Optical Metrology by Prescription Retrieval and Transverse-Translation Diversity Phase Retrieval

by

Dustin B. Moore

Submitted in Partial Fulfillment

of the

Requirements for the Degree

Doctor of Philosophy

Supervised by

Professor James R. Fienup

The Institute of Optics
Arts, Sciences and Engineering
Edmund A. Hajim School of Engineering and Applied Sciences

University of Rochester
Rochester, New York

2016
To Alice and April, who let me see clearly.
Biographical Sketch

Dustin B. Moore was born in Oregon City, Oregon on June 8th, 1980. In 1998, he was admitted to Reed College in Portland, Oregon. There, he majored in Physics and wrote his undergraduate thesis on experiments in computed tomography using neutrons. He received a Bachelor of Arts degree, Phi Beta Kappa, in 2002.

After graduation he worked at Pinecone Imaging/Venture Ad Astra, in West Linn, Oregon, to support work on hypothetical sparse-aperture telescopes. He did original research on deconvolution through image-based wavefront sensing, and developed optical models for novel imaging systems. In 2004, he worked at the Immersive Media Company of Portland, Oregon and supported their full-motion panoramic imaging technology with software and special projects. There, he worked on the effort that collected the first datasets that became Google Streetview™. In 2008, he joined ZuluTime LLC of Wilsonville, Oregon, to develop multipath mitigation strategies for WiFi-based device location.

In 2009, Dustin entered the PhD program of The Institute of Optics at the University of Rochester. He was awarded a Robert L. and Mary L. Sproull University Fellowship and studied phase retrieval in the context of optical metrology under advisor James R.
Fienup. For this research, he solved phase retrieval challenges related to ground and flight testing of components of the James Webb Space Telescope. In 2011, he was a co-winner of the Robert S. Hilbert Memorial Optical Design Competition.

**Peer-reviewed Journal Publications**


**Conference Publications**


Acknowledgments

I want to principally thank my advisor, Professor James R. Fienup, for his innumerable contributions to this work. The depth of his insight into the intersection of nonlinear estimation and imaging has little competition. The editorial clarity and finesse he brings to problems spanning a wide range of applications is remarkable, and he is joy to work with.

I would also like to thank Professors Miguel A. Alonso, Julie L. Bentley, Thomas G. Brown and Kevin J. Parker for their useful guidance during the production of this work. I would like to thank Professor Ji Liu for chairing the defense.

Much of the art of nonlinear optimization phase retrieval is passed by verbal tradition from senior group members to the next generation. I dearly want to thank all the other research group members past and present of the Fienup group. These include: Matt Bergkoetter, Nathan Clark, Zack DeSantis, Wesley Farriss, Manuel Guizar-Sicairos, Alex Iacchetta, Abhishek Kumar, Aaron Michalko, Sapna Shroff, Abbie Tippie Watnik, and Tom Zielinski. I would especially like to thank Alden Jurling for enjoyable discussions about the endless minutiae surrounding the art of
phase retrieval. I also want to thank his wife Jenny Jurling, for making my family feel at home in Rochester.

I would also like to thank the technical staff of the Wavefront Sensing Group at NASA’s Goddard Spaceflight Center which includes Bruce Dean and David Aronstein. I want to thank Bruce for introducing me to the pulse of phase retrieval applications at NASA and David for introducing me to the real-world challenges associated with the testing of the James Webb Space Telescope.

I would like to thank the staff of The Institute of Optics which has included: Kari Brick, Lissa Cotter, Gina Kern, Lori Russell, Maria Schnitzler, Gayle Thompson, and Noelene Votens.

I also want to thank my co-workers at Pinecone Imaging/Venture Ad Astra, the Immersive Media Company and Zulutime LLC, for inspiring me to dive into the deep part of the optics pool. I would particularly like to thank Geoff Rhoads who introduced me to all sorts of interesting problems, including phase retrieval.

This thesis would not have been possible without the background I learned from the wonderful faculty of the Physics Department of Reed College. I want to thank Professors John Essick, David Griffith, Mary James, Robert Reynolds and Nicholas Wheeler, each of whom made a unique impact on my professional life in their own way. Similarly, I want to thank the members of the North West Astronomy Group including Marc Chamberlin, Marvin DeSautel, Kent Johnson, Sandy Mikalow, Rick Read, and Phil Reid, for bringing computers, radio waves and telescopes into my world.
I cannot possibly understate the impact my father, Mike Moore, and my mother, Joan Moore, have had on enabling my success. Whatever the challenge—conceptual, mathematical, electrical, or logistical—they are always prepared. They are so kind, so supportive and have never inspired any less than my very best.

Finally, I want to thank my loving wife April Younglove, and my amazing daughter Alice Moore, for providing the light when the nights are dark, and bringing clarity when the ideas are unclear.
Abstract

Phase retrieval (PR) has been integral to the success of wavefront sensing for alignment of space telescope systems, like NASA’s Hubble Space Telescope and the James Webb Space Telescope, and it has shown promise for optical shop testing. It requires that an optical system image a target such as a star or an illuminated pinhole onto a detector array. The measured intensity, the point spread function (PSF), becomes input to a nonlinear optimization process which finds wavefront aberrations that predict PSFs consistent with the measured PSFs.

Transverse-translation-diversity phase retrieval (TTDPR) is a ptychographic algorithm in which multiple PSFs are acquired for various translations of a subaperture. The subaperture translates in a plane conjugate to the exit pupil. We assess information-theoretic limits on the ability to estimate subaperture translation, essential for accurate metrology, as a function of the aberrations of the system. Attention is given to the case that an unknown linear phase aberration, or equivalent detector or target motion, is present that varies with each PSF.

We then introduce unknown-transverse translation diversity phase retrieval: an algorithm that estimates the direction of translation and the distance traveled by the
subaperture from one PSF to the next. Uniquely, it estimates translations without prior information about the approximate direction of translation. It also estimates unknown point target motion and rotations of the subaperture between PSF acquisitions from the PSF data.

In prescription retrieval, the unknowns being searched for are the prescription parameters of a sequential raytrace of the optical system. This search was previously hampered by the use of gradients calculated by the method of finite-differences (FD). Gradients calculated using FD become slow for problems involving many unknowns, and/or a large amount of PSF data to be fit. We use the reverse mode of algorithmic differentiation (RMAD) to find fast analytic gradient computations for metrics depending on raytraces of systems including freeform surfaces. This is combined with a customized diffraction integral model to perform prescription retrieval using computationally-efficient gradient functions. Finally, we show a simulated example where the use of RMAD gradients speeds computation by a projected 10,000 times.
Contributors and Funding Sources

Professors James R. Fienup (advisor, Optics), Miguel A. Alonso (Optics), Julie L. Bentley (Optics), Thomas G. Brown (Optics) and Kevin J. Parker (Electrical and Computer Engineering) supervised the production of this thesis.

The Lyot stop pattern shown in Fig. 1.3 was supplied by Thomas P. Zielinski. Michael Moore and April Younglove provided proofreading assistance.

The contents of Chapters 2 and 3 are based on material published earlier in Applied Optics.

This work was funded by grants NNX13AC99A and NNX15AE94A from the National Aeronautics and Space Agency with additional funding provided by a Robert L. and Mary L. Sproull University Fellowship awarded by the University of Rochester.

I conducted the research herein described.
Table of Contents

Biographical Sketch iii

Acknowledgments vi

Abstract ix

Contributors and Funding Sources xi

List of Tables xv

List of Figures xvii

List of Algorithms xxii

1 Introduction 1

1.1 Imaging as an Interferometric Measurement .......................... 1

1.2 Imaging Systems that Measure Themselves .......................... 4

1.3 The James Webb Space Telescope ........................................ 10

1.4 Outline ............................................................................. 17

References .............................................................................. 18
## Subaperture Translation Estimation Accuracy in Transverse-Translation Diversity Phase Retrieval

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1 Introduction</td>
<td>27</td>
</tr>
<tr>
<td>2.2 Subaperture Translation Estimation Accuracy</td>
<td>30</td>
</tr>
<tr>
<td>2.3 Subaperture Translation Estimation with Shared Linear Phase</td>
<td>46</td>
</tr>
<tr>
<td>2.4 Subaperture Translation Estimation with Unshared Linear Phase</td>
<td>49</td>
</tr>
<tr>
<td>2.5 Monte-Carlo Experiments</td>
<td>50</td>
</tr>
<tr>
<td>2.6 Comparison with Phase Retrieval Simulation</td>
<td>55</td>
</tr>
<tr>
<td>2.7 Conclusion</td>
<td>61</td>
</tr>
<tr>
<td>References</td>
<td>66</td>
</tr>
</tbody>
</table>

## Unknown Transverse-Translation Diversity Phase Retrieval

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.1 Introduction</td>
<td>71</td>
</tr>
<tr>
<td>3.2 Discrete Model and Error Metric</td>
<td>77</td>
</tr>
<tr>
<td>3.3 Conventional TTDPR</td>
<td>87</td>
</tr>
<tr>
<td>3.4 Unknown TTDPR</td>
<td>90</td>
</tr>
<tr>
<td>3.5 Laboratory Experiment</td>
<td>97</td>
</tr>
<tr>
<td>3.6 Summary and Conclusion</td>
<td>107</td>
</tr>
<tr>
<td>3.7 Bilinear Interpolation</td>
<td>110</td>
</tr>
<tr>
<td>3.8 Gradient of the Error Metric</td>
<td>117</td>
</tr>
<tr>
<td>3.9 Coefficients of the Bias Independent Metric</td>
<td>123</td>
</tr>
<tr>
<td>References</td>
<td>123</td>
</tr>
</tbody>
</table>
4 Efficient Prescription Retrieval through Reverse-Mode Algorithmic Differentiation

4.1 Promise and Challenge of Prescription Retrieval 132

4.2 Raytrace Calculation and Gradient 143

4.3 The Diffraction Integral 181

4.4 Error Metric for Prescription Retrieval 196

4.5 Simulated Prescription Retrieval Example 197

4.6 Conclusion 212

References 215

5 Conclusion and Future Work 225

5.1 Key Contributions 225

5.2 Future Work 231

References 233
## List of Tables

2.1 Zernike polynomials in Cartesian coordinates with $j$, the index of the polynomial and $n$, the radial-degree. .................................................. 32

2.2 Non-zero terms of the Zernike translation matrix $\gamma_{qj}(q,r)$ for $j < 12$. 39

2.3 Non-zero terms of the linearized Zernike translation matrix $\gamma_{qj}^q$ for $j < 12$. ................................................................. 43

2.4 Non-zero terms of the linearized Zernike translation matrix $\gamma_{qj}^r$ for $j < 12$. ................................................................. 43

2.5 Mean wavefront sensing error variances estimated from the simulated phase retrieval experiment in Sect. 2.6. ............................... 53

4.1 Prescription parameters of the raytrace for the $j$th surface. .......... 150

4.2 Nominal prescription for 1 m F/21 Ritchey-Crétien telescope with refractive field-flattener. Object is at infinity and the full-field of view is 0.4 degrees. ......................................................... 198
4.3 Listing of defocus, field location, and wavelength for each PSF in prescription retrieval grouped by defocus and wavelengths. Defocus is a translation of image plane along the $z$ axis relative to the nominal image location. .................. 205

4.4 Listing of time to compute gradient with respect to various combinations of prescription parameters. Values with * are extrapolated from the time it took to evaluate the error metric .................. 211
List of Figures

1.1 Diagram of a NIRCam light path showing normal operation. NIRcam relays the internal image passed by ISIM to the detector when the clear aperture filter is selected. Marginal rays of a near-on axis point are drawn in red. Not shown: Fold mirror between image and collimating lens, another wheel called the filter wheel, and a dichroic beamsplitter between collimating lens and wheels. (Not to scale.) 12

1.2 Diagram of a NIRCam light path showing TTDPR operation with Lyot mask. NIRcam views the pinhole when the Lyot mask on the prism substrate is selected. Marginal rays from a point on the pinhole are drawn in red. Not shown: Fold mirror between image and collimating lens, another wheel called the filter wheel, and a dichroic beamsplitter between collimating lens and wheels. (Not to scale.) 15

1.3 Example transmission function similar to a “round” style Lyot stop of NIRCam. 16
2.1  (a) Sample exit pupil phase with linear phase subtracted and a colorbar
in radians, (b) sample wavefront phase having also had the defocus
subtracted with a colorbar in radians, (c) map of the subaperture/pupil
plane with a colorbar indicating number of subaperture positions overlapping
the pupil and (d) simulated noisy PSF from the sample exit pupil phase
and one of the subaperture positions. . . . . . . . . . . . . . . . . . . 52

2.2  Log-histogram of var(\hat{q}) over the population of coefficient vectors in
the SLP and USLP cases. . . . . . . . . . . . . . . . . . . . . . . . . 54

2.3  Log-log scatter plot of MSE(s) of simulated TTDPR experiments for
a given aberration vector versus the \( \sigma^{-2} \text{var}(\hat{q}) \) for the aberration for
both the SLP and USLP cases. Includes line representing the predicted
Cramer-Rao lower bound of Eq. (2.34). . . . . . . . . . . . . . . . . 59

3.1  Schematic of TTDPR laboratory experiment. Not shown: neutral
density filter and focusing microscope objective between the laser and
the spatial filter. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 98
3.2 (a) Example PSF of the aberrated system: intensity plotted to the 0.33 power. (b) Scatter plot of sub-aperture translations for the first experiment: red crosses are UTTDPR recovered values and black circles indicate true translations commanded to the X-Y stage. (c) Retrieved pupil phase for first experiment using UTTDPR with defocus and linear phase removed. (d) Phase difference (error) between UTTDPR and conventional TTDPR with knowledge of the subaperture translations for the first experiment.

3.3 (a) Areas of pupil accessed during subaperture motion in the second experiment. (b) Scatter plot of sub-aperture translations for the second experiment: red crosses are UTTDPR recovered values and black circles indicate true translations estimated from knowledge of rotation angle and geometry. (c) Fixed amplitude to the 0.5 power recovered by UTTDPR for the second experiment. (d) Fixed amplitude to the 0.5 power recovered by sieve amplitude TTDPR for the second experiment. (e) Retrieved pupil phase for the second experiment using UTTDPR with defocus and linear phase removed. (f) Phase difference (error) between UTTDPR and sieve amplitude TTDPR for the second experiment.
3.4  (a) Retrieved pupil phase for the second experiment using UTTDPR with defocus and linear phase removed. (b) Phase difference (error) between UTTDPR and sieve amplitude TTDPR for the second experiment. 107

4.1  Drawing of Ritchey-Crétien telescope simulated. 199

4.2  Transverse ray aberration plot of Ritchey-Crétien telescope employed to demonstrate prescription retrieval. 199

4.3  Stop amplitude model $A$ for a) the nominal system and b) the simulated perturbed system having additional features unknown to prescription retrieval. 200

4.4  Coefficients of the Zernike polynomial terms in the simulated data for prescription retrieval in (a) the primary surface, (b) the secondary surface and (c) the 7th surface. Polynomial terms are indexed by $j$ in the Noll ordering [60] and normalized over the unit circle of each surface’s clear aperture. 202
4.5 Surface sag deformation simulated in the PSF data for prescription retrieval in (a) primary surface, (b) secondary surface and (c) surface 7. Surface sag error estimated by prescription retrieval for the (d) primary surface, (e) secondary surface and (f) surface 7. Error in prescription retrieval result relative to the sag deformation simulated in the PSF data for (g) primary surface, (h) secondary surface and (i) surface 7. Each colorbar has a difference range of values, and positive sag indicates the surface is displaced in the positive z direction relative to the nominal hyperbola or spherical surface. Not to scale: the outer circle represents the clear aperture of each surface which is different between primary, secondary and the 7th surface. 

4.6 Simulated data PSFs. Intensity shown stretched to the one-quarter power.

4.7 Stop amplitude $A$ retrieved by prescription retrieval.
List of Algorithms

3.1 Calculation of $\overline{A_S}$, the gradient with respect to subaperture transmission function. We define $+= \text{operator that adds the right operand to the left operand and stores the result in the left operand.} \ldots 118$

4.1 Iterative estimation of ray intersection: Finds points $x_j$ on the surface and on each ray as well as the distance $L_j$ along the rays from last intersection points $x'_{j-1}$. \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots 171$

4.2 Notional freeform ray transfer computation to be differentiated using RMAD. Assumes existence of $L_j$ and finds outputs equivalent to the actual ray transfer computation. \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots 175$

4.3 Calculation of $\overline{\mathcal{f}}$ and $\overline{x}$ from $\overline{G}$. \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots 191
1

Introduction

1.1 Imaging as an Interferometric Measurement

From the microscope to the telescope, we use imaging systems to record moments that we do not wish to forget and to form images of targets that may be otherwise inaccessible to the eye. Light from a target that falls into an aperture of the imaging system, called the entrance pupil, becomes focused to an image plane. Once there, it is detected with an array detector such as a charge-coupled device (CCD) or other scientific instrument like a spectrometer. Physical models for the focusing process include geometrical optics, which views light as rays, and physical optics, which considers light to be a propagating wave. In the geometrical optics sense, rays of light originating from a point on the target cross the pupil and are then bent by the system to come to focus at a point. From the physical optics view, waves emitted by the object are described by a complex field which propagates to the entrance pupil. One fundamental statistic of this field is its spatial and temporal coherence: a measure of
the ability for the field at one place and time in the pupil to interfere constructively or destructively with the field measured somewhere else in the pupil at some other time. From the coherence function of the field in the pupil, an image of the target can be calculated [1] but, in a conventional optical system, this occurs without an explicit computation. Conversion of the pupil field into an image happens through (i) free space propagation and (ii) materials in geometries that impart spatially-varying delays to the field as it propagates to the image plane. When the squared magnitude of image plane field is time-averaged, this yields an image of the target. Incidentally, to be consistent with common usage in the computational optics community, we refer to this time averaged value as intensity although the proper term from radiometry is irradiance [2]. Usually, and quite usefully, the output of the array detector requires no additional computations to be interpreted as an image of the target.

As a contrast to the conventional optical imaging systems described above, consider the radio interferometers used in radio astronomy which use large arrays of antennas and radio receivers. These make direct measurements of the electric field that are digitally correlated to estimate the coherence statistics of the field. Sophisticated algorithms can then be applied to the coherence statistics to calculate an image. That conventional optical imaging systems accomplish the equivalent to this analysis with just glass and mirrors is miraculous. It is also necessary since the extreme frequency, hundreds of terrahertz (THz), of optical fields prevents the field correlations from being formed electronically for broadband light. Methods like the frequency mixing and heterodyne detection that allow radio interferometry are very challenging given
the bandwidth of typical optical fields. However, if the target is illuminated by a monochromatic and spatially-coherent source, this bandwidth issue can be resolved. If some of the light from the source can be delayed to within a temporal coherence length of the light reflected from the target, a superposition of this source field and the target field will form intensity fringes that are stationary long enough to be recorded. This process of coherent detection allows holographic imaging and direct determination of the fields in the aperture at the expense of needing to illuminate the object with very narrowband light. Most imaging applications prefer passive illumination of the target and, consequently, direct field determination by the holographic method is less common than conventional imaging systems that focus light to an image plane.

As the spectral bandwidth and the desired resolution increases, optical systems become increasingly large and complicated combinations of precision elements like lenses, mirrors, and structural components. The limiting resolution of an optical system is determined by over how wide an aperture it can constructively interfere light [3]. This effect, known as diffraction, requires that the image can never be as detailed as the target for a finite aperture width. Defects in the optical surfaces, as well as their design or placement, can also make the image much worse than diffraction allows with ideal optical surfaces. Consequently, minimizing these additional defects drives the precision of the shapes and materials used in the optical system. Optical systems must be constructed to tolerances that are fractions of an optical wavelength.

Since measurements using light can be very sensitive to errors, optical metrology is often executed with non-contact methods involving light. Classic interferometric
tools, like the common phase-shifting Fizeau interferometer, actively illuminate the system under test with coherent light [4]. As in holographic imaging, the light returning from the surface or system will interfere with a reference beam that is derived from the same source and is temporally coherent. Consequently, static fringes can be measured in some plane conjugate to the surface or the aperture of an imaging system under test. From these fringes, the phase differences that indicate the actual surface profile or system aberration may be inferred. Though the results are straightforward to interpret, optical metrology using conventional interferometric methods is not without complication. Additional optics are required to move the reference beam into a configuration where it and the aberrated beam from the system can be interfered. For certainty, the interferometric measurement should measure the system along the optical paths and at the conjugates that the imaging system will be used. This can necessitate that the additional precision optics be quite large, sometimes larger than the entrance pupil of the system under test. Such additional precision optics require their own optical metrology procedures and test apparatus, adding considerable expense. Also, the paths of both the aberrated beam and the reference beam must be thoroughly understood and care applied to avoid non-common path and retrace errors.

