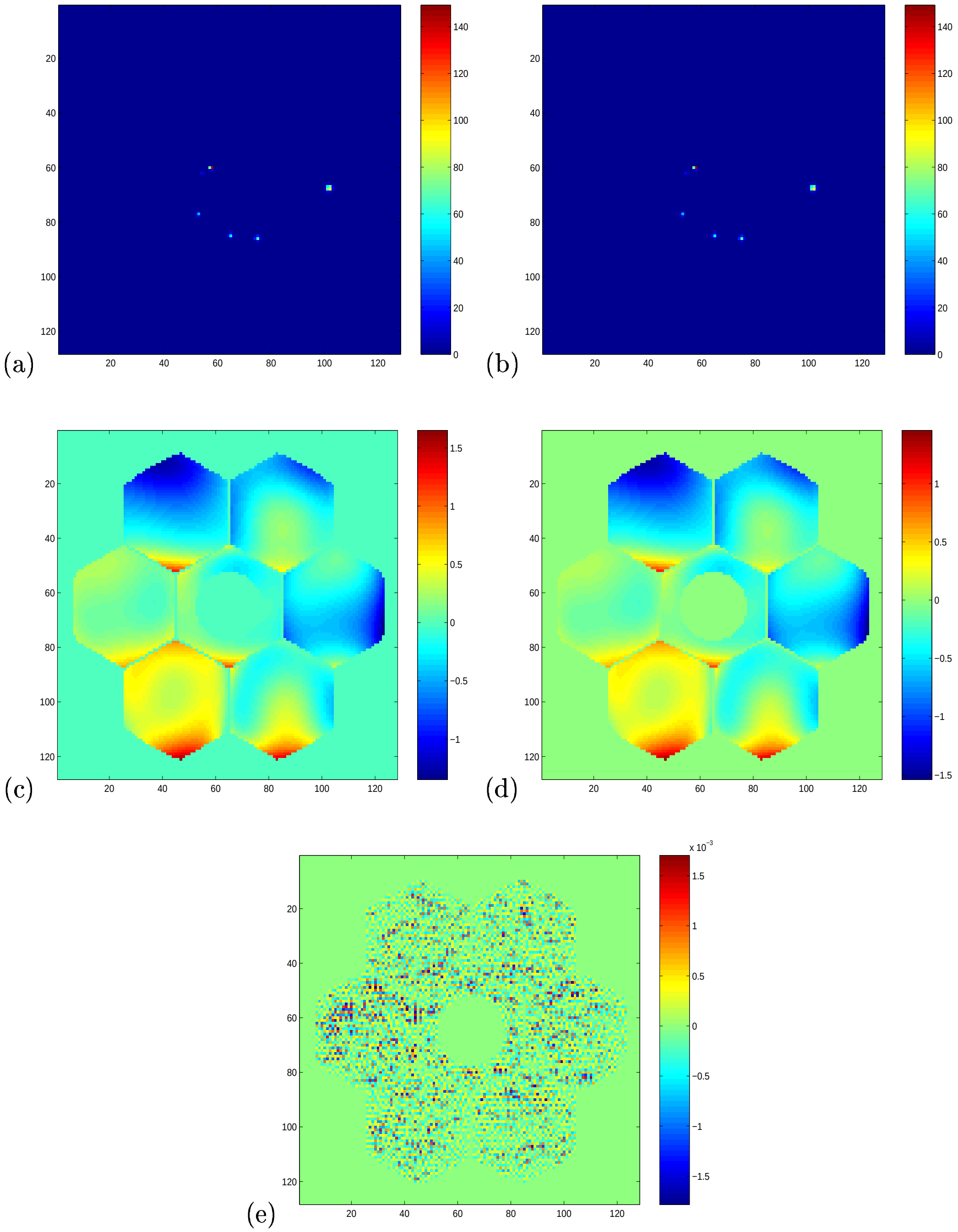


Figure 6.12: Recovered object (a) and wavefront (c) from noiseless data shown in Fig.(6.11). The true object is shown in frame (b). The true phase (d) is compared to the recovered phase. The error (e) is in units of wavelength.

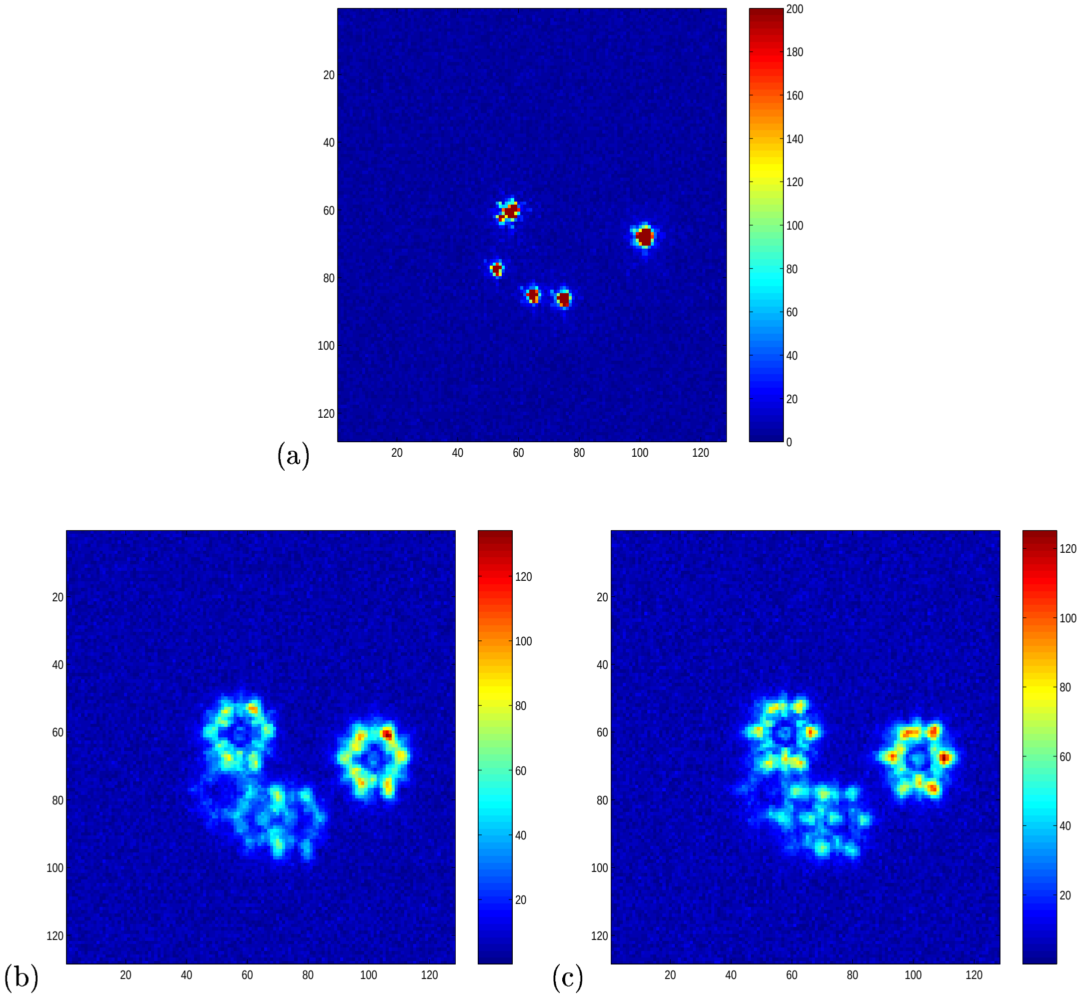


Figure 6.13: Noisy image data. Three images of the same object taken at 3 different focus settings.