1.2 Imaging Systems that Measure Themselves

Since a conventional imaging system already interferes light in the image plane, it is reasonable to wonder if the expense associated with interfering light in a plane conjugate to the pupil, as in a classic interferometric measurement, is always required.
For many applications, we would like to require a minimal amount of additional precision testing hardware and, for that case, image-based wavefront sensing (IBWFS) methods provide useful alternatives. IBWFS takes advantage of an imaging system’s inherent interferometric nature that accepts a field in the pupil and interferes that field with itself, under perturbations, in the image plane. If the target being imaged by the system is an unresolved point, the detected image is the point spread function (PSF) of the optical system. The PSF is a very sensitive measurement of the effects of aberrations of the imaging system and of diffraction. From one or more PSFs of the system, IBWFS techniques may be applied to infer the underlying aberrations of the imaging system. This process has the advantage of not requiring significant optics beyond what can already be expected to be present in an imaging system. It can also test the paths through the instrument which will be utilized during regular imaging, mitigating potential non-common-path issues. The most primitive IBWFS methods include non-quantitative measurements such as the star test [4] and contrast-sensitive autofocus algorithms present in consumer cameras [5]. Quantitative aberration measurements can be made through phase retrieval. Of these phase retrieval methods, curvature wavefront sensing [6] and transport of intensity imaging [7, 8] use linearizing assumptions to simplify the problem. The advantage of these techniques is that they can find an approximate answer in a single non-iterative computation, but they are not without limitations. For increased flexibility and accuracy, the nonlinear problem of finding an unknown complex field in one plane consistent with measurements and constraints made in other planes must be solved. Techniques for this problem involve iterative
searching that may or may not find a suitable answer. These methods include iterative transform algorithms, such as the Gerchberg-Saxton [9, 10] and descendants like the hybrid-diversity algorithm [11, 12], and gradient-based nonlinear optimization (NLO) methods. The NLO methods include those popularized by Gonsalves [13, 14], Fienup [15–17], Paxman [18] and others [19–23].

Phase retrieval using NLO involves a “forward model” computation, an error metric and a particular optimization technique. The forward model involves a parameterized representation of the amplitude and phase in the exit pupil and a propagation that evolves this complex field to the image plane. Then, the field in the image plane is magnitude-squared to find a prediction for the intensity PSFs given the model parameters. This predicted PSF can be compared with the measured data using an error metric like mean squared error (MSE) or the detector bias and gain invariant metric [24]. A standard nonlinear optimization technique then varies the model parameters until the error metric is minimized and correspondence between the model PSFs and measured PSFs is achieved. Such optimization techniques include steepest descent, conjugate gradient and the Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithm [25]. The BFGS algorithm distinguishes itself by using gradient information to construct an approximation of the Hessian of the error metric. In the neighborhood of a minimum, this second derivative information allows for convergence in a small number of error metric and gradient evaluations for general optimization problems [26, 27], as well as phase retrieval [22, 28]. All the techniques listed above evaluate the gradient of the error metric many times during an iterative solution process. MSE metrics can
also be minimized by another iterative technique, the Levenberg-Marquardt (LM) algorithm, which is also known as damped least squares. The LM algorithm is extensively utilized in lens design optimizations [29–31] and some work has shown its benefit for phase retrieval [32]. We distinguish the LM algorithm from the above optimization methods since it requires the Jacobian of the forward model and not the gradient of the error metric.

Returning to gradient techniques, these techniques evaluate the error metric and the gradient of the metric at a starting point. The gradient is then used to define a search direction along a line in the space of all parameters, and the error metric is evaluated at points along this line. After finding an approximate minimum along the line, the technique evaluates the gradient of the error metric again to find a new search direction along which to search. This process is repeated until a minimum has been found where the gradient approaches zero and there are no further directions which can be searched without raising the error metric. There is no guarantee that the minimum of the error metric found by the optimizer will be the global minimum associated with finding the correct unknown parameters. Whether the optimizer has found the global or local minimum can often be assessed by comparing the model-predicted PSFs with the measured data. For instance, if a MSE error metric is used and a local minimum has been found, the value of the metric will be larger than the variance associated with the noise of the measurement. It may then be inferred that the optimizer found a local minimum rather than the global minimum. Stagnation of the optimization at an undesired local minimum is a common challenge in phase retrieval.
In NLO phase retrieval methods, a good optimizer will go from the starting parameters to the nearest minimum in parameter space. We have observed that trying retrievals with randomly generated starting parameter selections is advantageous for circumventing local minimums [33]. Global search techniques such as simulated annealing assist with avoiding local minima [34–36], as does changing the parameterization of the model to exclude diffraction effects [37, 38]. Similarly, switching phase retrieval methods from NLO to an iterative transform algorithm may be useful as these have different stagnation properties [15, 39, 40].

For phase retrieval with more than a few unknown parameters, it is desirable to have a fast computation for the analytic gradient of the error metric with respect to the model parameters. This is because the gradient is critical to minimizing the error metric, and computing the gradient using the method of finite differences [25] becomes untenable for functions of many unknown parameters. Similarly, computing the Jacobian matrix required for the LM algorithm is computationally infeasible for models predicting many pixels depending on many unknown parameters. Fienup [15, 16] showed that it was possible to write analytic functions for the gradients of some phase retrieval models that were computationally efficient. Jurling [41] observed that this is a consequence of the reverse-mode of algorithmic differentiation (RMAD) and that the gradient computation has a complexity on the order of that of the forward model and error metric. The forward model can be constructed as a composition of many simpler component models, each having an associated RMAD gradient calculation. Further, the RMAD gradient formula for a component model
can be independently tested against a finite-differenced approximation of the Jacobian for that component model to isolate any programming errors. This is a considerable improvement over the earlier monolithic phase retrieval gradients which were much more difficult to debug. Chapters 3 and 4 make extensive use of RMAD to find efficient derivatives of complicated forward models involving many different parameters. Specific examples include Section 3.7, where we find the gradient of a general-purpose bilinear resampling of an array, and Section 4.2, where we differentiate the sequential raytrace of an optical system.

This thesis focuses on the NLO strategy because of its general robustness, accuracy, and ability to model unusual measurement conditions with only modest complication. For example, it is easily adapted to broadband light [42, 43], PSF data undersampled relative to the Nyquist rate [42, 44, 45] and various forms of amplitude models such as [46] and that developed in Section 3.2. It also adapts to the unusual diversities and varied types of unknowns that must be estimated, capabilities which this thesis uses extensively. In phase retrieval, diversity is the notion that multiple PSFs can be obtained under perturbations of the optical system to make solving the inverse problem more well-posed. The most common diversity, focus diversity, involves changing the defocus of the optical system from one PSF to the next.

Much of this thesis considers an alternate form of diversity: transverse-translation diversity phase retrieval (TTDPR) [47–49]. It is an image-based wavefront sensing method similar to the coherent diffractive imaging technique known as ptychography [50–54]. In TTDPR, point spread functions (PSFs) of an aberrated optical system
are acquired as a subaperture moves in a plane perpendicular to the optical axis somewhere inside the system. For wavefront sensing, this plane is usually conjugate to a pupil plane of the system. The subaperture motion allows each PSF to come from the light passing through different regions of the plane but with some overlap. The wavefront over the union of these regions is then reconstructed by a phase retrieval algorithm. The measured PSF data, knowledge of the transmission function of the subaperture and the subaperture translation for each PSF can be used in a TTDPR algorithm to find the exit pupil phase and amplitude.

1.3 The James Webb Space Telescope

Our motivation for studying phase retrieval is National Aeronautics and Space Administration’s (NASA’s) next great space observatory, the James Webb Space Telescope (JWST). The goals for this observatory are to see the first stars, to infer the history of the early universe, and to study how galaxy and planetary system formation has changed over time [55]. To achieve high resolution imaging in the infrared, the JWST must interfere light collected over a massive 6.5 meter-diameter primary mirror. Since there is no launch vehicle with a fairing that large, that mirror is composed of 18 segments which deploy on orbit.

Optical metrology must also be performed in space and far from Earth to bring those segments into the proper alignment after the disturbances of launch. A suite of metrology methods will bring the mirrors into alignment [56], with the most precise fine-phasing being made using phase retrieval measurements [11, 12, 57]. These
methods will guide the primary mirrors and other surfaces into a prescription with an
overall optical path-length error of less than 131 nm root mean squared (RMS) over
the aperture [55]. The JWST is a perfect application for phase retrieval as it requires
minimal additional hardware and allows for testing the imaging system on-orbit at
the intended conjugates. This capability was also employed for JWST’s predecessor,
the Hubble Space Telescope, for which phase retrieval [17, 22, 58–60] was famously
used to measure a figure error that was later corrected in a servicing mission. Such
an error would be unacceptable on the JWST as the observatory will be at the second
Lagrange point (L2), over a million miles from earth, and inaccessible to manned
service missions. Consequently, corrections must be made remotely from measure-
ments derived from phase retrieval after coarser phase measurement methods. There is
an expansive ground testing plan to ensure that the aberrations measured and corrected
on orbit will be within the performance limits of the actuators. The plan makes
extensive use of conventional active interferometric methods as well as phase retrieval
to characterize the flight hardware. For instance, the Near-Infrared Camera (NIRCam)
instrument [61] was measured with phase retrieval at its place of manufacture by the
Lockheed Martin Corporation. NASA’s Goddard Space Flight Center tested NIRCam
before and after it was included in the Integrated Science Instrument Module (ISIM).
More phase retrieval will be used during cryogenic testing of the entire telescope
assembly in the large vacuum and cryogenic chamber of NASA’s Johnson Space
Center [62].
Figure 1.1: Diagram of a NIRCam light path showing normal operation. NIRcam relays the internal image passed by ISIM to the detector when the clear aperture filter is selected. Marginal rays of a near-on axis point are drawn in red. Not shown: Fold mirror between image and collimating lens, another wheel called the filter wheel, and a dichroic beamsplitter between collimating lens and wheels. (Not to scale.)
In this thesis, we are primarily interested in testing of the NIRCam instrument [61] on the JWST. This instrument is critical to the IBWFS efforts as its shortwave channel is used for focus-diverse phase retrieval for alignment and fine phasing of the primary mirror segments. There are two redundant NIRCam modules labeled A and B that utilize different regions of the telescope’s field of view. Each of these modules has two optical paths: the “shortwave” channel and the “longwave” channel. Each channel of both modules has dedicated filter wheels and detectors. In normal operation, each of the channels behaves approximately as diagrammed in Fig. 1.1, though this diagram is extremely simplified. Each of the lenses is actually multiple elements and there are fold mirrors and dichroic elements that are not drawn for simplicity. Light from the telescope assembly travels through ISIM to form an internal image just ahead of NIRCam. Each NIRCam optical path processes the internal image through two filter wheels before re-imaging it onto a detector. One of the wheels is referred to as the pupil wheel while the other is the filter wheel, though the pupil wheel also has narrow-band filters [61]. For focus-diverse phase retrieval using the shortwave detectors, weak defocus filters can be selected in both wheels to provide various shifts in focus. This conventional phase retrieval cannot distinguish between aberrations of the primary mirror of the ISIM module and the aberrations intrinsic to NIRCam since all of these aberrations would appear superimposed in NIRCam’s exit pupil. While we do not expect NIRCam to have significant aberrations, it is useful to have a risk-mitigation strategy against the potential for ground testing and launch to alter NIRCam in an unknown fashion.
A light path that does not go through the main telescope and the segmented primary mirror is very useful for detecting the aberrations of just NIRCam using phase retrieval. There is a small pinhole that can be illuminated by an LED on the coronagraphic occulter assembly of NIRCam [63], and it is even very near the plane of the internal image. Unfortunately, it is outside the field of view of the detector in normal operation, and thus inaccessible to focus-diverse phase retrieval. As shown in Fig. 1.2, it is possible to view these pinhole sources if a Lyot stop is selected in the pupil wheel. The Lyot stop transmission patterns are on a prism substrate that shifts the LED pinhole sources into the field of view of the detector.

The Lyot stops are intended for use in coronagraphy [61, 64] and one of the filters has a transmission function similar to the pattern shown in Fig. 1.3. However, selecting a Lyot stop in the pupil wheel significantly limits the ability to do focus-diverse phase retrieval at the same time. If a Lyot stop is selected in the pupil wheel of a shortwave channel, there is just one weak defocus lens left in the other wheel for focus-diverse phase retrieval. This provides a more limited form of diversity than is desirable. Alternately, having selected the Lyot stop in the pupil wheel of a longwave detector, focus-diverse phase retrieval cannot be employed at all as there are no weak defocus lenses available in the longwave channel. However, the pupil wheel can be rotated in small increments, thus translating the Lyot stop through the pupil plane to yield a form of TTDPR. Zielinski [48] examined TTDPR using stars, as regions of the sky are still visible through the telescope when the Lyot stop is in use. In this work, though, we focus on methods that enable TTDPR using the internal pinholes, shown
Figure 1.2: Diagram of a NIRCam light path showing TTDPR operation with Lyot mask. NIRcam views the pinhole when the Lyot mask on the prism substrate is selected. Marginal rays from a point on the pinhole are drawn in red. Not shown: Fold mirror between image and collimating lens, another wheel called the filter wheel, and a dichroic beamsplitter between collimating lens and wheels. (Not to scale.)
Figure 1.3: Example transmission function similar to a “round” style Lyot stop of NIRCam.

in Fig. 1.2, as well as on ground testing where test targets are presented to NIRCam at the internal image ahead of NIRCam.

TTDPR with NIRCam includes some unique challenges. Hardware and test specifics prevent the exact nature of the Lyot stop translation from being known to the necessary precision during early testing. Once thermal equilibrium is reached after launch, it is expected that the pupil wheel will give precise orientation feedback for which an accurate calibration can be developed. Before that, translation distances and rotation angles must be estimated from the PSF data. Further, other test constraints suggest that the target being imaged be allowed to move by an uncalibrated amount between each PSF acquisition to enable use of TTDPR with simulated sources during ground testing. Chapter 2 is devoted to understanding when joint estimation of subaperture translation and target motion is possible. Further complications arise in that there is a fold mirror between the pinhole and the collimating lens that induces an unknown
degree of vignetting. In Chapter 3, we develop a TTDPR algorithm that simultaneously estimates subaperture translation, target motion, pupil illumination and exit pupil phase. TTDPR provides a unique opportunity to track the aberrations of NIRCam (separate from the rest of the telescope) during ground testing and on orbit.

1.4 Outline

This thesis provides three major contributions to phase retrieval testing of the JWST and other imaging systems:

- In Chapter 2, we examine the basis for estimating subaperture translation in TTDPR. Information theoretic limits on the ability to estimate subaperture translation are developed as a function of the optical aberrations of the system being measured. Attention is given to the case that an unknown linear phase aberration, or equivalent detector or target motion, is present that varies with each PSF in the measured data. The theory is verified with high fidelity Monte-Carlo simulations.

- Chapter 3 introduces unknown-transverse translation diversity phase retrieval (UTTDPR): a ptychographic algorithm for optical metrology when a subaperture is translating through a plane conjugate to the exit pupil in a very poorly known fashion. The algorithm estimates the direction of translation and the distance traveled by the subaperture from one point spread function (PSF) to the next. It also estimates unknown point target motion and rotations of the subaperture between PSF acquisitions from the PSF data. Experimental results are presented.
• Prescription retrieval [59] is a process for estimating the parameters of an optical system, such as surface radii of curvature and thicknesses, directly from PSF data.

In Chapter 4, we find an efficient calculation for gradients of error metrics involving sequential raytraces of optical systems. We then use this calculation to construct an efficient prescription retrieval technique suitable for large numbers of unknown parameters and/or large amounts of PSF data. This technique is demonstrated on simulated data to retrieve multiple unknown freeform surfaces.

More detail on these contributions is provided in each chapter and in the conclusion: Chapter 5.

This introduction resolves on one final thought. This thesis provides a number of computer algorithms that enable testing and operation of sophisticated optical imaging systems like those discussed in Section 1.1. We are reminded, though, that the practice of optics is mostly concerned with engineering structures that passively bend electromagnetic radiation to remarkable effect. That these structures enable us to see so far, so large, and so small with just metal, glass and empty space is beautiful. We hope that our methods may enable these amazing structures to see what needs to be seen.

References


2

Subaperture Translation Estimation
Accuracy in Transverse-Translation Diversity Phase Retrieval

2.1 Introduction

Transverse-translation-diversity phase retrieval (TTDPR) [1–5] is an image-based wavefront sensing method similar to the coherent diffractive imaging technique known as ptychography [6–10]. In TTDPR, point spread functions (PSFs) of an aberrated optical system are typically acquired as a subaperture moves in a plane perpendicular to the optical axis somewhere inside the system. This plane is usually conjugate to a pupil plane of the system and for simplicity this discussion will be restricted to that case. The subaperture motion allows each PSF to come from the light passing through different regions of the plane but with some overlap. The wavefront over the union of these regions is then reconstructed by a phase retrieval algorithm. The measured PSF data, knowledge of the transmission function of the subaperture and the subaperture
translation for each PSF can be used in a TTDPR algorithm to find the exit pupil field. In this discussion, we assume it is just the phase of the complex field that is unknown, but it is also possible to estimate the amplitude and thus retrieve the complex field. If there is error in the subaperture translation knowledge, estimates of these values can be refined by including them as unknowns in the phase retrieval process [3]. In this case, the PSFs themselves must supply the information necessary to determine the missing subaperture translation knowledge, which is the central concern of this paper. We assume here that it is necessary to estimate subaperture translation from the contents of each PSF. For brevity, we will just refer to “translation” whenever translation of the subaperture is referenced.

Our motivation to understand translation estimation in TTDPR comes from monitoring the aberrations of the longwave channel on the Near-Infrared Camera (NIRCam) [11] of the James Webb Space Telescope. In this TTDPR application, a Lyot stop in a longwave pupil plane serves as the subaperture. The Lyot stops are intended for coronographic imaging [12] but also enable TTDPR wavefront sensing [4, 5]. The longwave channels of NIRCam’s two modules lack the weak defocus lenses present in the shortwave channels that permit conventional focus-diverse phase retrieval [11, 13]. TTDPR provides an opportunity to track the aberrations of NIRCam (separate from the rest of the telescope) during ground testing and on orbit. However, hardware and test specifics prevent the exact nature of the translation from being known to the necessary precision and must be estimated from the PSF data. Other test constraints require that the target being imaged be allowed to move by an uncalibrated amount.
between each PSF acquisition in some test applications. Provided such target motion is within the isoplanatic patch of the system, it can be modeled as a phase aberration linear in subaperture plane coordinates under the paraxial assumptions of Section 2.2. This additional unknown linear-phase component must be estimated by the phase retrieval algorithm separately for each PSF. We refer to TTDPR measurements where these linear-phase terms must be estimated separately for each PSF as the unshared linear phase (USLP) case. When the linear phase is the same for all PSFs, we refer to it as the shared linear phase (SLP) case.

For unobscured systems, like NIRCam imaging its internal source, most PSFs in a TTDPR measurement will derive from exit pupils lacking significant amplitude variations besides the imposed subaperture. Similarly, in the experimental results of [3], the subaperture did not overlap the edge of the aperture stop. Assuming other amplitude variations can be neglected, we will show that estimating translation from PSF data is equivalent to estimating the phase retrieved from the PSF. We will parameterize that phase by coefficients of a Zernike polynomial description [14], though other basis functions may be similarly useful. We then calculate the Cramer-Rao lower bound for unbiased estimators of translation as a function of the pupil phase aberrations of the system under test. We show how they differ significantly in the SLP and USLP cases. In the SLP case, the ability to assume that each PSF shares the same linear phase contributes subaperture translation information that is unavailable in the USLP case. Translation estimation in the USLP case is entirely dependent on the existence of aberrations higher in order than defocus and astigmatism (those higher than 2nd
radial-degree) in the optical system under test. Through Monte-Carlo simulation, we evaluate translation estimation for a specific population of aberrations to quantify the error in translation estimation for the two cases. In higher fidelity simulations, we performed TTDPR to determine the actual error in translation estimation. Both simulations show that a large defocus aberrations will significantly decrease translation estimation error in the SLP case but not for the USLP case which must rely on the higher order phase errors for translation estimation. Translation estimation is still possible in the USLP case when the higher order aberrations are sufficiently large in magnitude and the inherent wavefront sensing error due to noise is sufficiently low.

## 2.2 Subaperture Translation Estimation Accuracy

In this section, we will estimate the accuracy of translation estimation using PSFs derived from an aberrated optical system using a simplified model suitable for Cramer-Rao lower bounds. For translation estimation methods useful for actual translation computation, see [1–3, 5, 9, 10].

It is presumed that the optical system under test obeys the scalar linear-systems theory for isoplanatic imaging with the aberrations. If the target is an unresolved point, the observed noise-free irradiance for the kth PSF in the detector plane is

\[ I_k(x, y) = |G_k(x, y)|^2, \quad (2.1) \]

where \( G_k(x, y) \) is the field in the detector. Let \( g_k(u, v) \) be a generalized pupil function [15] whose phase, \( \phi_k(u, v) \), is the phase aberration of the system. The phase of the
USLP case will depend on the conditions in place while measuring the PSF, so $\phi_k(u, v)$ is subscripted with PSF index $k$. We will assume that $G_k(x, y)$ is well approximated by a Fraunhofer propagation $P$ of $g_k(u, v)$ with [15]

$$G_k(x, y) = P[g_k] = \frac{A}{\lambda L} \int_{-\infty}^{\infty} g_k(u, v) \exp \left[ -\frac{i 2 \pi \lambda}{L} (x u + y v) \right] du dv,$$  \hspace{1cm} (2.2)

where $A$ is a constant amplitude, $\lambda$ is the wavelength of light, and $L$ is the distance between the exit pupil and detector planes. Any quadratic phase in the detector field due to the propagation is omitted from Eq. (2.2) as it would not alter the PSF intensity. A derivation similar to the following can be done if $P$ is a Fresnel propagation but the result is complicated by the need to track the additional quadratic phase in $u$ and $v$. For simplicity, this derivation will be in the Fraunhofer limit, or equivalently, the quadratic phase term within the Fresnel integral is made part of the phase of $g_k(u, v)$.

Let $A_F(u, v)$ be the amplitude of the fixed field that would have been in the exit pupil had the subaperture not been introduced into the system. This fixed aperture arises from (i) the aperture stop of the system, (ii) the illumination provided by the unresolved target being imaged to yield PSFs, and (iii) modifications to the illumination made by vignetting surfaces of the system which may not be conjugate to the pupil. Further, assume for simplicity that the subaperture is in a plane conjugate to the exit pupil and that it is perfectly re-imaged to the exit pupil yielding a real, non-negative, transmission function $A_S(u, v)$ in the plane of the exit pupil. We model the amplitude of the exit pupil as the product of the translated subaperture and the
Table 2.1: Zernike polynomials in Cartesian coordinates with \( j \), the index of the polynomial and \( n \), the radial-degree.

<table>
<thead>
<tr>
<th>( j )</th>
<th>( n )</th>
<th>( Z_j(u,v) )</th>
<th>Aberration</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td>Piston</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>2u</td>
<td>( u ) linear</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>2v</td>
<td>( v ) linear</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>( \sqrt{3}(2u^2 + 2v^2 - 1) )</td>
<td>Defocus</td>
</tr>
<tr>
<td>5</td>
<td>2</td>
<td>( 2 \sqrt{6}uv )</td>
<td>Astig. 45°</td>
</tr>
<tr>
<td>6</td>
<td>2</td>
<td>( \sqrt{6}(u^2 - v^2) )</td>
<td>Astig. 0°</td>
</tr>
<tr>
<td>7</td>
<td>3</td>
<td>( \sqrt{8}(3u^2v + 3v^3 - 2) )</td>
<td>( v ) coma</td>
</tr>
<tr>
<td>8</td>
<td>3</td>
<td>( \sqrt{8}(3u^3 + 3uv^2 - 2) )</td>
<td>( u ) coma</td>
</tr>
<tr>
<td>9</td>
<td>3</td>
<td>( \sqrt{8}(3u^2v - v^3) )</td>
<td>( v ) trefoil</td>
</tr>
<tr>
<td>10</td>
<td>3</td>
<td>( \sqrt{8}(u^3 - 3uv^2) )</td>
<td>( u ) trefoil</td>
</tr>
<tr>
<td>11</td>
<td>4</td>
<td>( \sqrt{5}[6(u^2 + v^2)^2 - 6(u^2 + v^2) + 1] )</td>
<td>Spherical</td>
</tr>
</tbody>
</table>

amplitude of the fixed field, making the complex field of the generalized exit pupil

\[
g_k(u, v) = A_S(u - s, v - t)A_F(u, v) \exp \{i \phi_k(u, v)\}, \tag{2.3}\]

where \((s, t)\) is the translation in exit pupil coordinates. In general, \( s \) and \( t \) will depend on \( k \) but in this discussion we consider the estimation of translation for an arbitrary but specific PSF, so the index \( k \) is suppressed for brevity.
We will require that $\phi_k(u, v)$ be a slowly-spatially-varying phase function that can be expressed by a Zernike polynomial expansion

$$\phi_k(u, v) = \sum_{j=1}^{J} a_{j,k} Z_j(u, v),$$  \hspace{1cm} (2.4)

where the coefficients $a_{j,k}$ of the $j$th Zernike polynomial $Z_j(u, v)$ are indexed by PSF number $k$ as necessary for the SLP and USLP cases. The assumption that $\phi_k(u, v)$ is slowly varying is usually true in a well-corrected optical system. We use the Noll-ordered polynomials [14, 16], the first 11 terms of which are listed in Table 2.1. We will require that $u$ and $v$ are normalized such that the region of the exit pupil accessed by all subaperture translations is encompassed by the circle $u^2 + v^2 \leq 1$.

In the SLP case, the only parameters of the system changing between PSF acquisitions are the subaperture translations $(s, t)$. The phase coefficients $a_{j,k}$ will all be independent of $k$ and thus the same values are shared amongst all PSFs. In the USLP case, the linear phase terms $a_{2,k}$ and $a_{3,k}$ will be different for each PSF. We next consider two experimental conditions that lead to the linear phase terms becoming unshared. First, consider a detector that shifts in its plane between PSF acquisitions. This motion is equivalent to a modification of the linear Zernike coefficients by the Fourier shift theorem. If the detector plane is shifted by a vector $(\varepsilon_k, \eta_k)$, then from Eq. (2.2) the field in the detector plane is

$$G_k(x - \varepsilon_k, y - \eta_k) = \mathcal{P} \left\{ g_k(u, v) \exp \left[ \frac{i2\pi}{\lambda L} (u\varepsilon_k + v\eta_k) \right] \right\}$$

$$\equiv \mathcal{P} [h_k(u, v)],$$  \hspace{1cm} (2.5)
where \( h_k(u, v) \) is an instance of \( g_k(u, v) \) with the linear phase coefficients \( a_{2,k} \) and \( a_{3,k} \) increased by \( \pi \varepsilon_k / \lambda L \) and \( \pi \eta_k / \lambda L \), respectively. The PSF detected by a shifted detector is identical to the PSF due to \( h_k(u, v) \) with modified linear phase terms since by Eq. (2.5)

\[
|G_k(x - \varepsilon_k, y - \eta_k)|^2 = |P[h_k]|^2. \tag{2.6}
\]

Thus, a detector that moves between PSF acquisitions is equivalent to each PSF having arisen from its own unique linear phase coefficients giving rise to a USLP situation.

Second, suppose instead that the detector was fixed and it was the unresolved target being imaged that was translating. Also, let the amount of the target translation be bounded so that the target stays within an isoplanatic patch of the optical system under test. The effect of target translation inside the isoplanatic patch is a translation of the PSF but by an amount scaled by the magnification of the system. The target translation can be modeled as an exit pupil phase that has appropriately varying linear phase terms. Consequently, an unknown and varying target translation also yields a USLP situation.

The subaperture positions for the \( k \)th PSF can, in principle, be recovered if the PSF for that subaperture position is different from a PSF derived from identical conditions but with an erroneous subaperture position. Since the PSF arises exclusively from the field in the exit pupil, these differences must be due to changes in the amplitude and/or phase of \( g_k(u, v) \). If the fixed aperture has hard-edged features, these variations act as fiducials against which translation can be estimated. This is the case when
the moving subaperture overlaps a hard-edged aperture stop in \( A_F(u, v) \) during one or more PSF acquisitions. However, if the subaperture is smaller than the aperture stop and vignetting of the subaperture by the aperture stop is mild or nonexistent, the utility of amplitude variation is limited since those PSFs will come from regions of the fixed aperture not overlapping any such amplitude fiducials. For this reason we focus our attention on the more difficult case that the effects of \( A_F(u, v) \) on each PSF are negligible and

\[
A_S(u - s, v - t)A_F(u, v) \approx A_S(u - s, v - t). \tag{2.7}
\]

The fixed amplitude for the experiment in [3] was consistent with Eq. (2.7) aside from some very small imperfections, referred to as “digs,” in the surface under test.

The reduced role of fixed aperture features differentiates translation estimation in TTDPR from the established position refinement methods in ptychography for coherent diffractive imaging [9, 10]. In coherent diffractive imaging, specimens generally have a multitude of well defined edges in amplitude, in phase, or in both. These edges yield far-field diffraction patterns that change robustly with small variations in translation or “probe position.” For TTDPR wavefront sensing, there are typically few such amplitude fiducials contributing to the subaperture positions, and it is variations in the phase that must be relied upon for translation estimation.

Assuming Eq. (2.7), Eq. (2.3) can be approximated as

\[
g_k(u, v) = A_S(u - s, v - t) \exp [i\phi_k(u, v)]. \tag{2.8}
\]
Next, consider the PSF intensity that would be expected from an incorrect subaperture translation. Let the subaperture be translated by an additional error vector \((q, r)\), also implicitly dependent on \(k\), as in the exit pupil field

\[ g'_k(u, v) = A_s(u - s - q, v - t - r) \exp[i\phi_k(u, v)]. \]  

(2.9)

The PSF intensity is

\[ I'_k(x, y) = \left| \mathcal{P}\{A_s(u - s - q, v - t - r) \exp[i\phi_k(u, v)]\} \right|^2. \]  

(2.10)

Making the variable substitutions \(u \rightarrow u' + q\) and \(v \rightarrow v' + r\) in the integral implicit in Eq. (2.10), one gets

\[ I'_k(x, y) = \left| \frac{A}{\lambda L} \int_{-\infty}^{\infty} A_s(u' - s, v' - t) \exp[i\phi_k(u' + q, v' + r)] \right|^2 \times \exp\left[ -\frac{i2\pi}{\lambda L} \left( xq + yr \right) \right] \times \exp\left[ -\frac{i2\pi}{\lambda L} \left( xu' + yv' \right) \right] du' dv' \]

\[ = \left| \frac{A}{\lambda L} \int_{-\infty}^{\infty} A_s(u' - s, v' - t) \exp[i\phi_k(u' + q, v' + r)] \right|^2 \times \exp\left[ -\frac{i2\pi}{\lambda L} \left( xu' + yv' \right) \right] du' dv', \]  

(2.11)

where the complex exponential in \(xq + yr\) from the first line disappears because it does not affect the intensity. Since inside Eq. (2.11) is a propagation,

\[ I'_k(x, y) = \left| \mathcal{P}\{A_s(u - s, v - t) \exp[i\phi_k(u + q, v + r)]\} \right|^2. \]  

(2.12)
From Eq. (2.12), we observe that the change in the PSF due to a mis-estimation of the translation by \((q, r)\) is equivalent to a shift in the exit pupil phase \(\phi_k\) by \((-q, -r)\). Evaluating the shift in terms of the Zernike polynomials gives

\[
\phi_k(u + q, v + r) = \sum_{j=1}^{J} a_{jk} Z_j(u + q, v + r).
\] (2.13)

A shifted Zernike polynomial term, \(Z_j(u + q, v + r)\), can be exactly expressed as a weighted sum of un-shifted Zernike polynomials \(Z_j(u, v)\) by endnote 36 in [17]. That proof involves expanding \(Z_j(u + q, v + r)\) as a Taylor series about the point \((u,v)\) in the shift variables \((q, r)\) and various partial derivatives of the un-shifted Zernike polynomial [17]. Since derivatives of the un-shifted Zernike polynomial of an order higher than the polynomial’s radial-degree are zero [14], the Taylor series is finite in length. Of the terms in the Taylor series, those involving first derivatives of Zernike polynomials are themselves each finite series of Zernike polynomials of lower radial-degree [14, 18]. The Taylor series terms involving higher-order partial derivatives can be found by repeatedly differentiating a first derivative series to find the finite series of Zernike polynomials that equals the higher-order partial derivative. Since every term in the Taylor series can be expressed as a finite series in Zernike polynomials, the shifted Zernike polynomial can be written exactly as a weighted sum of un-shifted Zernike polynomials of lesser radial degree. We use this to write the apparent change in a Zernike polynomial as a function of translation estimation error \((q,r)\) as

\[
Z_j(u + q, v + r) = Z_j(u, v) + \sum_{j'=1}^{J} \gamma_{jj'}(q,r) Z_{j'}(u, v),
\] (2.14)
where $\gamma'_{j,j'}(q,r)$ is a matrix of polynomials in $q$ and $r$. The polynomials in the matrix $\gamma'_{j,j'}(q,r)$ for the significant terms up to spherical aberration were evaluated algebraically and are listed in Table 2.2. It is possible to evaluate $\gamma'_{j,j'}(q,r)$ for shifted-Zernike polynomials of arbitrarily high radial-degree but we stopped at spherical aberration for simplicity. Terms that when shifted yield only a constant phase are invisible to phase retrieval, so terms of $\gamma'_{j,j'}(q,r)$ where $j' = 1$ are omitted from Table 2.2. For example, shifted versions of the linear phase term are the same phase un-shifted but with an additional constant:

$$Z_2(s + q, t + r) = 2(s + q) = 2s + 2q$$

$$= Z_2(s, t) + qZ_1(s, t).$$ (2.15)

Substituting Eq. (2.14) into Eq. (2.13) yields

$$\phi_k(u + q, v + r) = \sum_{j=1}^{J} a_{jk}Z_j(u, v) + \sum_{j=1}^{J} a_{jk} \sum_{j'=1}^{J} \gamma'_{j,j'}(q,r)Z_{j'}(u, v)$$

$$= \sum_{j=1}^{J} a_{jk}Z_j(u, v) + \sum_{j'=1}^{J} \sum_{j=1}^{J} a_{j'k} \gamma'_{j',j}(q,r)Z_{j}(u, v)$$

$$= \sum_{j=1}^{J} \left[ a_{jk} + \sum_{j'=1}^{J} a_{j'k} \gamma'_{j',j}(q,r) \right] Z_j(u, v),$$ (2.16)

where we note that the labels $j$ and $j'$ have been permuted in the right-hand part of the sum in the second step to yield the simplification. From Eq. (2.12) it is observed that the change in the PSF due to a translation mis-estimation is equivalent to altering the phase that contributed to the PSF. According to Eq. (2.16), the apparent alteration in
Table 2.2: Non-zero terms of the Zernike translation matrix $\gamma_{j'j}(q,r)$ for $j < 12$.

<table>
<thead>
<tr>
<th>$j'$</th>
<th>$j = 2$</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>$2\sqrt{3}q$</td>
<td>$2\sqrt{3}r$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>$\sqrt{6}r$</td>
<td>$\sqrt{6}q$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>$\sqrt{6}q$</td>
<td>$-\sqrt{6}r$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>$6\sqrt{2}qr$</td>
<td>$3\sqrt{2}q^2 + 9\sqrt{2}r^2$</td>
<td>$2\sqrt{6}r$</td>
<td>$2\sqrt{3}q$</td>
<td>$-2\sqrt{3}r$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>$9\sqrt{2}q^2 + 3\sqrt{2}r^2$</td>
<td>$6\sqrt{2}qr$</td>
<td>$2\sqrt{6}q$</td>
<td>$2\sqrt{3}r$</td>
<td>$2\sqrt{3}q$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>$6\sqrt{2}qr$</td>
<td>$3\sqrt{2}(q^2 - r^2)$</td>
<td>$2\sqrt{3}q$</td>
<td>$2\sqrt{3}r$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>$3\sqrt{2}(q^2 - r^2)$</td>
<td>$-6\sqrt{2}qr$</td>
<td>$-2\sqrt{3}r$</td>
<td>$2\sqrt{3}q$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>$2\sqrt{5}q + 12\sqrt{5}(q^3 + qr^2)$</td>
<td>$2\sqrt{5}r + 12\sqrt{5}(q^2r + r^3)$</td>
<td>$4\sqrt{15}(q^2 + r^2)$</td>
<td>$4\sqrt{30}qr$</td>
<td>$2\sqrt{30}(q^2 - r^2)$</td>
<td>$2\sqrt{10}r$</td>
<td>$2\sqrt{10}q$</td>
</tr>
</tbody>
</table>
phase is equivalent to modifying the Zernike coefficients \( a_{j,k} \) by adding an erroneous contribution of \( \sum_{j' = 1}^{l} a_{j',k} \gamma_{j',j}(q, r) \) to each coefficient.

Suppose there is some calculation involving the PSFs with indices not equal to \( k \) that estimates \( a_{j,k} \). For instance, if we perform a successful TTDPR phase retrieval operation on all the PSFs that are not the \( k \)th PSF, and the \( k \)th PSF comes from the same optical system, then we would expect the TTDPR solution from the other PSFs to be a good estimate of \( a_{j,k} \). This is not possible for \( a_{2,k} \) and \( a_{3,k} \) in the USLP case and we address this later. Also suppose that it is possible to perform a single-PSF phase retrieval on the \( k \)th PSF assuming erroneous values for \((s,t)\) and that the results of this successful phase retrieval solution are polynomial coefficients \( c_{j,k} \) for the \( j \)th Zernike. Note that performing this single-PSF phase retrieval may be ill-posed if the subaperture extent does not allow unique estimation of each of the Zernike coefficients of interest. For this argument, we assume this single-PSF phase retrieval accesses enough of the field to uniquely estimate the necessary Zernike phase coefficients.

In this case, Eq. (2.16) suggests a way to solve for the unknown translation as follows: the error \((q,r)\) in the translations is governed by the nonlinear system of equations:

\[
c_{j,k} - a_{j,k} = \sum_{j' = 1}^{l} a_{j',k} \gamma_{j',j}(q, r)
= \sum_{j' = 2}^{l} a_{j',k} \gamma_{j',j}(q, r) \tag{2.17}
\]
according to Eq. (2.16) and the fact that \( \gamma_{1,j}(q, r) = 0 \) for all \( j \), since \( Z_1(u + q, v + r) = 1 = Z_1(u, v) + 0 \). If this system of equations is solvable for \((q,r)\), the resulting estimates of \((q,r)\) can be added to the erroneous values of \((s,t)\) to re-estimate the true \((s,t)\).

Incidentally, we suspect that the translation estimation problem as posed in this section is similar to translation estimation in some forms of subaperture stitching interferometry (SSI) where (i) the interferometer aperture is confined to translate in the plane of the acquired interferograms, (ii) the translation must be estimated from acquired interferograms and (iii) it is not possible to track the change in constant piston phase between acquisitions. A particular interferogram with unknown translation could then be processed to yield an estimate for \(c_{j,k}\). If \(a_{j,k}\) were estimable from the ensemble of interferograms, the system of equations represented by Eq. (2.17) would exist for the SSI translation estimation problem as it does in the TTDPR case.

Since \( \gamma_{j,j'}(q, r) \) involves nonlinear terms, it is difficult to predict the invertability of the system of equations in Eq. (2.17) for a general aberration. When the expected translation-estimation errors \((q,r)\) are small, however, it is possible to make a linear approximation that is more easily analyzed. Suppose that the guesses for \(s\) and \(t\) used to find \(c_{j,k}\) are sufficiently accurate that \(|q|\) and \(|r|\) are much smaller than unity. The linear terms will then dominate the terms in \( \gamma_{j,j'}(q, r) \) that are quadratic and higher in \(q\) and \(r\). In this case, the matrix of polynomials \( \gamma_{j,j'}(q, r) \) can be split into the portion listed in Table 2.3 that is linear in \(q\) and the portion listed in Table 2.4 that is linear in \(r\) which we label \( \gamma^q_{j,j'} \) and \( \gamma^r_{j,j'} \), respectively. Neglecting the higher order terms in \(q\) and
In Eq. (2.17), we can produce a linear approximation for small \(q\) and \(r\):

\[
c_{jk} - a_{jk} \approx \sum_{j'=2}^{j} a_{j',k}(\gamma_{j',j}^{q} q + \gamma_{j',j}^{r} r)
\]

\[
= qQ_{j} + rR_{j}
\]  

(2.18)

where

\[
Q_{j} = \sum_{j'=2}^{j} a_{j',k}\gamma_{j',j}^{q}
\]

\[
R_{j} = \sum_{j'=2}^{j} a_{j',k}\gamma_{j',j}^{r}
\]  

(2.19)

can be seen as vectors in the space of the differences of \(c_{jk} - a_{jk}\) that depend on the aberrations of the system.

Let \(c\), \(a\), \(Q\) and \(R\) be column vectors representing the values of \(c_{jk}\), \(a_{j}\), \(Q_{j}\) and \(R_{j}\), respectively, starting at \(j = 2\) in the SLP case and \(j = 4\) in the USLP case. The SLP case column vectors begin at \(j = 2\) because \(a_{0}\) in Eq. (2.18) is unknowable. In the USLP case, the vectors begin with \(j = 4\) because \(a_{2,k}\) and \(a_{3,k}\) vary with each PSF and are thus unknowable \textit{a priori}. For simplicity, we have terminated \(\gamma_{j',j}^{q}\), \(\gamma_{j',j}^{r}\), \(c_{jk}\), \(a_{j}\), \(Q_{j}\), \(R_{j}\) and their vector equivalents at spherical aberration but, as with \(\gamma_{j',j}^{q}(q, r)\), they could be extended to include aberrations with arbitrarily high Zernike polynomial index.

Given Eq. (2.18), one can approximate the process of estimating \(q\) and \(r\) given noisy measurements of \(c_{jk}\) using a linear model

\[
c = H\theta + a + \eta
\]  

(2.20)
Table 2.3: Non-zero terms of the linearized Zernike translation matrix $\gamma_{j,j}'$ for $j < 12$.