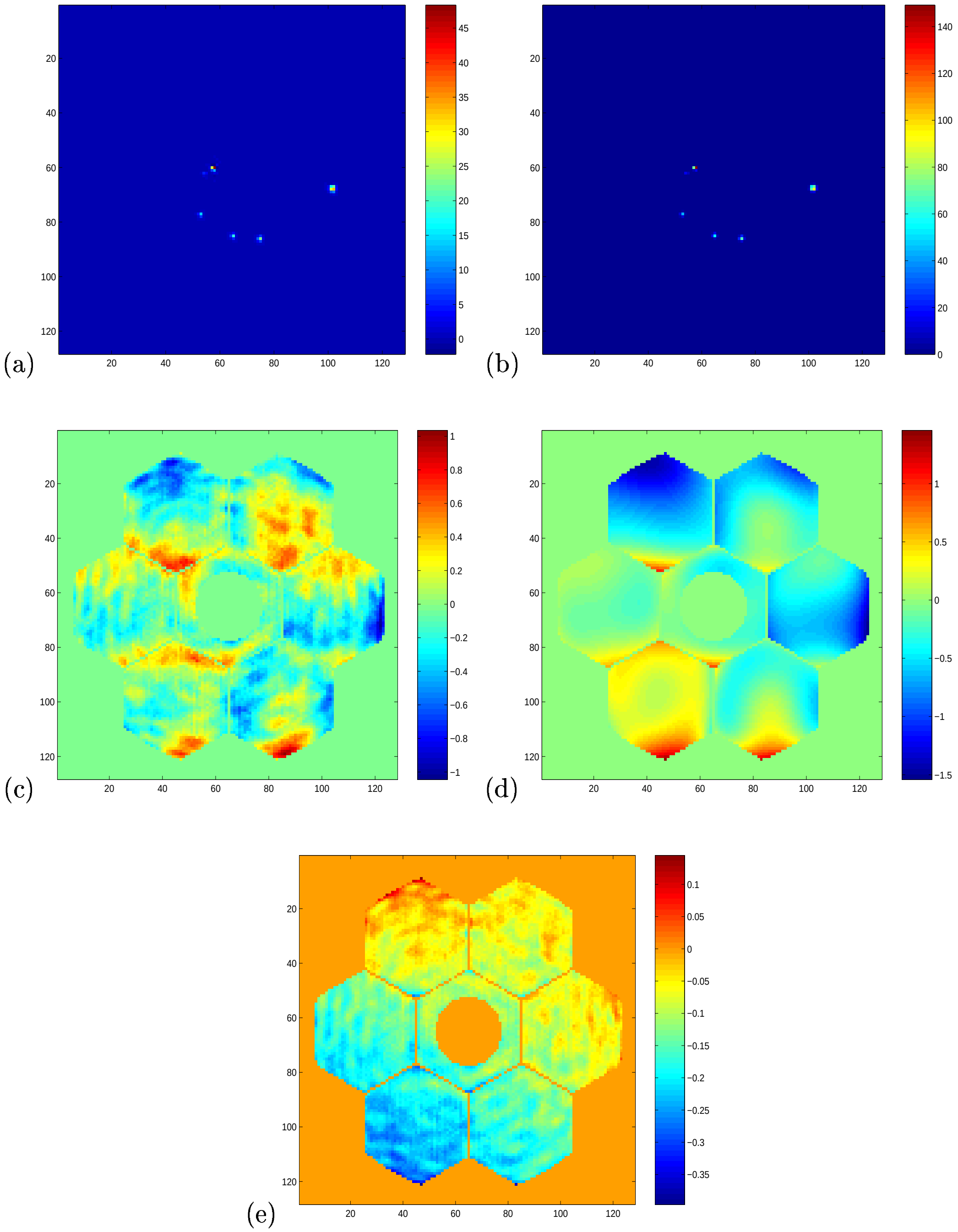


Figure 6.14: Recovered object (a) and wavefront (c) from noiseless data shown in Fig.(6.13). The true object is shown in frame (b). The true phase (d) is compared to the recovered phase. The error (e) is in units of wavelength.

Chapter 7

CONCLUSION

In this dissertation we have established a framework for the numerical theory of optical wavefront reconstruction and deconvolution. We briefly summarize our results in this final chapter.

7.1 Summary

In Chapter 2 of this thesis we provided a derivation of the optical phase retrieval problem from first principles. In Chapter 3 we presented a detailed study of the fundamentals of wavefront reconstruction. Many fundamental questions regarding convergence of projection algorithms remain. When the intersection of of even convex constraints is empty, convergence is an open question. This is often the case in image processing with noisy data. When algorithms stagnate it is impossible to tell if the method has found a local solution, or is stuck in what is often referred to as a tunnel. We noted that extensions to projection algorithms have been proposed to overcome stagnation [68]. These methods seem to be very robust and efficient in practice [177]. Their success warrants precise mathematical analysis which has yet to be done. In Section 3.2 of Chapter 3 we reviewed analytic perturbation approaches to the problem and quantified their relationship to geometric methods. Two performance measures were considered, their associated optimization problems were formulated in Pr.(3.33) and Pr.(3.98). The first measure is a perturbed weighted least squares measure. The second is a new approach which we call extended least squares. This objective allows us to adaptively correct for the relative variability in the diversity measurements, ψ_m. In Chapter 5 we reviewed two numerical methods. The first was a standard line search algorithm for which convergence to first-order necessary conditions for optimality was proven for the perturbed least squares and extended least squares objectives. The line search method is accelerated by a limited memory approach which allows us to efficiently approximate curvature information in large problems. The use of limited memory techniques for phase retrieval and deconvolution has appeared in recent work [115, 186]. The method is made robust with a novel use of explicit trust regions. The trust region strategy also allows for precise scaling of the step size, thus avoiding costly function evaluations that are common to more trial-and-error-type methods such as implicit trust regions and backtracking. The resulting algorithm was given in Alg.(5.2.2). In Chapter 6 we compared the performance of the different approaches on noiseless and noisy data. The results indicate that while certain implementations of iterative transform algorithms can be competitive (see the SP algorithm), their performance varies more from one example to the next than the algorithms based on analytic techniques. Other implementations of iterative transform algorithm such as averaged projections (AP) are clearly not competitive approaches. Limited memory and trust region techniques reduce the variability of performance without adding significant computational cost.

Further cpu speed up is possible with the introduction of multi-resolution techniques as discussed in [115, 135]. These are similar to windowing techniques used for noise filtering. In tests with Matlab we have achieved 17 fold speed up in time to convergence with the use of these techniques. With optimal parallelization and multi-resolution techniques we expect that the per iteration cpu time for a cluster of 16 PC's with three 512×512 diversity images could be brought down, conservatively, to a tenth of a second.

7.2 Extensions and Future Work

The extended least squares approach presented in Section 3.3.2 of Chapter 3 has great potential for future research. In our implementations we chose the simplest possible regularizing functional in Eq.(3.93), that is $G_m[\boldsymbol{u}] = const.$ Even this simple choice had a dramatic effect on the performance of the algorithms. This opens the door to a search for an optimal $G_m[\boldsymbol{u}]$. There are two different ways to interpret $G_m[\boldsymbol{u}]$, the first and perhaps most natural is statistical, the second is purely algorithmic. Under the statistical interpretation, $G_m[\boldsymbol{u}]$ is viewed as the variance or spatial correlation of the data sets. The method is very general and applies to a wide variety of observations and statistical models. Under the algorithmic interpretation, $G_m[\boldsymbol{u}]$ is a regularizing term in a penalty function and can be used to tackle the problem of algorithm stagnation in the middle iterations (see Fig.(6.5)). The adaptive weighting strategy allows one to include several different metrics in the same objective, one that is more effective for the middle regions and one that is more effective near a local solution.

Other directions for research include partial function evaluation algorithms similar to the sequential projection algorithms discussed in Section 3.1 of Chapter 3. The trust region methodology reviewed in Chapter 5 is a first step to stably implementing this strategy. Regularization techniques are also central to numerical methods for solving the more general problem of simultaneous wavefront reconstruction and deconvolution. The theory developed here is intended as a starting point for numerical solutions to both the phase retrieval problem and the more general phase diversity problem.

BIBLIOGRAPHY

[1] R. A. Adams. *Sobolev Spaces.* Academic Press, 1975.

[2] E. J. Akutowicz. On the determination of the phase of a Fourier integral, I. *Trans. Amer. Math. Soc.*, 83:179–192, 1956.

[3] E. J. Akutowicz. On the determination of the phase of a Fourier integral, II. *Proc. Amer. Math. Soc.*, 8:234–238, 1957.

[4] B. E. Allman, P. J. McMahon, K. A. Nugent, D. Paganin, D.L. Jacobson, M. Arif, and S. A. Werner. Imaging - phase radiography with neutrons. *Nature*, 408:158–159, 2000.

[5] R. S. Anderssen and P. Bloomfield. Numerical differentiation procedures for non-exact data. *Numer. Math.*, 22:157–182, 1974.

[6] R. S. Anderssen, F. R. de Hoog, and M. A. Lukas. *Application and numerical solution of integral equations: Proceedings of a Seminar held at the Australian National University, Canberra, November 29, 1978*, volume 6 of *Monographs and Textbooks on Menics of Solids and Fluids: Mechanics and Analysis.* Martinus Nijhoff, The Hague, 1980.

[7] R. S. Anderssen and P. M. Prenter. A formal comparison of methods proposed for the numerical solution of first kind integral equations. *J. Austral. Math. Soc. B*, 22:488–500, 1981.

[8] R. J. Aumann. Integrals of set-valued functions. *J. Math. Anal. Appl.*, 12:1–12, 1965.

[9] R. J. Aumann. Measurable utility and measurable choice theorems. In *La Décision*, pages 15–26, Paris, 1969. Colloque Internationaux du C.N.R.S.

[10] R. J. Aumann. An elementary proof that integration preserves uppersemicontinuity. *J. Math. Econ.*, 3:15–18, 1976.

[11] O. Axelsson and G. Lindskog. On the rate of convergence of the preconditioned conjugate gradient algorithm. *Numer. Math.*, 48:499–523, 1986.

[12] G. Ayers and J. Dainty. An iterative blind deconvolution algorithm and its applications. *Opt.Lett.*, 13:547–549, 1988.

[13] N. Baba and K Mutoh. Measurement of telescope aberrations through atmospheric turbulence by use of phase diversity. *Appl. Opt.*, 40(4):544–552, Feb 2001.

[14] R. Barakat and G. Newsam. Algorithms for reconstruction of partially known, band-limited Fourier-transform pairs from noisy data. *J. Opt.Soc. Am. A*, 2(11):2027–2038, 1985.

[15] R. Barakat and G. Newsam. Algorithms for reconstruction of partially known, band-limited Fourier-transform pairs from noisy data. II. the nonlinear problem of phase retrieval. *J. Int. Eq.*, 9(Suppl.):77–125, 1985.

[16] T. K. Barrett and D. G. Sandler. Artificial neural network for the determination of Hubble Space Telescope aberration from stellar images. *Appl. Opt.*, 32:1720–1727, 1993.

[17] M. J. Bastiaans. The Wigner distribution function and Hamilton's characteristics of a geometric-optical system. *Optics Communications*, 30(3):321–326, 1979.

[18] M. J. Bastiaans. Application of the Wigner distribution function to partially coherent light. *J.Opt.Soc.Am.A*, 3(8):1227–1238, 1986.

[19] H. H. Bauschke. The composition of finitely many projections onto closed convex sets in Hilbert space is asymptotically regular. *Proc. Amer. Math. Soc.* to appear.

[20] H. H. Bauschke and J. M. Borwein. On the convergence of von Neumann's alternating projection algorithm for two sets. *Set-Valued Anal.*, 1(2):185–212, 1993.

[21] H. H. Bauschke and J. M. Borwein. On projection algorithms for solving convex feasibility problems. *SIAM Review*, 38(3):367–426, 1996.

[22] H. H. Bauschke, J. M. Borwein, and A. S. Lewis. The method of cyclic projections for closed convex sets in Hilbert space. In *Recent developments in optimization theory and nonlinear analysis (Jerusalem, 1995)*, pages 1–38. Amer. Math. Soc., Providence, RI, 1997.

[23] B. Bell, J. V. Burke, and A. Shumitzky. A relative weighting method for estimating parameters and variances in multiple data sets. *Comp. Stat. Data Anal.*, 22:119–135, 1996.

[24] J. F. Benders. Partitioning proceedures fro solving mixed variables programming problems. *Numerische Mathematik*, 4:238–252, 1962.

[25] A. Bhatia and E. Wolf. On the circle polynomials of zernike and related orthogonal sets. *Proc. Camb. Phil. Soc.*, 50:40–48, 1954.

[26] M. Born and E. Wolf. *Principles of Optics.* Pergamon Press, New York, 6th edition, 1980.

[27] J. M. Borwein. Epi-Lipschitz-like sets in Banach spaces: theorems and examples. *Nonlinear Anal.*, 11:1207–1217, 1987.

[28] J.M. Borwein and H. M. Strojwas. Tangential approximations. *Nonlinear Anal.*, 9:1347–1366, 1985.

[29] R. H. Boucher. Convergence of algorithms for phase retrieval from two intensity measurements. volume 231 of *Proc. SPIE*, pages 130–141, 1980.

[30] L. M. Brègman. The method of successive projection for finding a common point of convex sets. *Soviet Mathematics - Doklady*, 6:688–692, 1965.

[31] Y. M. Bruck and L. G. Sodin. On the ambiguity of the image reconstruction problem. *Opt. Comm.*, 30(3):304–308, Sept. 1979.

[32] J. V. Burke. Sherman-Morrison-Woodbury formula for powers of the inverse. Submitted to the SIAM Journal on Optimization, July 1996.

[33] J. V. Burke and A. Wiegmann. Low-dimensional quasi-Newton updating strategies for large-scale unconstrained optimization. Submitted to the SIAM Journal on Optimization, July 1996.

[34] C. J. Burrows. Hubble Space Telescope optics status. volume 1567 of *Proc.SPIE*, pages 284–293, 1991.

[35] C. J. Burrows, J. A. Holtzman, S. M. Faber, P. Y. Beley, H. Hasan, C. R. Lynds, and D. Schroeder. The imaging performance of the Hubble Space Telescope. *Astrophys, J.*, 369(2):L21–L25, 1991.

[36] R. H. Byrd, J. Nocedal, and R. B. Schnabel. Representations of quasi-newton matrices and their use in limited memory methods. *Math. Prog.*, 63:129–156, 1994.

[37] R. Carreras, S. Restaino, G. Love, G. Tarr, and J. Fender. Phase diversity experimental results: deconvolution of ν Scorpii. *Opt.Comm.*, 130:13–19, Sept. 1996.