<table>
<thead>
<tr>
<th>$j'$</th>
<th>$j = 2$</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>2 $\sqrt{3}$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>$\sqrt{6}$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>$\sqrt{6}$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>2 $\sqrt{3}$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>2 $\sqrt{6}$</td>
<td>2 $\sqrt{3}$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>2 $\sqrt{3}$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>2 $\sqrt{3}$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>2 $\sqrt{5}$</td>
<td>2 $\sqrt{10}$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 2.4: Non-zero terms of the linearized Zernike translation matrix $\gamma_{j,j}'$ for $j < 12$.

<table>
<thead>
<tr>
<th>$j'$</th>
<th>$j = 2$</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>2 $\sqrt{3}$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>$\sqrt{6}$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>$-\sqrt{6}$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>2 $\sqrt{6}$</td>
<td>$-2 \sqrt{3}$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>2 $\sqrt{3}$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>2 $\sqrt{3}$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>$-2 \sqrt{3}$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>2 $\sqrt{5}$</td>
<td>2 $\sqrt{10}$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
where $H = [Q \ R]$, the values to be estimated are $\theta = [q \ r]^T$, and $\eta$ represents noise in the measurement $c$ as a result of the single-PSF phase retrieval. For simplicity, we will approximate the error in the phase retrieval measurement as an additive, white, Gaussian noise with variance $\sigma^2_j$ for the $j$th Zernike coefficient and with zero covariance between terms. The results of a phase retrieval measurement for $c_{jk}$ are not really going to be distributed in this manner [19], but it suffices for the purposes of understanding the basic sensitivity of estimators of $q$ and $r$ to aberrations $a_{jk}$. A minimum-variance unbiased estimator [20] can be calculated for Eq. (2.20) yielding an estimator $\hat{\theta}$ for $\theta$. The estimator will have a covariance matrix

$$C_{\hat{\theta}} = (H^T \ C_\eta^{-1} H)^{-1},$$

(2.21)

where $C_\eta$ is the covariance matrix of the noise $\eta$. Since we assume zero covariance between terms, $C_\eta$ and $C_\eta^{-1}$ are diagonal matrices with values of $\sigma^2_j$ and $\sigma^{-2}_j$ along their diagonals, respectively. Eq. (2.21) involves a 2-by-2 matrix inverse and a closed-form evaluation gives

$$C_{\hat{\theta}} = \frac{1}{\text{det}(H^T C_\eta^{-1} H)} \begin{bmatrix} R^T C_\eta^{-1} R & -Q^T C_\eta^{-1} R \\ -R^T C_\eta^{-1} Q & Q^T C_\eta^{-1} Q \end{bmatrix},$$

(2.22)

from which the individual variances for estimators of $q$ and $r$ can be extracted and written more simply as
\[
\text{var}(\hat{q}) = \left[ Q^T C_\eta^{-1} Q - \frac{(Q^T C_\eta^{-1} R)^2}{R^T C_\eta^{-1} R} \right]^{-1},
\]

\[
\text{var}(\hat{r}) = \left[ R^T C_\eta^{-1} R - \frac{(Q^T C_\eta^{-1} R)^2}{Q^T C_\eta^{-1} Q} \right]^{-1}. \tag{2.23}
\]

If the variances are small for a particular aberration and wavefront sensing accuracy, then translation estimation is expected to be robust. Since the coordinates \( u \) and \( v \) have been normalized as fractions of the radius of the circle \( u^2 + v^2 \leq 1 \) and are hence unitless, \( \text{var}(\hat{q}) \) and \( \text{var}(\hat{r}) \) are similarly unitless. A minimum-variance unbiased estimator of a linear model is efficient [20], so, assuming Eq. (2.20) and our simplified model for wavefront sensing error are true, \( \text{var}(\hat{q}) \) and \( \text{var}(\hat{r}) \) are Cramer-Rao lower bounds [20] for any unbiased estimator of translation.

The scalar values \( Q^T C_\eta^{-1} Q, R^T C_\eta^{-1} R \) and \((Q^T C_\eta^{-1} R)^2\) are always positive, so Eq. (2.23) has the lower bounds

\[
\text{var}(\hat{q}) \geq (Q^T C_\eta^{-1} Q)^{-1},
\]

\[
\text{var}(\hat{r}) \geq (R^T C_\eta^{-1} R)^{-1}. \tag{2.24}
\]

Thus a necessary condition for robust estimation of translations along the \( u \) and \( v \) coordinates is that the optical system have aberrations that yield \( Q^T C_\eta^{-1} Q \) and \( R^T C_\eta^{-1} R \) large in comparison to 1. To ensure robust estimation it is also required that \( (Q^T C_\eta^{-1} R)^2 \) is small in comparison to the product of \( Q^T C_\eta^{-1} Q \) and \( R^T C_\eta^{-1} R \), according to Eq. (2.23).
2.3 Subaperture Translation Estimation with Shared Linear Phase

In the SLP case for which the terms $a_{2,k}$ and $a_{3,k}$ are inferable from PSFs with all indices but $k$, evaluating the components of Eq. (2.19) for arbitrary aberrations $a_j$ yields

$$Q^T = [2 \sqrt{3} a_4 + \sqrt{6} a_6 + 2 \sqrt{5} a_{11}, \sqrt{6} a_5, 2 \sqrt{6} a_8, 2 \sqrt{3} (a_7 + a_9), 2 \sqrt{3} (a_8 + a_{10}), 0, 2 \sqrt{10} a_{11}]$$

$$R^T = [\sqrt{6} a_5, 2 \sqrt{3} a_4 - \sqrt{6} a_6 + 2 \sqrt{5} a_{11}, 2 \sqrt{6} a_7, 2 \sqrt{3} (a_8 - a_{10}), 2 \sqrt{3} (a_9 - a_7), 2 \sqrt{10} a_{11}, 0].$$

The components of Eq. (2.24) in the SLP case are therefore

$$Q^T C^{-1}_\eta Q = \sigma_2^{-2} (2 \sqrt{3} a_4 + \sqrt{6} a_6 + 2 \sqrt{5} a_{11})^2 + 6 \sigma_3^{-2} a_5^2 + 24 \sigma_4^{-2} a_8^2 + 12 \sigma_5^{-2} (a_7 + a_9)^2 + 12 \sigma_6^{-2} (a_8 + a_{10})^2 + 40 \sigma_7^{-2} a_{11}^2,$$

$$R^T C^{-1}_\eta R = 6 \sigma_2^{-2} a_5^2 + \sigma_3^{-2} (2 \sqrt{3} a_4 - \sqrt{6} a_6 + 2 \sqrt{5} a_{11})^2 + 24 \sigma_4^{-2} a_7^2 + 12 \sigma_5^{-2} (a_8 - a_{10})^2 + 12 \sigma_6^{-2} (a_9 - a_7)^2 + 40 \sigma_7^{-2} a_{11}^2.$$ 

Eq. (2.26) indicates that the lower bounds for the variance, Eq. (2.24), are inversely proportional to a series of sums depending on 2nd and higher radial-degree aberrations adding in quadrature. If there are no phase aberrations in the exit pupil, $Q = R = 0$ and the lower bounds in Eq. (2.24) are infinite. Hence, subaperture translation estimation is not practical for systems with completely unaberrated exit pupils.
or those with only piston or linear phase terms. If the translations were uncertain prior to TTDPR in these cases, no additional information about what region of the subaperture plane was unaberrated could be gained through TTDPR. Phase retrieval would indicate just that an unspecified region of the subaperture plane was aberration-free.

The expected variance and Cramer-Rao lower bound in Eq. (2.23) also require that the square of

\[ Q^T C_\eta^{-1} R = (\sigma_2^2 + \sigma_3^2)a_5(6 \sqrt{2}a_4 + 2 \sqrt{30}a_{11}) + 6(\sigma_2^2 - \sigma_3^2)a_5a_6 + 24\sigma_4^2 a_7a_8 \\
+ 12(\sigma_5^2 + \sigma_6^2)(a_8a_9 - a_7a_{10}) + 12(\sigma_5^2 - \sigma_6^2)(a_7a_8 - a_9a_{10}) \]

be small compared to the product of \( Q^T C_\eta^{-1} Q \) and \( R^T C_\eta^{-1} R \) in Eq. (2.26) for a low variance of the bound. If the wavefront sensing errors in linear phase and astigmatism are independent of rotation, then \( \sigma_3 = \sigma_2 \) and \( \sigma_6 = \sigma_5 \), and Eq. (2.27) simplifies to

\[ Q^T C_\eta^{-1} R = 4\sigma_2^2 a_5(3 \sqrt{2}a_4 + \sqrt{30}a_{11}) + 24\sigma_4^2 a_7a_8 + 24\sigma_5^2(a_8a_9 - a_7a_{10}). \]

For highly aberrated systems, which are described by many Zernike coefficients that are large in magnitude, \( Q^T C_\eta^{-1} Q \) and \( R^T C_\eta^{-1} R \) are likely large as well, and thus the lower bounds in Eq. (2.24) on translation estimation error are small. These bounds are more problematic on nearly-unaberrated systems where \( Q^T C_\eta^{-1} Q \) and \( R^T C_\eta^{-1} R \) will be much smaller; thus we concentrate on that case.

Cramer-Rao analysis of defocus-diverse phase retrieval shows that wavefront sensing in the presence of noise performs better when defocus has been induced in the system
[21, 22], so it can be expected that some defocus has been induced to improve wavefront sensing. When this intentionally added defocus dominates the other aberrations, the expected translation estimation error in the SLP case is approximately

$$\sigma_2^{-2}\text{var}(\hat{q}) \approx \frac{1}{12a_4^2},$$

$$\sigma_3^{-2}\text{var}(\hat{r}) \approx \frac{1}{12a_4^2},$$

(2.29)

according to Eqs. (2.26), (2.27) and (2.23), under the assumption that $a_j \ll 1$ for $j \neq 4$ and that the variance of wavefront sensing error is comparable for all aberrations. For example, if the aberrations are dominated by a defocus of $a_4 = 4.62$ radians RMS, equivalent to 2.55 waves center-to-edge of the circle $u^2 + v^2 \leq 1$, Eq. (2.29) is

$$\sigma_2^{-2}\text{var}(\hat{q}) = \sigma_3^{-2}\text{var}(\hat{r}) \approx 3.91 \times 10^{-3}\text{rad}^{-2}.$$  

(2.30)

In general, for any single aberration that dominates all of the others, the variance of translation is inversely proportional to the square of the dominating aberration with the constant of proportionality determined by the factors in Eq. (2.26).

If there is not just one dominating aberration such as defocus, Eqs. (2.26), (2.27) and (2.23) must be evaluated for the specific phase aberrations of the system and wavefront sensing error variances to predict translation estimation performance. If the specific aberration of interest is unknown, these equations can be evaluated over a specified population of aberrations to establish expected behavior. This is done in Section 2.5 for a specific population of aberrations in a Monte-Carlo experiment.
2.4 Subaperture Translation Estimation with Unshared Linear Phase

In the USLP case, \( c \), \( a \), \( Q \) and \( R \) become column vector representing the values of \( c_{jk} \), \( a_j \), \( Q_j \) and \( R_j \) for \( 4 \leq j \leq 11 \) since \( a_{2,k} \) and \( a_{3,k} \) are unshared and unknown. For this case, evaluating Eq. (2.19) yields

\[
Q^T = \begin{bmatrix}
2 \sqrt{6} a_8, 2 \sqrt{3} (a_7 + a_9), 2 \sqrt{3} (a_8 + a_{10}), 0, 2 \sqrt{10} a_{11}
\end{bmatrix},
\]

\[
R^T = \begin{bmatrix}
2 \sqrt{6} a_7, 2 \sqrt{3} (a_8 - a_{10}), 2 \sqrt{3} (a_9 - a_7), 2 \sqrt{10} a_{11}, 0
\end{bmatrix}.
\] (2.31)

Therefore

\[
Q^T C_\eta \eta^{-1} Q = 24 \sigma_4^{-2} a_8^2 + 12 \sigma_5^{-2} (a_7 + a_9)^2 + 12 \sigma_6^{-2} (a_8 + a_{10})^2 + 40 \sigma_8^{-2} a_{11}^2,
\]

\[
R^T C_\eta \eta^{-1} R = 24 \sigma_4^{-2} a_7^2 + 12 \sigma_5^{-2} (a_8 - a_{10})^2 + 12 \sigma_6^{-2} (a_9 - a_7)^2 + 40 \sigma_7^{-2} a_{11}^2,
\] (2.32)

and

\[
Q^T C_\eta \eta^{-1} R = 24 \sigma_4^{-2} a_7 a_8 + 12 (\sigma_5^{-2} + \sigma_6^{-2}) (a_8 a_9 - a_7 a_{10}) + 12 (\sigma_5^{-2} - \sigma_6^{-2}) (a_7 a_8 - a_9 a_{10}).
\] (2.33)

The aberrations that contribute to improved subaperture estimation in the USLP case are those of 3rd radial degree and above. Notably missing are the 2nd radial degree terms: defocus and the two astigmatisms. This is due to the first two entries in \( Q \) and \( R \) in the SLP case, which include the 2nd radial-degree terms, being absent from the \( Q \) and \( R \) of the USLP case. These terms are excluded from Eq. (2.20) in the USLP case because \( a_{2,k} \) and \( a_{3,k} \) are unknown and thus \( c_{2,k} - a_2 \) and \( c_{3,k} - a_{3,k} \) are unusable equations
of Eq. (2.18). In less mathematical terms, an errant subaperture translation samples a mean wavefront slope different than the mean slope across the subaperture in the correct translation. If the 2nd radial-degree aberrations are non-zero, the difference in mean wavefront slope has a component linear in pupil coordinates and that yields a translation of the PSF relative to where it would be expected, given \( a \). This PSF shift constitutes information about the subaperture translation and it is due to the existence of the 2nd radial-degree aberrations. However, this shift is indistinguishable from unknown detector or target motion since that also produces a translation of the PSF. Thus, the 2nd radial-degree aberrations in the optical system contribute no subaperture translation information if the detector or target are also translating. We observe that it is necessary to have aberrations of the exit pupil higher in order than defocus and astigmatism to estimate translation in a system for which the linear phase terms vary unpredictably between PSFs or, equivalently, when the detector or target are undergoing unknown translation. Similarly, in the USLP case with a dominant defocus or astigmatism aberration, it is much more difficult to estimate the variance of translation estimation without specifying the smaller aberrations of higher radial degree. In the next section, we turn to Monte-Carlo analysis to evaluate the behavior of the variance of translation estimation with a particular population of aberrations.

### 2.5 Monte-Carlo Experiments

We compared the Cramer-Rao lower bounds on translation estimation in the SLP and USLP cases for a specific aberration population using a Monte-Carlo exper-
iment. Five hundred random aberration coefficient vectors \( a_j \) were composed by drawing independent and normally distributed random numbers from distributions having standard deviations of 10, 0.2, 0.1, 0.1 and 0.3 radians RMS for the linear, astigmatism, coma, trefoil and spherical aberration terms, respectively. For each vector, the defocus term was assumed to be the arbitrary constant \( a_4 = 4.62 \) radians RMS used in Section 2.3 and the piston term was assumed zero. The mean wavefront aberration averaged across the population of aberration vectors was 4.64 radians RMS with linear phase excluded, but just 0.429 radians RMS with defocus phase also excluded. An example exit pupil phase synthesized from one such aberration vector is plotted in Fig. 2.1 (a) and in Fig. 2.1 (b) with the defocus subtracted. In both figures, the phase has been plotted over the circle \( u^2 + v^2 \leq 1 \).

In addition to an aberration coefficient vector, evaluation of Eq. (2.23) requires the variance of wavefront sensing error by Zernike coefficient for a single-PSF retrieval. Such variances will depend on the particular aberration coefficient vector but we assume that the variance of wavefront sensing error for a given Zernike coefficient averaged over the ensemble of coefficient vectors is adequate. In Sect. 2.6, a particular simulated wavefront sensing experiment is described for which ensemble-averaged variances have been estimated by a Monte-Carlo process described in that section. In this section, we take these mean wavefront-sensing error variances as a given and they are listed in Table 2.5. To the extent that the astigmatism variances are larger than that of defocus and the defocus variance is larger than the coma terms, these mean
Figure 2.1: (a) Sample exit pupil phase with linear phase subtracted and a colorbar in radians, (b) sample wavefront phase having also had the defocus subtracted with a colorbar in radians, (c) map of the subaperture/pupil plane with a colorbar indicating number of subaperture positions overlapping the pupil and (d) simulated noisy PSF from the sample exit pupil phase and one of the subaperture positions.
Table 2.5: Mean wavefront sensing error variances estimated from the simulated phase retrieval experiment in Sect. 2.6.

<table>
<thead>
<tr>
<th>Term</th>
<th>Mean Value (rad²)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\sigma_2^2$</td>
<td>$9.44 \times 10^{-5}$</td>
</tr>
<tr>
<td>$\sigma_3^2$</td>
<td>$9.49 \times 10^{-5}$</td>
</tr>
<tr>
<td>$\sigma_4^2$</td>
<td>$9.15 \times 10^{-5}$</td>
</tr>
<tr>
<td>$\sigma_5^2$</td>
<td>$1.58 \times 10^{-4}$</td>
</tr>
<tr>
<td>$\sigma_6^2$</td>
<td>$1.55 \times 10^{-4}$</td>
</tr>
<tr>
<td>$\sigma_7^2$</td>
<td>$1.84 \times 10^{-5}$</td>
</tr>
<tr>
<td>$\sigma_8^2$</td>
<td>$1.83 \times 10^{-5}$</td>
</tr>
</tbody>
</table>

The variances are consistent with the variances calculated numerically in [19], though the exact values and the experiment simulated differ.

For each of the coefficient vectors, the value of $\text{var}(\hat{q})$ was evaluated in the SLP and USLP cases using Eq. (2.23) assuming the wavefront sensing error variances in Table 2.5. A log-histogram of $\text{var}(\hat{q})$ values over the population of 500 vectors is shown in Fig. 2.2. Since the aberration statistics chosen do not have an orientation bias, a log-histogram of $\text{var}(\hat{r})$ would yield nearly identical distributions. Of particular note, the SLP cases are clustered around an average value of $3.58 \times 10^{-7}$. Substituting $\sigma_4^2$ from Table 2.5 into the defocus-dominant limit in Eq. (2.30) yields a prediction for $\text{var}(\hat{q})$ of $3.57 \times 10^{-7}$. This confirms that Eq. (2.30) is useful for predicting the expected variance of translation estimation in this defocus-dominated SLP case without requiring a Monte-Carlo experiment, just as Section 2.3 noted. The maximum variance of the population is about 1.3 times the mean variance and the
mean variance is about 2.9 times the minimum. The square-root of the mean variance yields a standard deviation of the subaperture translation estimation error of 0.03% of the width of the reference circle described by $u^2 + v^2 \leq 1$.

In contrast, the histogram for the USLP case in Fig. 2.2 is far more spread out and denotes variances much larger than the SLP case. The USLP variances range over almost four orders of magnitude with a maximum 44 times that of the mean and the mean variance is 214 times the minimum. The mean variance of $6.75 \times 10^{-5}$ in the USLP case is about 189 times larger than the mean in the SLP case. The standard deviation of subaperture translation estimation error corresponding to the mean variance is 0.41% of the width of the circle $u^2 + v^2 \leq 1$ while the maximum variance corresponds to a standard deviation of 2.7% of the width of the circle. A
worse-case translation error of 2.7% could be unacceptably large for some demanding applications.

Since Eqs. (2.26)-(2.28) and Eqs. (2.32) and (2.33) are linear in the inverse of the wavefront sensing error variances, Eq. (2.23) would be linear in any constant factor change that was applied to all wavefront sensing error variances equally. Supposing that the wavefront sensing errors in Table 2.5 were each 100 times larger, the variance statistics calculated in this section would also be 100 times larger and the standard deviations larger by a factor of 10. In this case, the worst aberration vector would yield a translation error of 27% which is almost certainly unacceptably large, since it would be very unclear to what part of the exit pupil each PSF’s phase measurements corresponded.

2.6 Comparison with Phase Retrieval Simulation

Next, we used Monte-Carlo analysis to examine whether the bound suggested by Eq. (2.23) is consistent with the variance of translation estimation error of phase retrieval in simulation. For each of the aberration coefficient vectors, 30 TTDPR experiments were simulated for the SLP and USLP cases with varying noise realizations. In each of these TTDPR experiments, seven PSFs were simulated from varying positions of a circular subaperture defined by \( u^2 + v^2 \leq 0.25 \). The subaperture positions were arranged as indicated by the map of subaperture overlap in Fig. 2.1 (c). This arrangement of the subapertures just inscribes the circle \( u^2 + v^2 \leq 1 \). To each of the PSFs, different realizations of simulated Poisson- and Gaussian-distributed detector
noise were applied. It was assumed that the peak pixel received 7,000 photoelectrons and there was a read noise of 30 photoelectrons. The simulated detector pixels were Nyquist sampled in intensity for a notional circular aperture defined by $u^2 + v^2 \leq 1$. No such circular aperture stop was simulated in the fixed aperture, though, to be consistent with the assumptions of Sect. 2.2 that specify no amplitude fiducials. TTDPR computation using the method of [3] was applied to each of the simulated data to estimate exit pupil phase and subaperture translation assuming the SLP and USLP cases.