[38] C. Castaing. *Sur les Multiapplications Measurables.* PhD thesis, Caën, 1967.

[39] J. N. Cederquist, J. R. Fienup, C. C. Wackerman, S. R. Robinson, and D. Kryskowski. Wave-front phase estimation from Fourier intensity measurements. *J. Opt. Soc. Amer. A*, 6(7):1020–1026, 1989.

[40] Y. Censor. Iterative methods for the convex feasibility problem. In M. Rosenfeild and J. Zaks, editors, *Convexity and Graph Theory*, pages 83–91. North-Holland, 1984.

[41] Y. Censor and A. Lent. Cyclic subgradient projections. *Math. Prog.*, 24:233–235, 1982.

[42] R. H. Chan, J. G. Nagy, and R. J. Plemmons. FFT-based preconditioners for toeplitz-block least squares problems. *SIAM J. Numer. Anal.*, 30(6):1740–1768, 1993.

[43] S. Chrétien and P. Bondon. Cyclic projection methods on a class of nonconvex sets. *Numer. Funct. Anal. and Optimiz.*, 17(1-2):37–56, 1996.

[44] F. H. Clarke. *Optimization and Nonsmooth Analysis*, volume 5 of *Classics in Applied Mathematics*. SIAM, 1990.

[45] F. H. Clarke, R. J. Stern, Yu. S. Ledyaev, and P. R. Wolenski. *Nonsmooth Analysis and Control Theory*. Springer Verlag, 1998.

[46] P. L. Combettes. Inconsistent signal feasibility problems: Least-squares solutions in a product space. *IEEE Trans. Signal Processing*, 42(11):2955–2966, 1994.

[47] P. L. Combettes. The convex feasibility problem in image recovery. In P. W. Hawkes, editor, *Advances in Imaging and Electron Physics*, volume 95, pages 155–270. Academic Press, 1996.

[48] P. L. Combettes. Hilbertian convex feasibility problem: convergence of projection methods. *Appl. Math. Optim.*, 35:311–330, 1997.

[49] P. L. Combettes and P. Bondon. Hard-constrained inconsistent signal feasibility problems. *IEEE Trans. Signal Processing*, 47(9):2460–2468, 1999.

[50] P. L. Combettes and H. J. Trussell. Method of successive projections for finding a common point of sets in metric spaces. *J. Opt. Theory App.*, 67(3):487–507, 1990.

[51] R. Courant and D. Hilbert. *Methods of Mathematical Physics*. Interscience Publishers, New York, 1953.

[52] J. C. Dainty and J. R. Fienup. Phase retrieval and image reconstruction for astronomy. In H. Stark, editor, *Image Recovery: Theory and Application*. Academic Press, 1987.

[53] M. M. Day. *Normed linear spaces*. Springer-Verlag, New York, 3rd edition, 73.

[54] G. Debreu. Integration of correspondences. In L. LeCam, J. Neyman, and E. L. Scott, editors, *Proc. Fifth Berkeley Symposium Math. Stat. Probability*, number 1 in II, pages 351–372, Berkeley, 1967. University of California Press.

[55] J. E. Dennis and R. Schnabel. *Numerical Methods for Unconstrained Optimization and Nonlinear Equations.* Prentice Hall, 1983.

[56] D. Dialetis and E. Wolf. The phase retrieval problem of coherence theory as a stability problem. *Nuovo Cimento (X)*, 47:113–116, 1967.

[57] D. C. Dobson. Phase reconstruction via nonlinear least squares. *Inverse Problems*, 8:541–558, 1992.

[58] B. Z. Dong, Y. Zhang, B. Y. Gu, and G.Z Yang. Numerical investigation of phase retrieval in a fractional Fourier transform. *J.Opt.Soc.Am.A*, 14(10):2709–2714, Oct 1997.

[59] A. J. J. Drenth, A. Huiser, and H. Ferwerda. The problem of phase retrieval in light and electron microscopy of strong objects. *Optica Acta*, 22:615–628, 1975.

[60] N. V. Efimov and S. B. Stechkin. Approximative compactness and chebyshev sets. *Soviet Math - Doklady*, 2:1226–1228, 1961.

[61] P. P. B. Eggermont. Maximum entropy regularization for fredholm integral equations of the first kind. *SIAM J. Math. Anal.*, 24(6):1557–1576, 1983.

[62] B. Ellerbroek and T. Rhoadarmer. Optimizing the performance of closed-loop adaptive optics control systems on the basis of experimentally measured performance data. *J.Opt.Soc.Am.A.*, 14(8):1975–1986, 1997.

[63] M. W. Farn. New iterative algorithm for the design of phase-only gratings. In *Proceedings of the SPIE*, volume 1555 of *Proc. SPIE*, pages 34–42, 1991.

[64] J. Fienup. Phase retrieval for Hubble Space Telescope using iterative propagation algorithms. In A. Tescher, editor, *Applications of Digital Image Processing XIV*, volume 1567 of *Proc.SPIE*, pages 327–332, 1991.

[65] J. R. Fienup. Reconstruction of an object from the modulus of its fourier transform. *Opt. Lett.*, 3(1):27–29, 1978.

[66] J. R. Fienup. Space object imaging through the turbulent atmosphere. *Opt. Eng.*, 18(5):529–534, 1979.

[67] J. R. Fienup. Iterative method applied to image reconstruction and to computer-generated holograms. *Opt. Eng.*, 19(3):297–305, 1980.

[68] J. R. Fienup. Phase retrieval algorithms: a comparison. *Appl.Opt.*, 21(15):2758–2769, 1982.

[69] J. R. Fienup. Phase-retrieval algorithms for a complicated optical system. *Appl. Opt.*, 32(10):1737–1746, 1993.

[70] J. R. Fienup, J. C. Marron, T. J. Schultz, and J. H. Seldin. Hubble Space Telescope characterized by using phase retrieval algorithms. *Appl. Opt.*, 32(10):1747–1767, 1993.

[71] J. R. Fienup and C. C. Wackerman. Phase retrieval stagnation problems and solutions. *J. Opt. Soc. Amer. A*, 3:1897–1907, 1986.

[72] J. Frank, P. Penczek, R. K. Agrawal, R. A. Grassucci, and A. B. Heagle. *Methods in Enzymology*, chapter 18. Three-dimensional cryoelectron microscopy of ribosomes, pages 276–291. Academic Press, San Diego, 2000.

[73] R. W. Gerchberg and W. O. Saxton. A practical algorithm for the determination of phase from image and diffraction plane pictures. *Optik*, 35:237–246, 1972.

[74] R. A. Gonsalves. Phase retrieval from modulus data. *J. Opt. Soc. Am.*, 66(9):961–964, Sept. 1976.

[75] R. A. Gonsalves. Phase retrieval and diversity in adaptive optics. *Opt.Eng*, 21(5):829–832, 1982.

[76] J. W. Goodman. *Introduction to Fourier Optics.* McGraw-Hill, 1968.

[77] J. W. Goodman. *Statistical Optics.* Wiley, 1985.

[78] L. Gubin, B. Polyak, and E. Raik. Method of projections for finding the common point of convex sets. *USSR Comput. Math and Math Phys.*, 7:1–24, 1967.

[79] 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, Aug 1996.

[80] T. E. Gureyev, A. Roberts, and K. A. Nugent. Phase retrieval with the transport of intensity equation: matrix solution with the use of Zernike polynomials. *J.Opt.Soc.Am.A*, 12(9):1933–1941, Sept 1995.

[81] P. Halmos. The range of a vector measure. *Bull. Amer. Math. Soc.*, 54:416–421, 1948.