TTDPR is a nonlinear-optimization technique that locates a minimum of an error metric given a starting point for the unknown phase aberrations and translations. For this analysis, we wished to assess the translation estimation accuracy of the TTDPR algorithm limited primarily by detector noise and independent of stagnation issues associated with converging to the correct exit pupil phase estimate. Therefore, we chose as starting aberration coefficient values the actual aberration coefficients used in the simulated data. Doing the same with the translation estimates, however, does not yield representative translation estimation errors in our testing. If we start the minimization process at the true subaperture translations, it has a tendency to find a minimum of the error metric near the true solution even when deeper minima of the metric exist somewhat further away from the true solution. In metrology where the true translation estimation is unknown, it is the translation estimation errors associated with these deeper minima that are of interest since these are likely to be found by the minimization process. To increase the chance of locating these deeper minima, twenty phase retrievals were done for each PSF noise realization with different initial
subaperture translation estimates perturbed from the true values by a Gaussian-distributed random error having a variance of $2.25 \times 10^{-2}$. Of these twenty retrievals, the retrieval with the lowest final error metric was selected as the result for a given PSF noise realization.

We evaluated $\text{MSE}(s)$, the mean squared difference between the retrieved translations and the actual translations present in the simulated PSF data, over the selected phase retrievals from each of the 30 noise realizations for each aberration vector. If the subaperture estimation in the TTDPR performed in simulation is unbiased, it would be expected that

$$\text{MSE}(s) \gtrsim \text{var}(\hat{\eta}),$$

(2.34)

where $\text{var}(\hat{\eta})$ is calculated using Eq. (2.23) since it is a Cramer-Rao lower bound. The validity of this bound is contingent on the assumptions of Sect. 2.2 including the approximation used to model the effects of detector noise on single-PSF phase retrieval. In particular, it was assumed that the noise in the PSF induces errors in the retrieved Zernike coefficients equivalent to uncorrelated, independently-distributed Gaussian noise. Though TTDPR applied to PSFs corrupted by Gaussian and Poisson noise yield different error statistics, we take the variance of actual single-PSF phase retrieval to be an estimate for the variances $\sigma^2_j$ in $C_\eta$. We assessed the variance of single-PSF phase retrieval through additional Monte-Carlo simulation experiments involving one PSF arising from an untranslated subaperture centered in the pupil. For each of the $\sigma^2_j$ in $C_\eta$, a Monte-Carlo experiment was performed to estimate the
variance of phase retrieval wavefront sensing error with respect to wavefront sensing error in just the $j$th Zernike term in the presence of PSF noise. Thirty noisy PSFs were synthesized for each of the randomly drawn aberration coefficient vectors used in the main simulation described earlier. For each noisy PSF and selected aberration type, phase retrieval was performed to estimate the $j$th Zernike coefficient while the other aberrations had their coefficients assumed to be the true value. Subaperture translation was not estimated and the phase retrieval algorithm used the correct translation. The squared errors of the wavefront sensing error for each Zernike term were averaged over the population of aberration coefficient vectors and noise realizations. These results are listed in Table 2.5. By estimating the error in each Zernike term individually, we excluded from our variance estimates the effects of the covariance of errors between different Zernike terms. This made our experimentally estimated variances consistent with the assumption of no covariance made in Section 2.2.

In Fig. 2.3, the value of $\text{MSE}(s)$ for each aberration coefficient vector is plotted versus $\text{var}(\hat{q})$ assuming the mean covariances in Table 2.5 for both the SLP and USLP case. Also plotted is a line representing the Cramer-Rao lower bound in Eq. (2.34). Each of the five hundred aberration coefficient vectors yielded a $\text{MSE}(s)$ consistent the Cramer-Rao lower bound inequality in Eq. (2.34) for the SLP case. In the USLP case, three of vectors out of the five hundred had $\text{MSE}(s)$ slightly less than the bound predicts, the least of which had a $\text{MSE}(s)$ 12% smaller than the bound. Still, the vast majority of aberration coefficient vectors had estimation errors consistent with the bound. This is remarkable given the significant approximations made in Sect. 2.2.
Figure 2.3: Log-log scatter plot of MSE(s) of simulated TTDPR experiments for a given aberration vector versus the $\sigma^{-2}\text{var}(\hat{q})$ for the aberration for both the SLP and USLP cases. Includes line representing the predicted Cramer-Rao lower bound of Eq. (2.34).
and simplification of the wavefront sensing noise to a Gaussian model that neglects covariance. It is also impressive that the translation errors of the simulated TTDPR are correlated to the simplified variance model in Eq. (2.23) as they are over the 3.8 orders of magnitude encompassed by Fig. 2.3. We found that 89% of the aberration vectors had a MSE(s) between $\text{var}(\tilde{q})$ and $4 \times \text{var}(\tilde{q})$ suggesting that Eq. (2.23) is a useful predictor of translation error for the actual phase retrieval simulated and not just an approximate lower bound. Since the TTDPR algorithm was estimating both the subaperture translations and the phase aberrations shared between PSFs, it would not expected to achieve the Cramer-Rao lower bound for estimating translation alone. That it approximately follows the bound Eq. (2.23) but with a constant scale factor is unexpected but potentially useful for predicting the behavior of real measurements that involve maximum-likelihood estimators rather than minimum variance estimators.

Next, we compared MSE(s) values in the experiment involving simulated TTDPR between the USLP and SLP cases. The MSE(s) across the population in the USLP case was $1.51 \times 10^{-4}$ and 230 times that of the MSE in the SLP case. The ratio predicted by comparing the mean values of $\text{var}(\tilde{q})$ between USLP and SLP cases in Section 2.5 was 189 so the actual ratio of 230 is only slightly more than expected. The error between 189 and 230 of just 21% suggests that the theory outlined in Section 2.5 is useful for comparing the behavior of SLP and USLP cases of real systems when averaged across an ensemble of potential aberrations.
Finally, we compared the wavefront sensing accuracy by finding the mean of the RMS error between the retrieved aberrations and the true aberrations for each simulated phase retrieval. These RMS errors were averaged to estimate the expected wavefront sensing error over the population of aberrations. In the SLP case, the mean RMS error was $7.6 \times 10^{-3}$ radians while in the USLP case it was $1.1 \times 10^{-2}$ radians. If the wavefront sensing requirements for the simulated wavefront sensing experiment were one-hundredth wave RMS, or $6.2 \times 10^{-2}$ radians RMS, TTDPR in both the SLP and USLP cases could be expected to succeed for the selected population of aberrations.

## 2.7 Conclusion

In transverse-translation diversity phase retrieval (TTDPR), point spread functions (PSFs) of an optical system are acquired for use in a phase retrieval algorithm to estimate the aberrations of that system. A subaperture in a plane conjugate to the exit pupil is translated to a different position for each PSF. In cases where the knowledge of this translation is insufficiently accurate or unavailable, it must be estimated from the PSF data. Misestimation of the translation of the subaperture leads to misestimation of the aberrations, so the accuracy of this translation estimation is critical. In Section 2.2, we developed analytic expressions for the variance of a minimum-variance unbiased estimator and a Cramer-Rao lower bound for subaperture translation estimation assuming various approximations. Notable among these approximations, we assumed that the exit pupil does not have fixed amplitude features that would
aide subaperture translation estimation. For unobscured imaging systems, most PSFs will derive from subaperture translations that do not overlap an amplitude feature that would aide translation estimation. The Cramer-Rao lower bound depends on wavefront sensing quality and the composition of phase aberrations of the system. Which aberrations contribute to the variance varies according to whether each PSF arose from identical linear phase terms or dissimilar linear phase terms due to unknown translations of the detector acquiring the PSF or of the target being imaged. We refer to these two cases the shared linear phase (SLP) and unshared linear phase (USLP) cases, respectively.

The expressions for the bound on translation estimation vary significantly between the SLP case evaluated in Section 2.3 and USLP cases evaluated in Section 2.4. The aberrations with radial degree of 2 contribute to lowering the variance of translation estimation in the SLP case but do not directly contribute in the USLP case according to the model of Section 2.2. In the common situation that a significant defocus has been introduced to improve wavefront sensing, USLP is at a significant disadvantage because this additional defocus does not directly aid translation estimation as it does in the SLP case. We note, though, that the model in Sect. 2.2 does not take into account the variation of wavefront sensing error $\sigma$ due to the composition of aberrations. In practice, increasing the magnitude of defocus and astigmatism will likely decrease wavefront sensing error for all aberrations and thus aide translation estimation even in the USLP case.
Evaluating whether translation estimation is viable for a particular experiment requires knowledge of the noise inherent in the wavefront sensing process and predictions for the actual aberrations expected in the subaperture plane. For instance, if the aberrations are dominated by large coma or spherical aberration terms, translation estimation may perform acceptably well in both SLP and USLP cases, depending on the wavefront sensing error. In Section 2.5, we chose a specific large population of randomly generated aberrations and evaluated the predictions of the theory to understand and evaluate the differences between the SLP and USLP cases assuming one had an efficient estimator that met the Cramer-Rao lower bound. From this analysis, we observed than the mean value of translation variance over this population of aberrations should be about 189 times larger in the USLP case than in the SLP case. In Section 2.6, we simulated TTDPR using simulated PSFs and actual TTDPR algorithm evaluations. Translation estimate errors from this actual phase retrieval algorithm would not be expected to meet the Cramer-Rao lower bound but they should not be lower than the Cramer-Rao bound derived in Section 2.2. This was the case for all of the aberrations examined in the SLP case and virtually all of the aberrations in the USLP case, suggesting that the bound was plausible. The 0.6% of the aberrations that violated the bound in the USLP case did so by at most 12% which is understandable given the nature of the approximations in the bound. For this set of experiments, the mean squared error of translation was $1.51 \times 10^{-5}$ in the USLP case where the units of translation are radii of the circle $u^2 + v^2 \leq 1$. This mean squared error was 230 times larger than the average SLP case or, alternately, more than 15 times larger
if the RMS error values were compared. Our comparison between theory-predicted Cramer-Rao bound for the variance and the actual mean squared error achieved by TTDPR suggested a strong correlation, as shown by Fig. 2.3. We found that 89% of the aberrations yielded translation variances between the predicted Cramer-Rao bound and four times the bound over a range of bound values encompassing 3.8 orders of magnitude. This level of agreement is remarkable given that the model for the Cramer-Rao bound does not take into account the variation of wavefront sensing error with aberration composition or the covariance of the wavefront sensing errors between the sensed wavefront coefficients. The simulation in Section 2.5 demonstrated a phase aberration sensing error better than one-hundredth wave RMS in both the SLP and USLP cases. On average, the USLP cases exhibited slightly higher wavefront sensing error than the SLP cases due to the effects of additional subaperture translation missettimation.

It is likely that these results derived for TTDPR have an analog in conventional subaperture stitching interferometry when the position of the subaperture have errors that must be estimated from interferogram data.

Finally, we return to evaluating TTDPR for use with the longwave channel of the Near-Infrared Camera (NIRCam) on the James Webb Space Telescope. The USLP case arises in ground testing involving external pupil illumination as well as when the internal light sources of NIRCam are used for ground testing or on-orbit testing of NIRCam. The results of Sect. 2.4 indicate that additional defocus does not directly aide translation estimation in the USLP case, which must rely soley on higher-order
aberrations. Consequently, the lack of weak defocus lenses in the longwave channel does not significantly hinder translation estimation for that case. If weak lenses were present, the improvement would only be that due to the indirect effect of increasing wavefront sensing accuracy. Instead, it is the 3rd radial degree aberrations like coma and trefoil in NIRCam that directly assist TTDPR. Fortunately, the NIRCam system develops such aberrations when the Lyot stops are introduced. This is because the Lyot stop is patterned on a prism substrate that shifts the field of view of NIRCam when selected in the filter wheel [12]. NIRCam involves a collimating lens assembly and a camera lens assembly whose individual aberrations cancel most ideally when the prism is not introduced between them. We have observed that introduction of the prism element unbalances the overall aberration correction in NIRCam enough for successful USLP TTDPR. In cases where NIRCam is presented with an input beam having well-defined fixed amplitude features such as the segment boundaries of the telescope’s primary mirror, estimation is likely to be even better than that achieved using pupil phase alone.

In conclusion, we have computed a Cramer-Rao lower bound for subaperture translation estimation in TTDPR aberration retrieval when amplitude fiducials are unavailable in the pupil. This estimation is essential to phase measurements when the translation must be estimated using the PSF data. The bound shows that Zernike coma, trefoil and those terms of higher radial degree are critical to estimation when the linear phase terms of the pupil are effectively unknown and also vary from PSF to PSF. Monte-Carlo experiments involving simulations of TTDPR were consistent
with the lower bound and also correlated with the bound, suggesting that it could approximate the average MSE expected. We believe that the insights provided by this approximate but analytic bound will be invaluable for designing robust optical metrology experiments using TTDPR in difficult situations.

References


3

Unknown Transverse-Translation Diversity Phase Retrieval

3.1 Introduction

As introduced in Chapters 1.1 and 2, ptychography [1–4] and transverse-translation diversity phase retrieval (TTDPR) [5–9] are techniques for inferring the unknown complex-valued field in a plane from intensity measurements collected in another plane. The unknown and measurement planes are typically connected by a propagation such as a Fraunhofer or Fresnel transform but, in most applications, only certain regions of the unknown field are allowed to propagate for any particular intensity measurement. Which regions are selected is controlled by translating some feature within the unknown field or the illumination prior to it. This translating feature takes different forms in various ptychography applications. In coherent diffractive imaging (CDI), the unknown field is due to a coherently-illuminated transmissive object. Ptychography has been demonstrated with partially coherent illumination
and with thick unknown objects where volume information is resolved, but we will consider just the coherent two-dimensional unknown case. If the coherent beam illuminates only a small region of the object, then this beam may be the feature translated; alternately the object may be translated through the beam. Intensities are collected for various translations of the beam illuminating different and overlapping regions of the object. Alternately, if the object is fully illuminated, a probe or sub-aperture with its own transmission function can be imposed before or after the object and translated instead of the illumination. Though ptychography and TTDPR solve similar mathematical problems, TTDPR uses the language of phase retrieval for wavefront sensing and the iterative nonlinear optimization of an error metric like [10, 11]. Algorithms associated with ptychography can be used for wavefront sensing [12] and the methods of TTDPR can be used to solve CDI problems often associated with ptychography [5]. This chapter utilizes the TTDPR approach to solve a wavefront sensing problem but the challenges faced would be similar to ptychography for CDI.

For the optical metrology of imaging systems [6, 8, 9, 12–14], the generalized exit pupil function [15], rather than an object, is the unknown complex field of interest. A subaperture or mask is introduced in a plane ideally conjugate to the pupil. It is translated to sample various overlapping regions of the pupil while point spread functions (PSFs) are collected by an array detector near a focal plane. As in other applications, an algorithm is applied to find the complex field consistent with all the measured PSF intensities. The classic ptychography algorithms include iterative transform algorithms [16] such as those derived from the ptychographic iterative
engine (PIE) [1, 2, 4, 17] and the nonlinear optimization technique of TTDPR [5–7].

For the simplest algorithms, both the transmission function of the subaperture or probe and the translation of that feature must be known with high accuracy [5, 18]. The exact translation distance is not always exactly known, though, and often only an approximate value is available. Here we call this approximate translation, known from the design of the experiment, the prior translation estimate (PTE). In CDI applications, errors in PTE can cause considerable degradation in the estimated fields [3, 5, 7, 17, 18] unless corrected. Similarly, errors in PTE degrade the accuracy of the retrieved exit pupil field in the optical metrology application. For this reason, some algorithms have the capability of refining subaperture translation estimates given the PTE and measured intensity data [5–7, 19–22]. In TTDPR, a parameterized model for the exit pupil field is fit to the measured PSFs by a gradient-based nonlinear optimization algorithm [5–7]. Translation estimation is accomplished by including the translation values in the parameters that the optimizer varies to improve the correspondence between modeled and measured PSFs [5–7]. Similarly, PIE algorithms have been developed that infer translation using gradient-free search techniques [19, 20] and the serial-correlation method [22]. For some ptychography applications, the errors in PTE can be parameterized using a simplified two-parameter drift model and recovered using gradient-free optimization [21]. The gradient-based nonlinear optimization for translation estimation used in TTDPR can also be used to periodically update translations used by a PIE algorithm [23].
The algorithms mentioned above depend on being supplied with PTE so that they can initially approximate the unknown complex field. The PTE is used as an initial estimate for the translation. Later stages of the algorithms then refine joint estimates of the translation and the complex field using the measured intensity or PSF data. With approximately correct PTE for the initial estimate, the final result of translation estimation has been observed to be robust [5, 7, 20, 22, 23]. When the PTE becomes sufficiently inaccurate, though, simulations in [22] showed a precipitous increase in the error of translation estimation. Above some threshold in PTE error, the algorithm is unable to correct errors in the PTE and converges to incorrect translation estimates. Simulations in [23] examined this phenomena as a function of noise in the detected intensity data and of the relative degree of overlap of the subaperture areas in neighboring translation positions. Higher signal-to-noise ratios and larger relative overlaps between subapertures delayed the onset of the precipitous increase in translation estimation error until higher PTE errors.

The existence of a precipitous increase in translation estimation errors for sufficiently imprecise PTE for CDI is consistent with our experience in optical metrology. However, the exit pupils reconstructed in the optical metrology application differ from the biological specimens examined in [22, 23]. When probing the central portion of the pupil, there are few sharp-edged amplitude or phase features to act as fiducials as is the case with CDI with biological or man-made specimens that have sharp-edged features. In Chapter 2, we calculated the Cramer-Rao lower bound (CRLB) for estimating subaperture translation from the measured PSF data in the optical
metrology application. The CRLB is in terms of (i) the aberration content of the exit pupil, (ii) the errors associated with retrieving the phase from a single PSF, and (iii) whether unknown motion of the point target being imaged must be estimated separately for each PSF. For example, consider a perfectly-corrected imaging system with no exit pupil phase aberrations where a subaperture is translating through a region of the pupil with constant amplitude. In-focus PSFs collected from such a system will be identical within the limits of paraxial optics and the CRLB for translation estimation tends to infinity [14]. If the point target is undergoing unknown motion, even systems with significant aberrations can have large CRLB bounds if those aberrations are just defocus and astigmatism. In this case, only the aberrations described by Zernike polynomial terms [24] whose radial degree is 3 or greater, such as coma, trefoil or spherical aberration, contribute subaperture translation information [14]. Specific pupil phase aberrations and point target translation estimation requirements can lead to poor subaperture translation estimation despite high subaperture overlap and favorable signal-to-noise ratio for the phase retrieval process alone. Though the CRLB analysis does not indicate how accurate PTE must be for success, it does indicate situations in which we expect errors in PTE to go uncorrected as there is insufficient information for joint solution of exit pupil phase and subaperture translation.

Our interest in unknown and potentially uncorrectable translation errors comes from analyzing wavefront sensing for the Near-Infrared Camera (NIRCam) instrument on the James Webb Space Telescope (JWST). Fine-phasing of the JWST is nominally
provided by focus-diverse phase retrieval using the shortwave channels of NIRCam [25]. This is not possible with the longwave channels, which lack the necessary weak lenses for defocus. They do have Lyot stops with special transmission functions intended for use in coronography [26], and rotation of the pupil wheel yields a mixture of translation and rotation of the Lyot stop relative to its nominal centered position. TTDPR can be applied to a collection of PSFs acquired for various small rotations of the pupil wheel, thus achieving wavefront sensing using the longwave channel of NIRCam [9, 13]. TTDPR for ground and on-orbit testing of NIRCam involves several unique challenges. For instance, calibration of the angle indication of the pupil wheel was potentially changing with each cryogenic cycling and vibration test during ground testing. Consequently, the PTE available during ground testing was limited to an approximate direction of translation and an even more approximate distance. Also, the ideally unresolved point target being imaged moved in an unknown fashion from PSF to PSF for some tests. Since this target motion must be estimated and because the residual aberrations are mostly of only 2nd radial degree like defocus and astigmatism, translation estimation is challenging consistent with the CRLB theory and simulations in Chapter 2. Consequently, we are concerned that systematic errors in the PTE arising from imprecise modeling of the Lyot stop motion could go uncorrected by the translation estimation steps in conventional TTDPR. We desire a TTDPR method impervious to errors in the PTE or that requires very little PTE information in a measurement situation unfavorable for estimating subaperture translation from the measured intensity data.
In this chapter, we propose a novel TTTDPR algorithm for the optical metrology application which we call unknown transverse-translation diversity phase retrieval (UTTDPR). By not assuming translation direction or distance PTE as part of the solution process, it avoids imprinting biases or systematic errors that might be in such explicit PTE. Subaperture translation values are estimated from the measured intensity data using an implicit PTE assumption that the PSFs are acquired sequentially from subaperture translations that are spatially contiguous. Initially, the algorithm solves for the exit pupil consistent with one "first" PSF. Later solutions are extended over larger and larger regions of the pupil by a multi-stage process that includes additional PSFs into the solution. The translation and rotation of the subaperture, the phase aberrations of the optical system, the amplitude of the exit pupil and the motion of the point target being imaged are all estimated during solution.

The algorithm consists of two major parts: (i) an error metric involving the unknown parameters of the TTDPR measurement described in Section 3.2 and (ii) and a multi-stage process for minimizing that error metric that accomplishes the necessary bootstrapping process. Section 3.3 outlines the conventional TTDPR solution using this metric for comparison with the proposed process described in Section 3.4. Experimental results from laboratory experiments are presented in Section 3.5.

### 3.2 Discrete Model and Error Metric

The UTTDPR algorithm optimizes an error metric quantifying the difference between PSFs from a parameterized model of the optical system and the actual measured PSFs.
This section discusses the construction of a parameterized model for the PSF intensity and the error metric. We begin with a model for the phase and amplitude of the exit pupil of the system.

Let \( Z_j \) be a discretely sampled two-dimensional array representing the \( j \)th Noll-ordered Zernike polynomial function [24]. Our parameterization of the phase of the exit pupil for the \( k \)th PSF is

\[
\phi_k = a_{2,k}Z_2 + a_{3,k}Z_3 + \sum_{j=4}^{J} a_j Z_j, \tag{3.1}
\]

where \( a_{2,k} \) and \( a_{3,k} \) are coefficients of the tip and tilt linear phase terms of the \( k \)th PSF, and \( a_j \) is the \( j \)th Zernike polynomial coefficient for all of the PSFs for \( j \geq 4 \). This model allows the linear phase terms to differ between PSFs in a condition we refer to as the unshared linear phase case, which accommodates for a translating detector or a target translating within an isoplanatic patch of the optical system as outlined in [14]. If the linear phase coefficients are known to be identical for all PSFs, the model reduces to the simpler expression

\[
\phi_k = \sum_{j=2}^{J} a_j Z_j, \tag{3.2}
\]

which we call the shared linear phase case and is independent of \( k \). The constant term \( Z_1 \) is not included in Eqs. (3.1) and (3.2) as it is undetectable by phase retrieval.

We refer to the amplitude in the exit pupil, when the subaperture is not present, as the fixed aperture. Furthermore, we presume that the fixed aperture must be estimated from PSF data. Earlier work [27, 28] observed difficulty in estimating amplitude from...
near-focus PSFs without multiple planes of focus-diverse PSF data. The methods proposed in [28] improve amplitude retrieval for near focus PSF data. However, these methods assume that the unknown amplitude is binary in nature (fully transmissive or completely opaque at each point). This assumption is appropriate for retrieving an unknown hard-edged stop in the subaperture plane. It is less suited to estimating a non-binary, smoothly varying shading of the fixed aperture. Fixed apertures with smoothly varying features can occur when the beam is vignetted by a surface that is not in the plane of the subaperture or due to the effects of a resolved, coherently or partially coherently illuminated target. We have observed that a fixed amplitude model parameterized by coefficients of a Zernike-polynomial can retrieve both smoothly-varying and hard-edged fixed amplitudes [29]. As in [29], we here model the fixed amplitude at array pixel \([n_r, n_c]\) by

\[
A_F[n_r, n_c] = \begin{cases} 
1.0 + C[n_r, n_c] & \text{if } C[n_r, n_c] > -1.0 \\
0.0 & \text{otherwise}, 
\end{cases} \quad (3.3)
\]

where

\[
C = \sum_{l=2}^{l} d_l Z_l, \quad (3.4)
\]

and \(d_l\) is the \(l\)th coefficient of a Zernike polynomial. The value of 1.0 in Eq. (3.3) is arbitrary but that value should be positive, and its negative should appear in the inequality of the if-condition. Parameterized models for the phase of an unknown field, like Eq. (3.2), are common in phase retrieval [10, 11] but similarly parameterized amplitude models are novel. This model encompasses three useful charac-
teristics for amplitude estimation. First, the model can vary from highly-regularized solutions when \( L \) is small to more detailed solutions when \( L \) is larger. This functionality will be used by the unknown TTDPR algorithm to avoid over-fitting the model to the data in situations where there is insufficient data to constrain a fixed amplitude model with many degrees of freedom. Second, Eq. (3.3) prohibits negative amplitudes which would yield undesirable \( \pi \)-phase shifts in the resulting exit pupil model. Third, since gradient formulas exist for the parameters in Eqs. (3.3) and (3.4) despite the discontinuity imposed by Eq. (3.3), efficient estimation of \( d \) can be performed jointly with the other unknowns during gradient-based optimization.

Having defined the phase and amplitude as above, a sampled array representing the field in the exit pupil absent the effects of the subaperture is

\[
  f_k = A_F \circ \exp (i\phi_k),
\]

where the complex exponential function is performed element-wise and the \( \circ \) operator represents the element-wise product of two arrays. The generalized exit pupil function [15] is

\[
  g_k = B_k \circ f_k,
\]

where the array \( B_k \) represents the transmission function of the subaperture after having been rotated by an angle \( \theta_k \) and translated by a distance \( (s_k, t_k) \).

Let \( A_S \) be a sampled version of the subaperture transmission function in a reference location having zero rotation and translation. Next, we define and justify a model for
how $B_k$ can be calculated from $\theta_k$, $(s_k, t_k)$ and $A_S$. Earlier TTDPR methods [6, 13], used the discrete Fourier shift theorem to evaluate translation. This was because Dirichlet kernel interpolation is effectively sinc interpolation if the arrays being interpolated are sampled finely enough to include the bandlimit of the continuous functions they sample. However, if the arrays to be interpolated are insufficiently sampled, the discrete Fourier shift theorem should not be applied. For example, the actual subaperture transmission function is likely to include sharp edges, making $A_S$ an aliased representation of the actual subaperture transmission function. Applying the discrete Fourier shift theorem to an array of aliased samples yields an inaccurate $B_k$ beset with Gibbs ringing along and near the edges of the subaperture. The earlier work [6] avoided this issue by applying the discrete Fourier shift theorem to the field incident on the subaperture under the assumption that the field would be bandlimited. This assumption usually holds when the subaperture does not overlap the edge of a hard-edged fixed aperture of the system. For the UTTDPR algorithm, however, we do not require that either the subaperture or the field impinging upon the subaperture be bandlimited functions. Hence we calculate $B_k$ from $A_S$ using bilinear interpolation, as detailed in Section 3.7. Bilinear interpolation is less affected by aliasing present in the sampled subaperture transmission function and it does not cause Gibbs ringing. Additionally, bilinear resampling allows the inclusion of sub-aperture rotation in the model.

Next, we consider the calculation of an array representing the field in the plane of the detector. The field in the plane of a detector near focus is a Fraunhofer propagation
\[ G_k(x, y) = \frac{A}{\lambda z} \int_{-\infty}^{\infty} g_k(u, v) \exp \left[ -\frac{i2\pi}{\lambda z}(xu + yv) \right] du \, dv, \]  

(3.7)

where \( A \) is a constant amplitude, \( \lambda \) is the wavelength of light, and \( z \) is the distance between the exit pupil and detector planes. A quadratic phase terms in \( x \) and \( y \) has been left out of Eq. (3.7) since we measure only intensity, and \( g_k \) is defined to include the quadratic phase in \( u \) and \( v \) associated with measurements not exactly at the focus plane. Let \( g_k(u, v) \) be sampled in two dimensions at intervals \( \Delta u \) to yield the \( N \) by \( N \) array \( g_k[n_r, n_c] \) with row and column indexes \( n_r \) and \( n_c \), respectively, according to

\[
  u = \Delta_u(n_c - n_0) \quad \text{and} \quad v = \Delta_u(n_r - n_0),
\]

(3.8)

where the sample at index \([n_0, n_0]\) corresponds to the optical axis where \((u, v) = (0, 0)\). Similarly, let \( G_k(u, v) \) be sampled at interval \( \Delta x \) to yield an \( M \) by \( M \) array \( G_k[m_r, m_c] \) according to

\[
  x = \Delta_x(m_c - m_0) \quad \text{and} \quad y = \Delta_x(m_r - m_0),
\]

(3.9)

where the index \([m_0, m_0]\) corresponds to the sample intersecting the optical axis. Substituting Eqs. (3.8) and (3.9) into Eq. (3.7), replacing the continuous functions with sampled equivalents and ignoring the leading constant yields

\[
  G_k[m_r, m_c] = \sum_{n_r=0}^{N-1} \sum_{n_c=0}^{N-1} g_k[n_r, n_c] \exp \left\{ -i2\pi \alpha \left[ (n_r - n_0)(m_r - m_0) + (n_c - n_0)(m_c - m_0) \right] \right\},
\]

(3.10)
where
\[ \alpha = \frac{\Delta u \Delta x}{\lambda z}. \] (3.11)

Eq. (3.10) is a discrete Fourier transform (DFT) between two uniformly sampled arrays. For simplicity of the formulae and consistency with [30], we will use \( \alpha \) in Eq. (3.11) rather than the more commonly used, and inversely-proportional, term \( \lambda / f_p \) defined in [31]. Such a DFT could be calculated by an operation involving zero-padding and the fast-Fourier transform (FFT) for integer values of \( \alpha^{-1} \) larger than \( N \) and \( M \). Alternately, the chirp-Z transform can be used for general \( \alpha \) [30] as can matrix multiplication [32–34]. The matrix multiplication DFT (MMDFT) is found as follows. For simplicity of the gradient calculation, let

\[ \omega[n, m] = -2.0 \pi \alpha (n - n_0)(m - m_0), \] (3.12)

for \( n = 0, ..., N - 1 \) and \( m = 0, ..., M - 1 \), and also define

\[ \Omega[n, m] = \exp (i\omega[n, m]). \] (3.13)

Rearranging Eq. (3.10) as

\[ G_k[m_r, m_c] = \sum_{n_r=0}^{N} \exp [-i2\pi \alpha (n_r - n_0)(m_r - m_0)] \sum_{n_c=0}^{N} g_k[n_r, n_c] \times \exp [-i2\pi \alpha (n_c - n_0)(m_c - m_0)], \] (3.14)
makes it clear that the two summations are equivalent to

\[ G_k = \Omega^\top \ast g_k \ast \Omega, \quad (3.15) \]

where the \( \ast \) operator here denotes matrix multiplication.

It is often desirable to estimate \( \alpha \) from the measured PSF data since the related quantities \( \Delta_x, \lambda \) or \( z \) may not be independently known to the required precision [30, 33, 35]. Earlier work in estimating the related quantities used tedious grid-search techniques or gradients with respect to \( \alpha \) evaluated using finite differences [35]. Later work found that the chirp-Z transform admitted analytic gradients with respect to \( \alpha \) [30], thus avoiding performance impact of a grid search or a finite-differences gradient method. We find that the MMDFT similarly admits an analytic gradient with respect to \( \alpha \), which is derived in Section 3.8. The simplicity of Eqs. (3.12)-(3.15) allows this formulation to be conceptually simpler than that of the chirp-Z transform. This simplicity also lends to improved computational performance. The chirp-Z transform is known to have better asymptotic complexity than the MMDFT[34], but it involves many high-level operations including (i) the calculation of three chirp phases, (ii) multiplication of the field by each of these chirp phases and (iii) two FFTs. Also, the FFTs and one of the chirp-multiplications are on large arrays of least \( N + M - 1 \) by \( N + M - 1 \) samples [30]. The MMDFT involves operations on smaller arrays that are at most \( N \) or \( M \) on a side. In terms of higher-level operations, the MMDFT also requires just the calculation of \( \Omega \) and two matrix-multiplies. Since the MMDFT utilizes smaller array sizes and just a few high-level operations, we have observed that
the MMDFT is computationally faster than the chirp-Z when $N$ and $M$ are too small to realize the asymptotic efficiency of the chirp-Z transform. This is the case for the TTDPR problem of interest where we expect the near-focus PSF of a well-corrected optical system to be encompassed by $M$ of just a few hundred pixels on a side.

The predicted intensity of the $k$th PSF is given by

$$I_k = |G_k|^2,$$  \hspace{1cm} (3.16)

where the absolute value and squaring operations are performed element-wise. Our model for the intensity measured by the detector, including an unknown detector scalar gain $\beta_k$ and bias $\gamma_k$, is

$$M_k = \beta_k I_k + \gamma_k \mathbf{1},$$  \hspace{1cm} (3.17)

where $\mathbf{1}$ is an $M$ by $M$ array of the value 1. Let $D_k$ be samples of intensity of the $k$th measured PSF, and let $w_k$ be a array of weightings on the detector coordinates $[m_r, m_c]$ that are zero for samples in $D_k$ containing bad pixels. A sum-of-squared-differences error metric between the model and measured data for the $k$th PSF is

$$E_k = \sum_{m_r, m_c} w_k \circ (M_k - D_k)^2 = \sum_{m_r, m_c} w_k \circ (\beta_k I_k + \gamma_k \mathbf{1} - D_k)^2.$$  \hspace{1cm} (3.18)

As noted in [36], metrics like Eq. (3.18) can be minimized with respect to detector gain and bias without treating $\beta_k$ and $\gamma_k$ as unknowns in the nonlinear optimization. Closed-form solutions for the values of $\beta_k$ and $\gamma_k$ that put $E_k$ at an extremum are derived in Section 3.9 and listed in Eq. (3.64).
Summing over the $E_k$ for individual PSFs yields a mean-squared error metric for $K$ PSFs

$$E = \sum_{k=k_{\text{min}}}^{k_{\text{max}}} E_k, \quad (3.19)$$

where $k_{\text{min}}$ and $k_{\text{max}}$ are the minimum and maximum index of the PSFs to be included in the error metric. Eq. (3.19) is implicitly a function of the unknown (i) phase parameters $a_{jk}$, (ii) amplitude parameters $d_l$, (iii) subaperture translations $s_k$ and $t_k$, (iv) subaperture rotations $\theta_k$ and (v) propagation scale term $\alpha$. The estimation process in the conventional and UTTDPR cases involves minimizations of the error metric in [5, 6] or Eq. (3.19) with respect to the respective unknowns of each problem. Such nonlinear optimization may be accomplished with a gradient-based optimization such as preconditioned conjugate-gradient method or the limited-memory Broyden-Fletcher-Goldfarb-Shanno (L-BFGS) algorithm [37, 38]. However, in addition to the error metric formula, these algorithms require a method for evaluating the gradient of the error metric with respect to the unknowns. Though this gradient can be estimated by the method of finite differences, it is more efficiently and accurately estimated using an analytic gradient formula. Gradients with respect to $s_k$, $t_k$, and $\theta_k$ are evaluated in Section 3.7 while the overall structure of the gradient calculation and specific gradients with respect to $a_{jk}$, $d_l$, and $\alpha$ are provided in Section 3.8.
3.3 Conventional TTDPR

Next, we review the optimization stages of conventional TTDPR [5, 6] utilizing the metric of Section 3.2 so that comparison may be made with the UTTDPR method discussed in Section 3.4. For conventional TTDPR, it would be assumed that initial values for subaperture positions \((s_k, t_k)\) could be given by the PTE, and subaperture rotation \(\theta_k\) and scaling \(\alpha\) would be assumed known. Similarly, a parameterized amplitude model such as Eqs. (3.3) and (3.4) does not appear in conventional TTDPR so it would be initially assumed that \(A_F\) is itself approximately known. Absent specific knowledge, it may be assumed that \(A_F\) is initially a uniform constant over a known aperture as in [6]. Finally, the initial value of \(a_{jk}\) can be selected from many appropriate starting guesses. In many cases, the initial assumption that \(a_{jk} = 0\) (an unaberrated wavefront) will lead to successful phase retrieval. If prior knowledge of the aberrations of the system, such as an approximately known defocus or astigmatism aberration, are available then that should be incorporated into \(a_{jk}\). Furthermore, we define \(J_L\) and \(J_H\) to be two free parameters of the algorithm corresponding to the largest index of Zernike coefficient to be used during low-order and higher-order phase estimation, respectively. Exact values are problem-dependent, but we assume that \(J_L \leq J_H\). Whichever is in use replaces the \(J\) appearing in the summations in Eqs. (3.1) and (3.2). For instance, using \(J_L = 11\) allows estimation of the 2nd and 3rd radial-degree terms as well as the 4th radial-degree spherical aberration term during the low-order phase estimation.
One selection of steps for conventional TTDPR consistent with earlier work [6] for solving for the unknown parameters is as follows:

1. **Linear phase estimation:** For each PSF, find the optimal translation of the model-predicted $I_k$ such that it best matches the measured $D_k$. This can be done using centroiding or the cross-correlation method of [39] to arbitrary precision, but 1/2 pixel is sufficient. For the unshared linear phase case, the discrete-Fourier shift theorem can then be used to infer new values of the linear $a_{2,k}$ and $a_{3,k}$ phase terms that yield such optimal shifts in the $I_k$. In the shared linear phase case, the mean of the optimal shifts can be used to update the single available set of the linear phase terms $a_2$ and $a_3$.

2. **Low-order phase estimation:** Let $k_{\text{min}}$ and $k_{\text{max}}$ be set to the minimum and maximum indices, respectively, of the available PSF data. Then, given the initial values for the parameters and newly-estimated linear phase terms from Step 1, minimize the error metric with respect to all linear phase parameters and low-order phase terms $a_j$ for $j \in [4, J_L]$.

3. **Phase and subaperture translation estimation:** Given the initial values for the parameters and newly estimated phase terms from Step 2, minimize the error metric with respect to the subaperture translations $(s_k, t_k)$ as well as all phase parameters $a_j$ for $j \in [4, J_{HL}]$. This allows for small errors in the subaperture translation estimate to be refined using the PSF data.

4. **Amplitude, phase and subaperture translation estimation:** If desired, add the per-pixel values of $A_F$ to list of parameters allowed to be altered during optimization.
and optimize to find a pixel-by-pixel estimate for the amplitude of the exit pupil and jointly-optimal phase and subaperture translation parameters. This results in a joint estimate for amplitude $A_F$, aberration coefficients $a_{jk}$ and subaperture translations $(s_k, t_k)$.

We have observed that this series of steps assists in alleviating stagnation of the optimization process in local minima of the error metric. In particular, Step 1 alleviates stagnation induced by the model PSF not having significant intensity overlapping regions of the measured PSF with significant intensity. Also, as noted in Chapter 2, the 2nd and 3rd radial degree phase aberrations have a critical role in contributing subaperture translation information to the subaperture translation estimation process. Consequently, we observe that estimating these lowest-order exit pupil phase in Step 2 prior to estimating subaperture translations in Step 3 aides successful retrieval. Finally, since estimating the pixel-by-pixel amplitude $A_F$ and estimating subaperture translation are both instances of estimating exit pupil amplitude, these estimations can have correlated errors. This is more noticeable when $A_F$ has a large number of degrees of freedom that can over-fit the data and lead to stagnation of the optimization at a local minimum. For example, the optimizer may find a solution that erroneously models vignetting induced by the subaperture using erroneous values in $A_F$ rather than the correct adjustment to the subaperture translation. By estimating subaperture translation in Step 3 and reserving the pixel-by-pixel amplitude for Step 4, we encourage the optimization process to explain as much of the observed exit pupil amplitude with subaperture translation adjustments as possible. Then, in Step 4, pixel-by-pixel
amplitude can be estimated with a reduced chance of a subaperture translation being mis-modeled as a variation in $A_F$.

### 3.4 Unknown TTDPR

In Section 3.1, we indicated the risks posed to joint translation and phase estimation by errors in the PTE going uncorrected in the final estimate. In this section, we outline unknown transverse-translation diversity phase retrieval (UTTDPR), a method that is independent of prior direction-of-translation information and requires only a bound on the unknown distance of subaperture translation from one PSF to the next. Since explicit PTE in the form of approximate translation directions and distances is not used, it does not have to be reliably measured or estimated before application of the algorithm. Also, UTTDPR will not stagnate in a local minimum due to the influence of PTE errors on the convergence of the optimization. Still, the optimization may fail to converge but such a failure can not be induced by errors in the PTE since explicit PTE is not utilized.

Estimating subaperture translation requires significant knowledge of the low-order exit pupil phase, as in Chapter 2, but one cannot estimate phase without approximately knowing the subaperture translations. This circular dependence between subaperture translation and phase means that normal algorithms like that described in Section 3.3 are prone to stagnation during low-order phase estimation in Step 2. Simple choices for initial $s_k$ and $f_k$ usually result in the optimizer settling in a local minimum of $E$ with completely erroneous values of the unknown subaperture translations. The problem of
simultaneously estimating phase and subaperture translation is acute in the unshared linear phase case as any residual defocus or astigmatism in the exit pupil wavefront does not aide subaperture translation estimation as it would in a shared linear phase case [14]. Additionally, in UTTDPR it is presumed that the pupil amplitude is also unknown, thus further complicating subaperture translation estimation for the reason mentioned in Section 3.3.