[82] I. Halpern. The product of projection operators. *Acta Sci. Math.*, 23:96–99, 1962.

[83] P. C. Hansen. *Rank-Deficient and Discrete Ill-Posed Problems.* SIAM, 1998.

[84] K. Hanson. Bayesian and related methods in image reconstruction from incomplete data. In H. Stark, editor, *Image recovery: Theory and Application*, pages 79–125. Academic Press, 1987.

[85] M. H. Hayes. *Signal Reconstruction from Phase or Magnitude.* PhD thesis, Massachusetts Institute of Technology, 1981.

[86] M. H. Hayes. The unique reconstruction of multidimensional sequences from Fourier transform magnitude or phase. In H. Stark, editor, *Image Recovery: Theory and Application.* Academic Press, 1987.

[87] M. H. Hayes and A. V. Oppenheim. Signal reconstruction from phase or magnitude. *IEEE Trans. Acc. Sp. and Sig. Proc.*, ASSP-28(6):672–680, Dec. 1980.

[88] G. T. Herman. *Image Reconstruction from Projections – The Fundamentals of Computerized Tomography.* Academic Press, New York, 1980.

[89] W. Hildenbrand. *Core and Equilibria of a Large Economy.* Princeton University Press, 1974.

[90] A. Huiser and H. Ferwerda. The problem of phase retrieval in light and electron microscopy of strong objects. II. on the uniqueness and stability of object reconstruction procedures using two defocused images. *Optica Acta*, 23:445–456, 1976.

[91] N. E. Hurt. *Phase Retrieval and Zero Crossings.* Kluwer Academic Publishers, 1989.

[92] A. D. Ioffe. Approximate subdifferentials and applications: II. *Mathematika*, 33:111–128, 1986.

[93] A. D. Ioffe. Absolutely continuous subgradients of nonconvex integral functionals. *Nonlinear Analysis, Theory, Methods and Applications*, 11(2):245–257, 1987.

[94] A. D. Ioffe. Approximate subdifferentials and applications: III. *Mathematika*, 36(71):1–38, 1989.

[95] A. D. Ioffe. Proximal analysis and approximate subdifferentials. *J. London Math Soc.*, 41:175–192, 1990.

[96] T. Isernia, G. Leone, R. Pierri, and F. Soldovieri. Role of support information and zero locations in phase retrieval by a quadratic approach. *J. Opt. Soc. Am. A*, 16(7):1845 – 1856, 1999.

[97] J. Jacod and P. Protter. *Probability Essentials.* Springer, 1991.

[98] A. Jain. *Fundamentals of Digital Image Processing.* Prentice Hall, 1989.

[99] F. Jones. *Lebesgue Integration on Euclidean Space.* Jones and Bartlett, 1993.

[100] Y. Kano and E. Wolf. Temporal coherence of black body radiation. *Proc. of the Phys. Soc. (London)*, 80:1273–1276, 1962.

[101] R. Kendrick, D. Acton, and A. Duncan. Phase diversity wave-front sensor for imaging systems. *Appl.Opt.*, 33(27):6533–6546, 1994.

[102] A. Kirsch. *An Introduction to the Mathematical Theory of Inverse Problems.* Springer Verlag, 1996.

[103] K. Knox. Image retrieval from astronomical speckle patterns. *J.Opt.Soc.Am.*, 66:1236–1239, November 1976.

[104] R. Kress. *Linear integral equations*, volume 82 of *Applied Mathematical Sciences.* Springer Verlag, New York, 2 edition, 1999.

[105] J. E. Krist and C. J. Burrows. Phase retrieval analysis of pre and post-repair Hubble Space Telescope images. *Appl. Opt.*, 34(22):4951–4964, 1995.

[106] A. Y. Kruger. Properties of generalized differentials. *Siberian Math. J.*, 26:822–832, 1985.

[107] K. Kuratowski and C. Ryll-Nardzewski. A general theorem on selectors. *Bull. Polish Acad. Sci.*, 13:397–411, 1965.

[108] D. J. Lee, M. C. Roggemann, B. M. Welsh, and E. R. Crosby. Evaluation of least-squares phase-diversity technique for space telescope wave-front sensing. *App. Opt.*, 36(35):9186–9197, 1997.

[109] G. Leone, R. Pierri, and F. Soldovieri. Reconstruction of complex signals from intensities of Fourier transform pairs. *J. Opt. Soc. Am. A*, 13(7):1546–1556, 1996.

[110] A. Levi and H. Stark. Image restoration by the method of generalized projections with application to restoration from magnitude. *J.Opt.Soc.Am.A*, 1(9):932–943, 1984.

[111] A. Liapunov. Sur les fonctions-vecteurs complètement additives. *Bull. Acad. Sci. USSR*, 4:465–478, 1940.

[112] H. M. Lloyd, S. M. Jefferies, J. R. P Angel, and E. K. Hege. Wave-front sensing with time-of-flight phase diversity. *Opt. Let.*, 26(7):402–404, 2001.

[113] P. D. Loewen. Optimization and nonlinear analysis. In A.D. Ioffe, L. Marcus, and S. Reich, editors, *Pitman Research Notes Math.*, 244, chapter Limits of Fréchet normals in nonsmooth analysis, pages 178–188. 1992.

[114] M. G. Lofdahl, G. B. Scharmer, and W. Wei. Calibration of a deformable mirror and strehl ratio measurements by use of phase diversity. *App. Opt*, 39(1):94–103, 2000.

[115] D. R. Luke, J. V. Burke, and R. Lyon. Fast algorithms for phase diversity and phase retrieval. In *Proceedings of the Workshop On Computational Optics And Imaging For Space Applications*, NASA/Goddard Space Flight Center, May 2000. Optical Society of America.

[116] R. Lyon. DCATT wavefront sensing and optical control study. Technical Report WFSC-0001, NASA/Goddard Space Flight Center, February 1999.

[117] R. Lyon, J. Dorband, and J. Hollis. Hubble Space Telescope faint object camera calculated point spread functions. *Appl.Opt.*, 36(8):1752–1765, 1997.

[118] R. Lyon, P. Miller, and A. Grusczak. Hubble Space Telescope phase retrieval: a parameter estimation. In A. Tescher, editor, *Applications of Digital Image Processing XIV*, volume 1567 of *Proc.SPIE*, pages 317–326, 1991.

[119] V. Mahajan. Zernike annular polynomials for imaging systems with annular pupils. *J.Opt.Soc.Am.*, 71(1):75–85, 1981.

[120] J. C. Mather. NGST. volume 4013 of *Proc. SPIE*, pages 2–16, 2000.

[121] A. B. Meinel, M. P. Meinel, and D. H. Schulte. Determination of the Hubble Space Telescope effective conic-constant error from direct image measurements. *Appl. Opt.*, 32(10):1715–1719, Apr. 1993.

[122] J. Miao, P. Charalambous, J. Kirz, and D. Sayre. Extending the methodology of x-ray crystallography to allow imaging of micrometre-sized non-crystalline specimens. *Nature*, 400:342–344, 1999.

[123] J. Miao, D. Sayre, and H. N. Chapman. Phase retrieval from the magnitude of the Fourier transforms of nonperiodic objects. *J.Opt. Soc. Am.*, 15(6):1662–9, 1998.

[124] R. Millane. Phase retrieval in crystallography and optics. *J.Opt. Soc. Am. A.*, 7:394–411, 1990.

[125] D. L. Misell. An examination of an iterative method for the solution of the phase problem in optics and electron optics I. test calculations. *J. Phys. D.*, 6(18):2200–2216, 1973.