UTTDPR addresses these challenges with a series of optimization steps similar to conventional TTDPR, but adjusted to confront the case that the exit pupil phase and the direction of subaperture translation from PSF to PSF is unknown. The distance translated is also unknown but presumed bounded. Let the PSFs be acquired in a sequential, contiguous, motion of the subaperture where the change in subaperture translation from PSF to PSF is a small fraction of the width of the subaperture. Similarly, we assume that the subaperture rotation varies by small amounts between PSFs with consecutive indexes. Additionally, the fixed amplitude may be assumed unknown but approximately constant over the subaperture of at least one PSF with index $k'$ for which the subaperture rotation $\theta_{k'}$ can be guessed to within a small number of degrees. Typically the PSF with index $k'$ would derive from a subaperture location near the center of the fixed aperture. If the index $k'$ of a PSF with such a subaperture translation and fixed amplitude situation is unknown, an arbitrarily chosen PSF appearing near the middle of the sequence may suffice.

As in conventional TTDPR, the initial values for $a_{jk}$ are presumed using the best available prior knowledge, or equal to zero or some random starting guess in the
absence of prior phase knowledge. We use the amplitude model of Eqs. (3.3) and (3.4) with initial values \( d_l = 0 \) which yields a constant \( A_F \). The initial value for \( \theta_{k'} \) is taken to be the guess mentioned above and the initial value of \((s_{k'}, t_{k'})\) is taken to be \((0, 0)\). It is also presumed that \( \alpha \) is known to within a few percent and that this approximate value can be used as an initial guess. We define \( L_L, L_M \) and \( L_H \), with \( L_L \leq L_M \leq L_H \), to be three additional free parameters of the algorithm corresponding to the largest index of a Zernike coefficient to be used during low-order, medium-order and higher-order amplitude coefficient estimation, respectively. The exact choice of these free parameters is problem dependent. Each of these three parameters replace the \( L \) in Eq. (3.4) at various stages of estimation. Finally, the initial values of \( k_{\text{min}} \) and \( k_{\text{max}} \) are assigned to be \( k' \), so that the error metric initially only includes the \( k' \)th PSF.

We have, through a process of trial-and-error, found a series of optimization steps based on the ideas of conventional TTDPR that often yields joint solutions for the UTTDPR problem given the starting estimates above. The unknowns to be found or refined are (i) phase parameters \( a_j, k \), (ii) amplitude parameters \( d_l \), (iii) subaperture translations \( s_k \) and \( t_k \), (iv) subaperture rotations \( \theta_k \) and the (v) propagation scale term \( \alpha \). The steps are as follows:

1. **Initial linear phase estimation:** The optimal translation of the model-predicted \( I_{k'} \) such that it best-matches the measured \( D_{k'} \) as outlined in Step 1 from conventional TTDPR in Section 3.3 is calculated. \( a_{2,k'} \) and \( a_{3,k'} \) are updated to incorporate these translations into the initial parameters.
2. **Low-order phase estimation**: The error metric is minimized with respect to the linear phase terms \( a_{2,k} \) and \( a_{3,k} \), and the low-order phase terms \( a_j \) for \( j \in [4, L_L] \).

Since \( k_{\text{min}} = k_{\text{max}} = k' \), only the \( k' \)th PSF contributes to the error metric. Thus, this step finds a low-order phase solution consistent with just the \( k' \)th PSF under the assumption that the subaperture is un-translated and the fixed amplitude is constant. The subaperture transmission function is the only distribution in the amplitude of the exit pupil model.

3. **Initial amplitude estimation**: The amplitude coefficients corresponding to \( d_l \) for \( l \in [2, L_L] \), in addition to the parameters estimated in Step 2, are allowed to vary during a minimization. This finds a joint solution for a few of the amplitude coefficients and the phase coefficients estimated in Step 2.

4. **Initial rotation estimation**: \( \theta_{k'} \) is added to the parameters estimated in Step 3 and minimization performed to find a joint solution for rotation of the subaperture during the \( k' \)th PSF and the previously estimated amplitude and phase coefficients.

5. **Initial \( \alpha \) estimation**: \( \alpha \) is added to the parameters estimated in Step 4 and minimization of the error metric executed to find a joint solution for \( \alpha, \theta_{k'}, \) and the previously estimated amplitude and phase coefficients.

6. **Include one lesser PSF**: If \( k_{\text{min}} \) can be decremented by one to yield the index of an available data PSF, this is done. Also, once decremented, values for \( \theta_{k_{\text{min}}}, s_{k_{\text{min}}} \) and \( t_{k_{\text{min}}} \) are initially set equal to \( \theta_{(k_{\text{min}}+1)}, s_{(k_{\text{min}}+1)} \) and \( t_{(k_{\text{min}}+1)} \). This makes the erroneous but useful first approximation that the \( k_{\text{min}} \)th PSF is in the same location as the PSF with index \( k_{\text{min}} + 1 \). The linear phases \( a_{2,k_{\text{min}}} \) and \( a_{3,k_{\text{min}}} \) are estimated.
by the correlation method outlined in Step 1 from conventional TTDPR in Section 3.3 assuming the approximate subaperture translation.

7. **Include one greater PSF:** If \( k_{\text{max}} \) can be incremented by one to yield the index of an available data PSF, this is done. Also, once incremented, values for \( \theta_{k_{\text{max}}} \), \( s_{k_{\text{max}}} \) and \( t_{k_{\text{max}}} \) are initially set equal to \( \theta_{(k_{\text{max}}-1)} \), \( s_{(k_{\text{max}}-1)} \) and \( t_{(k_{\text{max}}-1)} \). The linear phases \( a_{2,k_{\text{max}}} \) and \( a_{3,k_{\text{max}}} \) are estimated by the correlation method as in Step 6.

8. **Subaperture translation estimation:** Optimization is performed to minimize the error metric with respect to the translation and linear phase terms corresponding to the \( k_{\text{min}} \)th PSF if it was newly decremented in Step 6 and the \( k_{\text{max}} \)th PSF if it was newly incremented in Step 7. Other unknowns, including the phase coefficients of 2nd radial degree and higher, are not allowed to vary during this optimization.

9. **Rotation estimation:** Optimization is again performed with respect to the parameters estimated in Step 8 but also including subaperture rotation \( \theta_{k_{\text{min}}} \) if \( k_{\text{min}} \) was newly decremented in Step 6 and \( \theta_{k_{\text{max}}} \) if \( k_{\text{max}} \) was newly incremented in Step 7.

10. **Medium-order amplitude refinement:** The amplitude coefficients corresponding to \( d_l \) for \( l \in [2, L_M] \), in addition to the parameters estimated in Step 9, are allowed to vary during optimization.

11. **Low-order phase refinement:** The low-order phase terms \( a_j \) for \( j \in [4, J_L] \) are included in the list of parameters to optimize in addition to those in Step 10.

12. **Refine everything but higher-order amplitude and phase:** Subaperture translation, rotation and linear phase terms for PSFs with indexes from \( k_{\text{min}} \) through \( k_{\text{max}} \), as well as \( \alpha \) and the parameters of Step 11 are allowed to vary during optimization.
13. Repeat Steps 6 through 12, unless \( k_{\text{min}} \) and \( k_{\text{max}} \) correspond to the lowest and highest indexed data PSFs in the dataset.

14. **High-order amplitude estimation:** The amplitude coefficients corresponding to \( d_l \) for \( l \in [2, L_H] \), in addition to the parameters estimated in Step 12, are allowed to vary during a minimization.

15. **High-order phase estimation:** The error metric is minimized with respect to high-order phase terms \( a_j \) for \( j \in [4, J_H] \) in addition to the parameters refined in Step 14.

For UTTDPR in the shared linear phase case, several steps are modified. In Steps 1 and 2, it is \( a_2 \) and \( a_3 \) that are estimated rather than \( a_{2,k'} \) and \( a_{3,k'} \). The process of estimating linear phase by correlating model PSFs with data PSFs is eliminated entirely from Steps 6 and 7, and no attempt to optimize over linear phase is made in Steps 8-10. In Steps 11, 12, 14 and 15, \( a_2 \) and \( a_3 \) are included as parameters for optimization.

### 3.4.1 Discussion

As the subaperture translations of the \( k_{\text{min}} \)th and \( k_{\text{max}} \)th PSFs optimize in Step 8, the clear area of the subaperture overlaps regions of the exit pupil phase and fixed amplitude not accessed by the subaperture in the PSFs incorporated into the solution by earlier computations. Since the phase and fixed amplitude coefficients are held constant in Steps 6-9, these newly-accessed regions of phase and amplitude are extrapolations using the phase and fixed amplitude found earlier Steps. Once this translation is approximated using the extrapolated phase and fixed amplitude, the subaperture
rotation can be estimated in Step 9. Next, the amplitude and phase Zernike coefficients are refined in Steps 10 and 11 allowing the overall phase and fixed amplitude estimate to conform to the PSFs included in the solution. Then, most of the unknowns are refined in Step 12 to form a robust estimate for later iterations of Steps 6 through 12 incorporating PSFs that access additional new regions of the pupil. The process of estimating the translation for new PSFs using extrapolations from the model fitted to the old PSFs repeats until a joint solution for phase, amplitude and subaperture translation is found incorporating all PSFs. It is this process of incrementally solving larger and larger sub-problems of the overall TTDPR problem that is unique to UTTDPR and enables it to estimate joint solutions without initial estimates for important parameters like subaperture translation and linear phase.

We make no claim that the numerous steps of the UTTDPR algorithm above uniquely solve the problem. They were derived from practical experience accumulated through working with specific sets of simulated data, actual data collected in the laboratory and from the NIRCam instrument. Depending on the particular problem, some steps may be omitted, transposed or aggregated with other steps to simplify the algorithm. The listed steps are, however, the ordering that we have found to be the most robust for the data coming from NIRCam and NIRCam-like laboratory experiments and simulations. Skipping some of the steps occasionally led to stagnation without finding a solution.

There is no guarantee that an acceptable solution for subaperture translations can always be obtained even if the Cramer-Rao lower bound for translation estimation is
favorable. It may not be possible for Steps 6-8 to converge to a good joint estimate for the linear phase and subaperture translations from the estimates for exit pupil phase and amplitude derived from the earlier steps. Noise in the detected PSFs alters the location of global minimum of the error metric and may add new local minima as in [40]. It is also possible to have a PSF whose subaperture translation from its neighbor is so great that the optimizer would have to go up in error metric value before finding this better minimum of the error metric. This phenomena yields a constraint on the minimum required overlap of the subaperture transmission function between neighboring PSFs similar to those noted in [22, 23]. We have not assessed this distance as it is likely highly dependent on particular pupil phase and fixed amplitudes.

3.5 Laboratory Experiment

In [9], we reported results of applying algorithms essentially similar to UTTDPR to simulated data. For this chapter, we report laboratory results using the UTTDPR algorithm. The aberrated optical system shown schematically in Fig. 3.1 was assembled to produce data for TTDPR. Light from a helium-neon laser ($\lambda = 632.8$ nm) was focused onto a pinhole spatial filter and then collimated by Lens I. The collimated beam then passed through a moving subaperture. A Lyot stop transmittance function [13], shown in Fig. 1.3, similar to the Lyot stop used in NIRCam, was cut into sheet metal to form a subaperture having horizontal width 4.6 mm. It was affixed to a rotation stage allowing it to rotate about an axis parallel to $\hat{z}$. The distance from this true axis of rotation to the center of the subaperture was approximately 96 mm. The
Figure 3.1: Schematic of TTDPR laboratory experiment. Not shown: neutral density filter and focusing microscope objective between the laser and the spatial filter.

The rotation stage was mounted to an X-Y translation stage. After the sub-aperture, the light passed through Lens II and then Lens III which was offset in the \( \hat{y} \) direction and rotated about the \( \hat{y} \) axis by approximately 20 degrees in order to induce aberrations in the imaging of the pinhole to a CCD detector having 7.4 \( \mu \)m square pixels. An example PSF is shown in Fig. 3.2 (a).

In the first of two experiment performed with the apparatus, PSFs were acquired with subaperture translations on a 7-by-7 grid separated by approximately 766 \( \mu \)m through motion of the X-Y translation stage. Each position was randomly-perturbed by values drawn from a zero-mean Gaussian distribution having a standard deviation of 70 \( \mu \)m. These randomly-perturbed translations are plotted by the circle features in Fig. 3.2 (b). The translation distance from one subaperture location to the next was about 17% of the width of the subaperture: a substantial overlap between neighboring positions of the subaperture. PSFs were collected sequentially down one column in
$y$, then over one increment in $x$, then up the next column, and over one increment in $y$. This pattern of motions was repeated to achieve a series of PSFs arising from spatially-contiguous subaperture translations as required for UTTDPR. To simulate the effect of a moving detector or unshared linear phase in the pupil, the region of the detector pixels passed to UTTDPR was randomly offset a few pixels from one PSF acquisition to the next. This randomly-offset PSF data was processed by the unshared linear-phase variant of the UTTDPR algorithm described in Section 3.4. The number of Zernike phase terms were $J_L = 9$ and $J_H = 35$, while the number of amplitude parameters were $L_L = 5$, $L_M = 9$ and $L_H = 35$. Minimization of the error metric was accomplished using the L-BFGS algorithm [37]. The algorithm had no knowledge of the specific subaperture translations but it did have access to the approximate values of defocus phase, subaperture rotation of the initial PSF and $\alpha$ to use as initial estimates for optimization. The subaperture translations retrieved are shown by the “x” features in Fig. 3.2 (b). The error between the TTDPR-estimated subaperture translations and the truth from the X-Y stage was $36 \mu \text{m}$ root mean squared (RMS). After normalization by the reference translation amounts, the normalized root mean squared error (NRMSE) was 0.024. This error was about 0.8% of the width of the sub-aperture or 0.2% of the width of the entire retrieved pupil. The pupil phase retrieved by UTDDPR is shown in Fig. 3.2 (c) with the linear phase terms and 4.6 waves peak-to-valley (P-V) of defocus subtracted. Over regions of the pupil encompassed by any particular subaperture translation, the amount of defocus was approximately 0.5 waves P-V.
Figure 3.2: (a) Example PSF of the aberrated system: intensity plotted to the 0.33 power. (b) Scatter plot of sub-aperture translations for the first experiment: red crosses are UTTDPR recovered values and black circles indicate true translations commanded to the X-Y stage. (c) Retrieved pupil phase for first experiment using UTTDPR with defocus and linear phase removed. (d) Phase difference (error) between UTTDPR and conventional TTDPR with knowledge of the subaperture translations for the first experiment.
Since for these experiments we actually had good knowledge of the translations, conventional TTDPR that used this PTE was performed on the same data but without the random shifting of readout of the detector pixels. This permitted the use of shared linear phase conventional TTDPR which is known to increase subaperture translation estimation accuracy in the presence of defocus and astigmatism [14]. Subaperture translations and rotations were estimated using conventional TTDPR given the prior-translation estimate of the reliable X-Y stage motion. The translations retrieved using conventional TTDPR were used to fit an unknown rotation between the CCD camera and the axes of subaperture motion. The truth translation position shown in Fig. 3.2 (b) and our root mean squared error (RMSE) and NRMSE for UTTDPR are reported relative to this rotation correction. The difference between the UTTDPR phase and conventional TTDPR reference phase is shown in Fig. 3.2 (d) and had an average error of $3.7 \times 10^{-3}$ waves RMS over the accessed regions of the pupil.

For the second experiment, we chose a circular motion similar to the motion of the Lyot stop about the axis of the pupil wheel in NIRCam. The transmissive parts of the Lyot stop pattern shown in Fig. 1.3 swept through the regions of the pupil shown in Fig. 3.3 (a). To accomplish this sweeping motion with the apparatus, the rotation stage rotated from $-4$ to $+4$ degrees in evenly-spaced increments to yield a motion of the center of the subaperture represented by the circle features in Fig. 3.3 (b). These positions are approximate due to the practical difficulty of estimating the lever-arm vector between the axis of rotation and the subaperture as projected into the plane perpendicular to the beam. Our best estimate for the location of the pivot point was
Figure 3.3: (a) Areas of pupil accessed during subaperture motion in the second experiment. (b) Scatter plot of sub-aperture translations for the second experiment: red crosses are UTTDPR recovered values and black circles indicate true translations estimated from knowledge of rotation angle and geometry. (c) Fixed amplitude to the 0.5 power recovered by UTTDPR for the second experiment. (d) Fixed amplitude to the 0.5 power recovered by sieve amplitude TTDPR for the second experiment. (e) Retrieved pupil phase for the second experiment using UTTDPR with defocus and linear phase removed. (f) Phase difference (error) between UTTDPR and sieve amplitude TTDPR for the second experiment.
(s, t) = (41.5, −74.4) mm. The average distance translated by the subaperture from
one PSF to the next was about 11% of the width of the subaperture.

A hard-edged circular aperture was imposed into the collimated space between the
subaperture and Lens II to make the second experiment more representative of TTDPR
with NIRCam which may include unknown pupil-shear and vignetting. The effect
of this additional aperture was to completely extinguish the beam along the upper-
right corner of the pupil shown in Fig. 3.3 (a). This unknown fixed-amplitude effect
allowed the second experiment to demonstrate the fixed amplitude retrieval capability
of UTTDPR in a challenging situation.

Like the first experiment, the PSF data from the second was processed by the
unshared linear-phase variant of the UTTDPR algorithm but with the number of ampli-
tude parameters \( L_M = 20 \) and \( L_H = 119 \). Since \( L_H \) was much larger than \( L_M \), Step 14
was amended to include additional optimizations with 35, 54, 77 and 91 amplitude
terms before finally optimizing with \( L_H \). Again, the UTTDPR algorithm had no
knowledge of the specific subaperture translations in each PSF or knowledge that
the subaperture motion lied on a circular arc. It only had access to approximate values
of defocus phase, \( \alpha \) and an intelligent guess for the rotation of the \( k' \)th PSF to use
as starting points. The subaperture translations retrieved by UTTDPR are shown by
the “x” features in Fig. 3.3 (b). Relative to our estimate for the true translations, the
UTTDPR retrieved translations had a RMSE of 45 \( \mu \)m or a NRMSE of 0.017. As a
fraction of the width of the accessed regions of the pupil, this represents a translation
estimation error of 0.3%. A portion of these translations errors are consistent with
the pivot point of the rotation not being where we estimated it to be. A best-fit to the UTTDPR translations suggests motion about a pivot point of $(s, t) = (42.2, -73.9)$ mm which was 0.88 mm away from our estimate. If we take this pivot point estimated from the UTTDPR translations as correct, the residual error of the UTTDPR solution was $37 \, \mu\text{m}$ RMS, which is consistent with the $35 \, \mu\text{m}$ RMSE observed for the first experiment. As a fraction of the width of the overall accessed region of the pupil, this represented a 0.24% translation error.

The fixed amplitude retrieved by UTTDPR is shown in Fig. 3.3 (c). Since we can not know the amplitude in regions of the pupil where the transmissive portion of the subaperture has not been, the fixed amplitude is only shown over the areas outlined by Fig. 3.3 (a). Comparing that amplitude with Fig. 3.3 (a) indicates that UTTDPR has retrieved a sharp-edged vignetting feature in the upper-right corner of the fixed amplitude consistent with the unknown aperture we imposed. It is remarkable that 119 Zernike polynomial coefficients can, in conjunction with the thresholding of Eq. (3.3), represent a hard-edged vignetting surface and this is consistent with the simulations in [29].

Since for this experiment we also had good knowledge of the translations, classic TDDPR was performed using our best estimates for the subaperture motion. The amplitude of the fixed pupil, $A_F$, was allowed to vary in a pixel-by-pixel manner as is classic TTDPR to recover the fixed pupil effects of the additional hard-edged aperture. However, the resulting estimates for pupil amplitude were corrupted by salt-and-pepper noise to an unacceptable extent as in [29]. We have found this failure
mode to be typical of pixel-by-pixel amplitude retrieval involving a single plane very
near focus and other work has encountered similar challenges as mentioned in Section
3.2. The experiment had little defocus to approximate the NIRCam application where
amplitude retrieval is challenging and to demonstrate the capabilities of UTTDPR
utilizing the Zernike amplitude model.

To generate a reference more suitable for comparison with UTTDPR, we amended
our conventional TTDPR algorithm to use the method of sieves [41] to regularize
the retrieved pixel-by-pixel fixed-amplitude. We refer to this modified conventional
TTDPR algorithm as sieves amplitude TTDPR. It begins with performing conven-
tional TTDPR with the starting guess that $A_F[n_r, n_c] = 1.0$ and allowing the pixels
to vary on the interval $A_F[n_r, n_c] \geq 0.0$. The resulting estimate for $A_F$ was blurred
with a wide Gaussian kernel and then used as the starting point for a second conven-
tional TTDPR optimization. This cycle of blurring and optimization was repeated
with Gaussian kernels of ever-decreasing widths to find the retrieved amplitude shown
in Fig. 3.3 (d). This amplitude also showed a hard-edged circular vignetting surface
in the same place as the UTTDPR solution. However, it also showed many pixels
with amplitudes between $0.1$ and $0.3$ in the area beyond the edge that should be near
zero. Conventional TTDPR without the method of sieves also yielded significant
values in the region that should be zero. We regard this as a limitation of pixel-by-
pixel amplitude models, like those used in conventional TTDPR and sieves amplitude
TTDPR, because the circular aperture imposed was utterly opaque. The UTTDPR
fixed-amplitude solution shows no such feature; the values of $A_F$ from the Zernike
coefficient model in the retrieval were exactly zero in the vignetted regions beyond the circular edge. Similarly, the UTTDPR amplitude in Fig. 3.3 (c) was smooth in the evenly-illuminated regions of the pupil while the conventional TTDPR solution had spurious fine features near the Lyot stop’s edges in each of the PSFs. We attribute these fine features to artifacts of the pixel-by-pixel amplitude model that are excluded by the Zernike amplitude model used in UTTDPR. Both the UTTDPR and sieves amplitude TTDPR amplitudes showed a darkening of the amplitude at the bottom of the displayed pupil. This was consistent with the pupil illumination due to a coherently-illuminated resolved pinhole. It may be possible to construct a UTTDPR-like algorithm with a different amplitude model than the Zernike polynomial model. However, we have found the smooth amplitude extrapolations yielded by the Zernike model useful for improving the probability of convergence of UTTDPR, and we do not know whether other alternate amplitude models will yield a similar benefit.