[126] B. S. Mordukhovich. *Approximation Methods in Problems of Optimization and Control.* Nauka, Moscow, 1988. Russian.

[127] B. S. Mordukhovich and Y. Shao. On nonconvex subdifferential calculus in Banach spaces. *J. of Convex Anal.*, 2:211–228, 1995.

[128] B. S. Mordukhovich and Y. Shao. Extremal characterizations of Asplund spaces. *Proc. Amer. Math. Soc.*, 124:197–205, 1996.

[129] B. S. Mordukhovich and Y. Shao. Nonsmooth sequential analysis in asplund spaces. *Trans. Amer. Math. Soc.*, 328(4):1235–1280, 1996.

[130] J. J. Moré and D. C. Sorenson. Computing a trust region step. *SIAM J. Sci.Comput.*, 4:553–572, 1983.

[131] Z. MouYan and R. Unbehauen. Methods for reconstruction of 2-d sequences from Fourier transform magnitude. *IEEE Trans.Im.Proc.*, 6(2):222–234, Feb 1997.

[132] J. Nocedal. Updating quasi-newton matrices with limited storage. *Math. Prog.*, 35:773–782, 1980.

[133] J. Nocedal and S. Wright. *Numerical Optimization.* Springer Verlag, 2000.

[134] R. J. Noll. Zernike polynomials and atmospheric turbulence. *J. Opt. Soc. Am.*, 66:207–211, 1976.

[135] Y Ohneda, N Baba, N Miura, and T Sakurai. Multiresolution approach to image reconstruction with phase-diversity technique. *Opt. Rev.*, 8(1):32–36, 2001.

[136] E. L. O'Neill and A. Walther. The question of phase in image formation. *Optica Acta*, 10:33–40, 1963.

[137] A. V. Oppenheim and R. W. Schafer. *Digital Signal Processing.* Prentice-Hall, 1975.

[138] S. S. Oren and E. Spedicato. Optimal conditioning of self-scaling variable metric algorithms. *Math.Prog.*, 10:70–90, 1976.

[139] R. G. Paxman, T. J. Schultz, and J. R. Fienup. Joint estimation of object and aberrations by using phase diversity. *J.Opt.Soc.Am.A*, 9(7):1072–1085, 1992.

[140] R. G. Paxman, J. H. Seldin, M. G. Lüfdahl, G. B. Scharmer, and C. U. Keller. Evaluation of phase-diversity techniques for solar-image restoration. *Astrophys.J.*, 466:1087–1099, Aug 1996.

[141] R. R. Phelps. personal communication.

[142] R. R. Phelps. Convex sets and nearest points, I. *Pro. Amer. Math. Soc.*, 8:790–797, 1957.

[143] R. R. Phelps. Convex sets and nearest points, II. *Pro. Amer. Math. Soc.*, 9:867–873, 1957.

[144] R. J. Plemmons and V. P. Pauca. Some computational problems arising in adaptive optics imaging systems. *J. Comp. and Appl. Math.*, 123:467–487, 2000.

[145] T. Quatieri and A.V. Oppenheim. Iterative techniques for minimum phase signal reconstruction from phase or magnitude. *IEEE Trans. on Acc.,Sp.and Sig. Proc.*, ASSP-29(6):1187–1193, Dec. 1981.

[146] D. Redding, S. Basinger, D. Cohen, A. Lowman, F. Shi, P. Bely, C. Bowers, R. Burg, L. Burns, P. Davila, B. Dean, G. Mosier, T. Norton, P. Petrone, B. Perkins, and M. Wilson. Wavefront control for a segmented deployable space telescope. Technical report, Jet Propulsion Labs, NASA/Goddard Space Flight Center, and Space Telescope Science Institute, 2000.

[147] D. Redding, S. Basinger, A. Lowman, A. Kissil, P. Bely, R. Burg, R. Lyon, G. Mosier, M. Wilson, M. Femiano, M. Wilson, G. Schunk, L. Craig, D. Jacobson, J. Rakoczy, and J. Hadaway. Wavefront sensing and control for a Next Generation Space Telescope. volume 3356 of *Proc. SPIE*, 1998.

[148] D. Redding, B. M. Levine, J.W. Yu, and J.K. Wallace. Hybrid ray-trace and diffraction propagation code for analysis of optical systems. In Y. Kohanzadeh, G. N. Lawrence, G. McCoy, and H. Weichel, editors, *Design, Modeling, and Control of Laser Beam Optics*, volume 1625 of *Proc. SPIE*, pages 95–107, 1992.

[149] C. H. Reinsch. Smoothing by spline functions, ii. *Numer. Math.*, 16:451–454, 1971.

[150] H. Richter. Verallgemeinerung eines in der Statistik benötigten Satzes der Masstheorie. *Math. Annalen.*, 150:85–90, 1963.

[151] R. T. Rockafellar. *Convex Analysis.* Princeton University Press, 1970.

[152] R. T. Rockafellar. Integral functionals, normal integrands and measurable selections. In *Nonlinear Operators in the Calculus of Variations*, number 543 in Lecture Notes in Mathematics, pages 157–207. Springer-Verlag, 1976.

[153] R. T. Rockafellar. Generalized directional derivatives and subgradients of nonconvex functions. *Can. J. Math.*, 32:257–280, 1980.

[154] R. T. Rockafellar and R. J. Wetts. *Variational Analysis.* Springer, 1998.

[155] C. Roddier and F. Roddier. Combined approach to the Hubble Space Telescope wave-front distortion analysis. *Appl. Opt.*, 32:2992–3008, 1993.

[156] F. Roddier. Curvature sensing and compensation: a new concept in adaptive optics. *Appl. Opt.*, 27:1223–1225, 1988.

[157] F. Roddier. Wavefront sensing and the irradiance transport equation. *Appl. Opt.*, 29:1402–1403, 1990.

[158] P. Roman and A. S. Marathay. Analyticity and phase retrieval. *Nuovo Cimento (X)*, 30:1452–1464, 1963.

[159] W. Rudin. *Real and Complex Analysis.* McGraw-Hill, 2nd edition, 1974.

[160] W. Rudin. *Functional Analysis.* International Series in Pure and Applied Mathematics. McGraw-Hill, San Francisco, 1991.

[161] D. Schmeidler. Fatou's lemma in several dimensions. *Proc. Amer. Math. Soc.*, 24(2):300–306, 1970.

[162] D. Schmeidler. A remark on the core of an atomless economy. *Econometrica*, 40(3):579–580, 1972.

[163] B. D. Seery and E. P. Smith. NASA's Next Generation Space Telescope visiting a time when galaxies were young. volume 3356 of *Proc. SPIE*, pages 2–13, 1998.

[164] J. H. Seldin and J. R. Fienup. Numerical investigation of the uniqueness of phase retrieval. *J. Opt. Soc. Amer. A*, 7(3):412–27, 1990.

[165] M. I. Sezan and H. Stark. Applications of convex projection theory to image recovery in tomography and related areas. In H. Stark, editor, *Image recovery: Theory and Applications*, pages 415–462. Academic Press, 1987.

[166] D. F. Shanno and K. Phua. Matrix conditioning and nonlinear optimization. *Math. Prog.*, 14:149–160, 1978.

[167] G. B. Sharmer. Object-independent fast phase-diversity. *Astronomical-Society-of-the-Pacific-Conference-Series*, 183:330–341, 1999.

[168] I. Singer. Some remarks on approximative compactness. *Rev. Roumaine Math. Pures Appl.*, 9:167–177, 1964.

[169] I. Singer. *ISNM*, volume 20, chapter On Set-Valued Metric Projections, pages 217–233. Birkh auser, 1972.

[170] A. Sommerfeld. *Optics.* Academic Press, 1954.

[171] H. Stark, editor. *Image Recovery: Theory and Application.* Academic Press, 1987.

[172] H. Stark and M. I. Sezan. *Real-time optical information processing*, chapter Image processing using projection methods, pages 185–232. Academic Press, London, UK, 1994.

[173] W. J. Stiles. Closest point maps and their product, II. *Nieuw Archief voor Wiskunde*, 13:212–225, 1965.

[174] J. W. (Lord Rayleigh) Strutt. On the interference bands of approximately homogeneous light; in a letter to Prof. A. Michelson. *Phil.Mag.*, 34:407–411, 1892.

[175] W. Swantner and Weng W. Chow. Gram-schmidt orthonormalization of Zernike polynomials for general aperture shapes. *Appl.Opt.*, 33(10):1832–7, Apr. 1994.

[176] A. Szoke. Holographic microscopy with a complicated reference. *J. Imaging Sci.Tech.*, 41:332–341, 1997.

[177] H. Takajo, T. Shizuma, T. Takahashi, and S. Takahata. Reconstruction of an object from its noisy Fourier modulus: ideal estimate of the object to be constructed and a method that attempts to find that object. *Appl. Opt.*, 38(26):5568–5576, Sept 1999.

[178] H. Takajo, T. Takahashi, R. Ueda, and M. Taninaka. Study on the convergence property of the hybrid input-output algorithm used for phase retrieval. *J. Opt. Soc. Am. A*, 15(11):2849–2861, Nov 1998.

[179] M. R. Teague. Deterministic phase retrieval: A Green's function solution. *J.Opt.Soc.Am.*, 73(11):1434–1441, Nov 1983.

[180] F. G. Tricomi. *Integral Equations.* Dover, New York, 1957.

[181] R. Tyson. *Principles of Adaptive Optics.* Academic Press, 1991.

[182] P. van Toorn and H. Ferwerda. The problem of phase retrieval in light and electron microscopy of strong objects. III. developments of methods for numerical solution. *Optica Acta*, 23:456–468, 1976.

[183] P. van Toorn and H. Ferwerda. The problem of phase retrieval in light and electron microscopy of strong objects. IV.checking algorithms by means of simulated objects. *Optica Acta*, 23:468–481, 1976.

[184] J. Véran, F. Rigaut, H. Maître, and D. Rouan. Estimation of the adaptive optics long-exposure point-spread function using control-loop data. *J.Opt.Soc.Am.A.*, 14(11):3057–3068, 1997.

[185] L. P. Vlasov. Approximative properties of sets in normed linear spaces. *Russian Math. Surveys*, pages 1–66, 1973.

[186] C. R. Vogel. A limited memory BFGS method for an inverse problem in atmospheric imaging. In P .C. Hansen, B.H. Jacobsen, and K. Mosegaard, editors, *Methods and Applications of Inversion*, volume 92 of *Lecture Notes in Earth Sciences*, pages 292–304. Springer-Verlag, 2000.

[187] C. R. Vogel, T. Chan, and R Plemmons. Fast algorithms for phase diversity-based blind deconvolution. In *Adaptive Optical System Technologies*, volume 3353 of *Proc. SPIE*, 1998.

[188] C. R. Vogel, T. Chan, and R Plemmons. Fast algorithms for phase diversity-based blind deconvolution. Technical report, Department of Mathematical Sciences, Montana State University, 1999.

[189] J. von Neumann. On rings of operators, reduction theory. *Ann. Math.*, 50:401–485, 1949.

[190] J. von Neumann. *Functional Operators, Vol II*, volume 22 of *Ann. Math Stud.* Princeton University Press, 1950.

[191] D. H. Wagner. Survey of measurable selection theorems. *SIAM J. Contr. Opt.*, 15:859–903, 1977.

[192] A. Walther. The question of phase retrieval in optics. *Optica Acta*, 10:41–49, 1963.

[193] D. Werner. *Funktional analysis.* Springer, 1995.

[194] E. Wigner. On the quantum correction for thermodynamic equilibrium. *Physical Review*, 40:749–759, 1932.

[195] E. Wigner. *Quantum Mechanical Distribution Functions Revisited*, chapter 4, pages 25–36. MIT Press, Cambridge, 1971.

[196] E. Wolf. Is a complete determination of the energy spectrum of light possible from measurements of degree of coherence? *Proc. of the Phys. Soc. (London)*, 80:1269–1272, 1962.

[197] J. W. Wood, M. A. Fiddy, and R. E. Burge. Phase retrieval using two intensity measurements in the complex plane. *Optics Letters*, 6(11):514–516, Nov 1981.

[198] G. Z. Yang, B. Z. Dong, B. Y. Gu, J. Y. Zhuang, and O. K. Ersoy. Gerchberg-Saxton and Yang-Gu algorithms for phase retrieval in a nonunitary transform system: a comparison. *Applied-Optics*, 33(2):209–219, Jan 1994.

[199] Kôsaku Yosida. *Functional Analysis.* Springer-Verlag, New York, 2nd edition, 1968.

[200] D. C. Youla. Mathematical theory of image restoration by the method of convex projections. In H. Stark, editor, *Image Recovery: Theory and Applications*, pages 29–77. Academic Press, 1987.

[201] D. C. Youla and V. Velasco. Extensions of a result on the synthesis of signals in the presence of inconsistent constraints. *IEEE Trans. Circuits Syst.*, 33(4):465–468, 1986.

[202] D. C. Youla and H. Webb. Image restoration by the method of convex projections: Part I - theory. *IEEE Trans Med. Im.*, MI-1(2):81–94, Oct. 1982.

[203] E. H. Zarantonello. Projections on convex sets in Hilbert space and spectral theory. In E. H. Zarantonello, editor, *Contributions to Nonlinear Functional Analysis*, pages 237–424. Academic Press, 1971.

[204] F. Zernike. Beugungstheorie des schneidenverfahrens und seiner verbesserten form, der phasenkontrastmethode. *Physica*, 1:689–794, 1934.

[205] X. Zhuang, E. Ostevold, and R. Haralick. The principle of maximum entropy in image recovery. In H. Stark, editor, *Image Recovery: Theory and Applications*, pages 157–193. Academic Press, 1987.

[206] G. Zoutendijk. Maximizing a function in a convex region. *J. Roy. Statist. Soc. Ser. B*, 21(2), 1959.

Appendix A

CONVOLUTION AND AUTOCORRELATION PROPERTIES

Define the correlation operator, $\star$, by

$$f \star g(\boldsymbol{x}) \equiv \int_{\mathbb{R}^n} f(\boldsymbol{x}')g(\boldsymbol{x}+\boldsymbol{x}')d\boldsymbol{x}'. \tag{A.1}$$

Define the conjugate correlation operator, $\Box$, by

$$f \Box g(\boldsymbol{x}) \equiv \int_{\mathbb{R}^n} f(\boldsymbol{x}')g(\boldsymbol{x}'-\boldsymbol{x})d\boldsymbol{x}'. \tag{A.2}$$

Denote the complex conjugate of a function $f : \mathbb{R}^n \to \mathbb{C}$ by $\overline{f}$. The inner product is defined as $\langle f, g\rangle \equiv \int_{\mathbb{R}^n} f(\boldsymbol{x})\overline{g}(\boldsymbol{x})d\boldsymbol{x}$.

Property A.0.1 *For* f, g, *and* $h : \mathbb{R}^n \to \mathbb{C}$

(a) $f \star g = [f^{\vee} g^{\wedge}]^{\vee} = f^{\wedge\wedge} * g$;

(b) $f \Box g = [f^{\wedge} g^{\vee}]^{\vee} = f * g^{\wedge\wedge}$;

(c) $f \Box g = g \star f$;

(d) $f \Box g = (f \star g)^{\wedge\wedge}$;

(e) $\langle f \star g, h\rangle = \langle \overline{h} \star g, \overline{f}\rangle$;

(f) $\langle f \Box g, h\rangle = \langle f \Box \overline{h}, \overline{g}\rangle$;

(g) if $h^{\vee}$ *is real-valued then* $\langle f \Box g, h\rangle = \langle h \star g, \overline{f}\rangle$;

(h) $\langle f * g, h\rangle = \langle g, \overline{f} \star h\rangle$. *In particular, the adjoint of the convolution operator* L *with convolution kernel* f *is the* autocorrelation *operator with autocorrelation kernel* $\overline{f}$, i.e. *if* $Lg \equiv f * g$ *then* $L^{*}g \equiv \overline{f} \star g$.

DETAIL:

(a)

$$\begin{aligned} f^{\wedge\wedge} * g(\boldsymbol{x}) &= \int_{\mathbb{R}^n} f^{\wedge\wedge}(\boldsymbol{x}')g(\boldsymbol{x}-\boldsymbol{x}')d\boldsymbol{x}' \\ &= \int_{\mathbb{R}^n} f(-\boldsymbol{x}')g(\boldsymbol{x}-\boldsymbol{x}')d\boldsymbol{x}' \\ &= \int_{\mathbb{R}^n} f(\boldsymbol{y})g(\boldsymbol{x}+\boldsymbol{y})d\boldsymbol{y} \end{aligned}$$

where $\boldsymbol{y} = -\boldsymbol{x}$. By the Convolution Theorem,

$$\begin{aligned} f^{\wedge\wedge} \star g &= [f^{\wedge\wedge\wedge} * g^{\wedge}]^{\vee} \\ &= [f^{\vee} g^{\wedge}]^{\vee} \end{aligned}$$

(b) (Same as above).

(c)

$$\begin{aligned} f\Box g(\boldsymbol{x}) &= \int_{\mathbb{R}^n} f(\boldsymbol{x}')g(\boldsymbol{x}'-\boldsymbol{x})d\boldsymbol{x}' \\ &= \int_{\mathbb{R}^n} f(\boldsymbol{y}+\boldsymbol{x})g(\boldsymbol{y})d\boldsymbol{y},\ (\boldsymbol{y}=\boldsymbol{x}'-\boldsymbol{x}) \\ &= g \star f(\boldsymbol{x}). \end{aligned}$$

(d)

$$\begin{aligned} (f \star g)^{\wedge\wedge} &= (f^{\wedge\wedge} * g)^{\wedge\wedge} \\ &= [(f^{\wedge\wedge\wedge} \cdot g^{\wedge})^{\vee}]^{\wedge\wedge} \\ &= (f^{\wedge\wedge\wedge})^{\wedge} * g^{\wedge\wedge} \\ &= f * g^{\wedge\wedge}. \end{aligned}$$

(e)

$$\begin{aligned} \langle f\Box g, h\rangle &= \int_{\mathbb{R}^n}\left[\int_{\mathbb{R}^n} f(\boldsymbol{y})g(\boldsymbol{x}+\boldsymbol{y})d\boldsymbol{y}\right]\overline{h}(\boldsymbol{x})d\boldsymbol{x} \\ &= \int_{\mathbb{R}^n}\left[\int_{\mathbb{R}^n} \overline{h}(\boldsymbol{x})g(\boldsymbol{x}+\boldsymbol{y})d\boldsymbol{x}\right] f(\boldsymbol{y})d\boldsymbol{y} \\ &= \langle \overline{h} \star g, \overline{f}\rangle. \end{aligned}$$

(f) This is a consequence of property (3) and (5).

(g)

$$\begin{aligned}\langle f \star g, h\rangle &= \int_{\mathbb{R}^n}\left[\int_{\mathbb{R}^n} f(\boldsymbol{y})g(\boldsymbol{y}-\boldsymbol{x})d\boldsymbol{y}\right]\overline{h}(\boldsymbol{x})d\boldsymbol{x}\\ &= \int_{\mathbb{R}^n}\left[\int_{\mathbb{R}^n} \overline{h}(\boldsymbol{x})g(\boldsymbol{y}-\boldsymbol{x})d\boldsymbol{x}\right] f(\boldsymbol{y})d\boldsymbol{y}.\end{aligned}$$

Now

$$\begin{aligned}h^{\vee}\in\mathbb{R} \Rightarrow h(\boldsymbol{x}) &= [h^{\vee}]^{\wedge}(\boldsymbol{x})\\ &= [\overline{h^{\vee}}]^{\wedge}(\boldsymbol{x})\\ &= \overline{h}^{\wedge\wedge}(\boldsymbol{x})\\ &= \overline{h(-\boldsymbol{x})}.\end{aligned}$$

Thus

$$\begin{aligned}\int_{\mathbb{R}^n}\overline{h}(\boldsymbol{x})g(\boldsymbol{y}-\boldsymbol{x})d\boldsymbol{x} &= \int_{\mathbb{R}^n} h(-\boldsymbol{x})g(\boldsymbol{y}-\boldsymbol{x})d\boldsymbol{x}\\ &= \int_{\mathbb{R}^n} h(\boldsymbol{x}')g(\boldsymbol{y}+\boldsymbol{x}')d\boldsymbol{x}'\\ &= h\star g(\boldsymbol{y}).\end{aligned}$$

(h)

$$\begin{aligned}\langle f * g,\ h\rangle &= \langle (f*g)^{\wedge},\ h^{\wedge}\rangle\\ &= \langle f^{\wedge}g^{\wedge},\ h^{\wedge}\rangle\\ &= \langle g^{\wedge},\ \overline{f^{\wedge}}h^{\wedge}\rangle\\ &= \left\langle g^{\wedge},\ \overline{f}^{\vee}h^{\wedge}\right\rangle\\ &= \left\langle g,\ [\overline{f}^{\vee}h^{\wedge}]^{\vee}\right\rangle\\ &= \langle g,\ \overline{f}\star h\rangle.\end{aligned}$$

□

Remark: The condition $h^{\vee}\in\mathbb{R}$ for property (A.1.g) is also known as the Hermetian property of h. This is the continuous analog of the Hermetian property for matrices.

Remark: The $\star$ and $\square$ operators can be efficiently calculated in $\mathcal{O}(N\log N)$ operations with three FFT's each since, by the convolution theorem,

$$f\star g = [f^{\vee}\cdot g^{\wedge}]^{\vee}$$
$$f\square g = [f^{\wedge}\cdot g^{\vee}]^{\wedge}.$$

VITA

David Russell Luke was born in Clifton Forge, Virginia on April 20, 1969. He grew up in central Ohio, graduating from Granville High School in 1987. Russell attended college at the University of California, Berkeley where he graduated with honors in Applied Mathematics in 1991. After four years of "real world school", part spent making documentary films (assistant editor on *The Ride to Wounded Knee*, 1992) and part spent organizing Self-Help Housing for low income residents of Eastern Washington, Russell returned to mathematics at the University of Washington's Department of Applied Mathematics in 1995. On the way toward a Ph.D. he received a Master of Science in Applied Mathematics from the University of Washington in 1997.