The phase retrieved by UTTDPR is shown in Fig. 3.4 (a) with 7.7 waves P-V of defocus subtracted. It was plotted over the regions of the pupil which UTTDPR indicate have an amplitude greater than 0.01. This retrieved phase was consistent with an extension of the aberrations retrieved from the first experiment shown in Fig. 3.2 (c) extended over a larger area of the pupil, as expected. The difference between the UTTDPR phase and the sieves amplitude TTDPR phase is shown in Fig. 3.4 (b). The magnitude of this error was $3.0 \times 10^{-3}$ waves RMS over the accessed regions of the pupil.
Figure 3.4: (a) Retrieved pupil phase for the second experiment using UTTDPR with defocus and linear phase removed. (b) Phase difference (error) between UTTDPR and sieve amplitude TTDPR for the second experiment.

### 3.6 Summary and Conclusion

Transverse-translation diversity phase retrieval (TTDPR), also known as ptychography, is a technique that can infer the complex-valued exit pupil of an optical system from point spread functions (PSFs) of that system. The PSFs differ according to the translation of a subaperture placed in a plane approximately conjugate to the exit pupil. Algorithms can find a complex exit pupil field consistent with the measured PSF data though they are sensitive to errors in prior knowledge of the translation of the subaperture for each PSF, which we term the prior-translation estimate (PTE). Algorithms like [5, 6, 19, 20, 22] refine estimates of the translation using the measured PSF data given approximately correct PTE.
These previous forms of ptychography have assumed some approximate knowledge of where the subaperture or probe is relative to the ensemble of intensity measurements. We consider optical metrology when this explicit PTE is unavailable or untrusted due to the potential for introducing systematic error. Instead, we presume a sequence of PSFs are acquired from spatially-contiguous subaperture translations in an unknown arrangement in the exit pupil which is a very weak form of PTE. Our optical metrology technique, unknown transverse-translation diversity phase retrieval (UTTDPR), uses the error metric of Section 3.2 and a multi-stage bootstrapping process described in Section 3.4. The exit pupil phase and amplitude are initially estimated from phase retrieval performed on a single “initial” PSF. These subaperture positions in neighboring PSFs include regions of the complex exit pupil unknown to the algorithm given the phase retrieval of the first PSF. In these unknown regions, models for the complex exit pupil provide values for phase and amplitude that are extrapolations of the phase and amplitude retrieved from the single PSF. A novel amplitude model based on Zernike-polynomials aided this extrapolation and the estimation of pupil amplitude. If the unknown subaperture translations of neighboring PSFs keep the subaperture inside the region of validity of the extrapolated exit pupil, acceptable subaperture translation estimates can usually be obtained by assuming the extrapolated exit pupil. Later steps find a joint phase, amplitude and subaperture translation solution across all three PSFs and the extrapolation is eliminated using the newly-included PSF data. This process of extrapolation, translation estimation and refinement is repeated using additional PSFs not previously included in the solution until all the PSFs are part of a joint solution.
The UTTDPR algorithm also simultaneously estimated: (i) a moving point target or unknown varying linear phase in the exit pupil, (ii) unknown subaperture rotation, and (iii) unknown scale factor inherent in the propagation from the generalized exit pupil function to the PSF plane. Estimation of subaperture rotation and translation was aided by the use of bilinear interpolation while estimation of the scale factor employed the matrix-multiply discrete Fourier transform.

In Section 3.5, we demonstrated the algorithm using two simple laboratory experiments with a NIRCam-like subaperture and simulated unknown point target motion. In both experiments, UTTDPR recovered the unknown subaperture translations despite the target motion and without specific knowledge of the translation of the subaperture in each PSF. In the first experiment, the subaperture was raster-scanned in two dimensions; while in the second experiment, the subaperture translated along an arc while rotating in a way similar to the Lyot stop of NIRCam. The retrieved subaperture translations had RMSE of 0.2% and 0.3% of the width of the overall pupil for the first and second experiments respectively. We compared the UTTDPR phase retrieval results to classic TTDPR assuming a fixed target and known subaperture translation. Relative to classic TTDPR, UTDDPR yielded exit pupil phase errors of $3.7 \times 10^{-3}$ and $3.0 \times 10^{-3}$ waves RMS for the first and second experiments, respectively. Further, the second experiment demonstrated that UTTDPR could reconstruct the fixed amplitude due to an unknown hard-edged aperture imposed in the system. Its reconstruction using the special Zernike amplitude model was superior to that of conventional TTDPR involving a pixel-by-pixel fixed amplitude model regularized by the method of sieves.
In conclusion, we have proposed and demonstrated an extremely flexible TTDPR algorithm for ptychographic optical metrology of imaging systems. Unlike earlier algorithms, the algorithm does not need to know the approximate translation of the subaperture relative to the ensemble of subaperture translations nor an approximate direction of translation from one PSF to the next. In the cases evaluated, the algorithm recovered the subaperture translations and rotations from the PSF data. It simultaneously estimated significant nuisance parameters such as: (i) an unknown fixed pupil amplitude involving hard-edged features, (ii) the arbitrary motion of the point target and (iii) the scale factor associated with the propagation from the aperture stop to PSF plane.

### 3.7 Bilinear Interpolation

In this section, we describe the bilinear interpolation of an array $A_S$ to calculate the translated and rotated array, $B_k$, used in Eq. (3.6). This description is structured to allow us to derive an analytic gradient formula with respect to the angle of rotation and translation distances applied by the interpolation.

First, consider the mapping of physical coordinates $(u', v')$ of $A_S$ to physical coordinates $(u, v)$ of $B_k$ due to a rotation about the origin by angle $\theta_k$ followed by a translation. Let the translation be by a vector $[s_k, t_k]$ having units of pixels of $A_S$. In physical coordinates, this translation is by a vector $(\Delta_u s_k, \Delta_u t_k)$ where $\Delta_u$ is the pupil sampling spacing defined in Section 3.2. The mapping can be represented by
the matrix multiplication [42]
\[
\begin{bmatrix}
  u \\
  v \\
  1
\end{bmatrix} =
\begin{bmatrix}
  C_k & S_k & \Delta_u s_k \\
  -S_k & C_k & \Delta_u t_k \\
  0 & 0 & 1
\end{bmatrix}
\begin{bmatrix}
  u' \\
  v' \\
  1
\end{bmatrix},
\]
(3.20)

where \( C_k = \cos \theta_k \) and \( S_k = \sin \theta_k \). An inverse mapping [42] can be found for Eq. (3.20) by matrix inversion:
\[
\begin{bmatrix}
  u' \\
  v' \\
  1
\end{bmatrix} =
\begin{bmatrix}
  C_k & -S_k & -C_k \Delta_u s_k + S_k \Delta_u t_k \\
  S_k & C_k & -S_k \Delta_u s_k - C_k \Delta_u t_k \\
  0 & 0 & 1
\end{bmatrix}
\begin{bmatrix}
  u \\
  v \\
  1
\end{bmatrix},
\]
(3.21)

Next, we find the inverse mapping of integer sample coordinates \([n_r, n_c]\) in \( B_k \) to non-integer sample coordinates \([n'_r, n'_c]\) in the array \( A_S \) as required for bilinear interpolation [42]. Assume that the sample of \( B_k \) corresponding to the physical origin is at index \([n_0, n_0]\) and that \((u, v)\) can be found from \([n_r, n_c]\) by Eq. (3.8). Similarly let

\[
u' = \Delta_u (n'_c - n'_0) \quad \text{and} \quad v' = \Delta_u (n'_r - n'_0),
\]
(3.22)

where \([n'_0, n'_0]\) is the index of the sample corresponding to the physical origin in \( A_S \). Substituting Eqs. (3.8) and (3.22) into Eq. (3.21) and simplifying yields
\[
\begin{bmatrix}
  n'_c \\
  n'_r \\
  1.0
\end{bmatrix}^T = R^{-1}_k \begin{bmatrix}
  n_c \\
  n_r \\
  1.0
\end{bmatrix}^T,
\]
(3.23)
where

\[
R_k^{-1} = \begin{bmatrix}
C_k & -S_k & n'_0 - C_k(s_k + n_0) + S_k(t_k + n_0) \\
S_k & C_k & n'_0 - S_k(s_k + n_0) - C_k(t_k + n_0) \\
0 & 0 & 1
\end{bmatrix}.
\]  

(3.24)

The coordinates \(n'_c\) and \(n'_r\) in Eq. (3.23) implicitly depend on PSF index \(k\).

Let \(B_k\) have integer row and column coordinates on the interval 0 to \(N - 1\). Also, let \(B_k\) be stored in linearly-addressed memory in row-major order such that the linear address of a sample in memory is \(n = Nn_r + n_c\). Further, let \(n\) be a 3 by \(N^2\) array defined by

\[
n[\rho, n] = \begin{cases}
  n \mod N & \text{for } \rho = 0 \\
  \lfloor n/N \rfloor & \text{for } \rho = 1 \\
  1.0 & \text{for } \rho = 2,
\end{cases}
\]

(3.25)

where \(\mod\) represents the modulo operator and \([x]\) denote rounding \(x\) down to the nearest integer. Eq. (3.25) defines \(n\) to have the useful property that the first two rows of the \(n\)th column corresponds to the row index \(n_r\) and column index \(n_c\) of the sample of \(B_k\) at linearly-addressed memory position \(n\). By Eq. (3.23), then, the matrix-multiplication

\[
n'_k = R_k^{-1} * n,
\]

(3.26)

has similar useful properties. By Eq. (3.23), the \(n\)th column of \(n'\) has \(n'_c\) in the first row and \(n'_r\) in the second row corresponding to the coordinate of \(A_S\) to which the \(n\)th linearly-addressed sample of \(B_k\) is inverse-mapped.
In general, the coordinates in $n'_k$ do not fall on the integer positions associated with samples of $A_S$ and thus some form of interpolation must be used. For TTDPR, we prefer bilinear interpolation as described in Section 3.2. Next, assume that $A_S$ has row and column coordinates with indices on the interval $0$ to $N' - 1$, inclusive. For coordinates $[n_r, n_c]$ corresponding to $n$ satisfying

\[ 0 \leq n'_k[0, n] < N' - 1 \quad \text{and} \quad 0 \leq n'_k[1, n] < N' - 1, \quad (3.27) \]

let

\[
B_k[n_r, n_c] = \sum_{c=0}^{1} \sum_{d=0}^{1} \left\{ \delta_{0,c} + (-1)^{c+1} F_{k\downarrow}[n_r, n_c] \right\} \times \left\{ \delta_{0,d} + (-1)^{d+1} F_{k\rightarrow}[n_r, n_c] \right\} \times A_S \left[ [n'_k[1, n]] + c, [n'_k[0, n]] + d \right], \quad (3.28)
\]

where $\delta$ is the Kronecker delta and

\[
F_{k\downarrow}[n_r, n_c] = n'_k[1, n] - [n'_k[1, n]],
\]

\[
F_{k\rightarrow}[n_r, n_c] = n'_k[0, n] - [n'_k[0, n]]. \quad (3.29)
\]

For $[n_r, n_c]$ that does not satisfy Eq. (3.27), set $B_k[n_r, n_c]$ to 0.

For example, suppose that the sample of $B_k$ at coordinate $[n_r, n_c] = [8, 9]$ corresponded to an $n = 809$, and the 810th column of $n'_k$ was $[35.3, 78.4, 1]^T$, the bilinearly
interpolated value would be

\[ B_k[8, 9] = (1 - 0.3)(1 - 0.4)A_S[35, 78] + (1 - 0.3)(0.4)A_S[35, 79] \]
\[ + (0.3)(1 - 0.4)A_S[36, 78] + (0.3)(0.4)A_S[36, 79]. \] (3.30)

Since estimation of the unknown translation and rotation of the sub-aperture is required, we derive the steps for accumulating the reverse mode of algorithmic differentiation (RMAD) gradient [43] of the interpolation. Let \( X \) be defined as \( \partial E / \partial X \) where \( E \) is defined in Eq. (3.19). The input to the gradient of the bilinear interpolation is the array \( B_k \) with values

\[ B_k[n_r, n_c] = \partial E / \partial B_k[n_r, n_c]. \] (3.28)

The first step of the gradient is finding the arrays \( F_k \downarrow \) and \( F_k \rightarrow \) since the dependence of \( B_k \) on \( \theta_k \) and \( (s_k, t_k) \) comes primarily by way of \( F_k \downarrow \) and \( F_k \rightarrow \). Evaluating the sum over \( c \) in Eq. (3.28) shows that \( B_k \) can be expressed in terms of an element-wise addition of two terms, one involving \( F_k \downarrow \) element-wise multiplied with other quantities and one not involving \( F_k \downarrow \) at all. Consequently, by Eqs. (47) and (49) of [43],

\[ F_k \downarrow [n_r, n_c] = B_k[n_r, n_c] \sum_{c=0}^1 \sum_{d=0}^1 \left\{ \delta_{0,d} + (-1)^{d+1}F_{k \rightarrow}[n_r, n_c] \right\} \]
\[ \times (-1)^{c+1}A_S\left[ [n'_k[1, n]] + c, [n'_k[0, n]] + d \right]. \] (3.31)

for the \([n_r, n_c]\) satisfying Eq. (3.27). For the remaining \([n_r, n_c]\), \( F_k \downarrow [n_r, n_c] = 0 \), since for those indices \( B_k \) has no dependence on \( F_k \downarrow \). By a similar argument,

\[ F_k \rightarrow [n_r, n_c] = B_k[n_r, n_c] \sum_{c=0}^1 \sum_{d=0}^1 \left\{ \delta_{0,c} + (-1)^{c+1}F_{k \downarrow}[n_r, n_c] \right\} \]
\[ \times (-1)^{d+1}A_S\left[ [n'_k[1, n]] + c, [n'_k[0, n]] + d \right]. \] (3.32)
and \( F_{k \to [n_r, n_c]} = 0 \) for the remaining \([n_r, n_c]\).

Evaluating the gradient of Eq. (3.29) is conceptually complicated by the discontinuous nature of the derivative of the floor function when its argument is an integer. For our purposes, the discontinuous nature of the true derivative does not significantly affect optimization if we assume the floor functions \([n'_k[0, n]]\) and \([n'_k[1, n]]\) are constant for small variations of \(n'_k[0, n]\) and \(n'_k[1, n]\), respectively. Presuming \([n'_k[0, n]]\) and \([n'_k[1, n]]\) to be essentially constant yields a gradient rule for Eq. (3.29) of

\[
\overline{n'}_k[n, n] = \begin{cases} 
F_{k \to [n_r, n_c]} & \text{for } \rho = 0 \\
F_{k \downarrow [n_r, n_c]} & \text{for } \rho = 1 \\
0 & \text{for } \rho = 2, 
\end{cases}
\]  

(3.33)

where \(n_r = \lfloor n/N \rfloor\) and \(n_c = n \mod N\) as in Eq. (3.25). Similarly, for the purpose of computing derivatives, we will ignore the dependence of \(B_k\) on \(\theta_k\) and \((s_k, t_k)\) arising from \([n'_k[0, n]]\) and \([n'_k[1, n]]\) appearing explicitly in Eq. (3.28).

Applying Eq. (50) in [43] to Eq. (3.26) yields the gradient step

\[
\overline{R}^{-1}_k = \overline{n'}_k \ast n^T.
\]  

(3.34)

By Eq. (3.24), \(\overline{R}^{-1}_k\) has two explicit dependencies on \(s_k\). Applying Eq. (49) and (59) from [43] to these dependences and accumulating them yields

\[
\overline{s}_k = -C_k \overline{R}^{-1}_k[0, 2] - S_k \overline{R}^{-1}_k[1, 2].
\]  

(3.35)
By a similar argument, 
\[ \bar{t}_k = s_k \overline{R^{-1}_k}[0, 2] - c_k \overline{R^{-1}_k}[1, 2], \] (3.36)
and the gradients of Eq. (3.24) with respect to \( c_k \) and \( s_k \) are
\[ \overline{c}_k = \overline{R^{-1}[0, 0] + R^{-1}[1, 1]} - (s_k + n_0) \overline{R^{-1}[0, 2]} - (t_k + n_0) \overline{R^{-1}[1, 2]}, \]
\[ \overline{s}_k = \overline{R^{-1}[1, 0] - \overline{R^{-1}[0, 1]}} + (t_k + n_0) \overline{R^{-1}[0, 2]} - (s_k + n_0) \overline{R^{-1}[1, 2]}. \] (3.37)

We will use special notation to denote the two quantities that need to be accumulated due to \( \theta_k \) appearing in the definitions of both \( c_k \) and \( s_k \). For simplicity, denote the contribution to \( \overline{\theta}_k \) due to \( c_k \)'s dependence on \( \theta_k \) as \( \overline{\theta}_k^{(c_k)} \). The gradient rules for the sine and cosine do not appear in [43] but they can be derived from the chain rule as follows:
\[ \overline{\theta}_k^{(c_k)} = \frac{\partial E}{\partial \theta_k} = \frac{\partial E}{\partial \cos(\theta_k)} \frac{\partial \cos(\theta_k)}{\partial \theta_k} = \frac{\partial E}{\partial c_k} \frac{\partial \cos(\theta_k)}{\partial c_k} = \overline{c}_k (-\sin \theta_k), \] (3.38)
and for contribution to \( \overline{\theta}_k \) due to \( s_k \)'s dependence on \( \theta_k \):
\[ \overline{\theta}_k^{(s_k)} = \frac{\partial E}{\partial \theta_k} = \frac{\partial E}{\partial \sin(\theta_k)} \frac{\partial \sin(\theta_k)}{\partial \theta_k} = \frac{\partial E}{\partial s_k} \frac{\partial \sin(\theta_k)}{\partial s_k} = \overline{s}_k \cos \theta_k. \] (3.39)

Finally, since \( \theta_k \) appears in both \( c_k \) and \( s_k \), the overall gradient for rotation is
\[ \overline{\theta}_k = \overline{\theta}_k^{(c_k)} + \overline{\theta}_k^{(s_k)}. \] (3.40)
Though this section has derived equations for the gradient of interpolation with respect to just sub-aperture translation and rotation, it is easily adapted to estimate more general transformations of $A_S$ including scaling and shearing. Eq. (3.24) can be altered to include the geometric effects of additional parameters. The gradient propagation steps for these new transformation parameters can then be calculated from the new expression in terms of $R^{-1}$. For even more general translations that do not fit the model of Eq. (3.26), new formulas for $n'$ can be found in terms of the desired geometry parameters and the gradients of these parameters found in terms of $n'$.

We have not attempted to estimate $A_S$ in a UTTDPR algorithm, but, like the algorithms of [4, 39], it is possible to estimate the subaperture transmission function using the metric in Section 3.2. The gradient term with respect to $A_S$ can be found by applying Eqs. (47), (49), and (60) in [43] to Eq. (3.28). The gradient computation is listed in Algorithm 3.1 since it is not a 1 to 1 function.

### 3.8 Gradient of the Error Metric

In this section, we apply the rules for RMAD to find the accumulated gradient terms for the model parameters beginning with $\alpha$ from Eq. (3.12). First, $E = \partial E/\partial E = 1$. Applying Eq. (47) from [43] to Eq. (3.19) yields $E_k = E$ for $k$ in the interval between $k_{\min}$ and $k_{\max}$, inclusive. For other $k$, the gradient steps need not be evaluated as those PSFs do not contribute to the error metric. Applying Eqs. (49) and (52) from [43] to
Algorithm 3.1 Calculation of $A_S$, the gradient with respect to subaperture transmission function. We define $+=$ to be an operator that adds the right operand to the left operand and stores the result in the left operand.

1: for all $n'_r$ and $n'_c$ such that $0 \leq n'_r < N'$ and $0 \leq n'_c < N'$ do
2: 
3: end for

4: for all $n$ such that Eq. (3.27) is True do
5: 
6: for $c = 0$ to 1 do
7: 
8: Q ← $\left\{ \delta_{0,d} + (-1)^{d+1} F_{k\rightarrow}[n_r,n_c] \right\}$
9: end for
10: end for
11: end for
12: return $A_S$

Eq. (3.18) produces

$$M_k = 2w_k \circ (M_k - D_k) \circ E_k = 2w_k \circ (M_k - D_k). \quad (3.41)$$

Evaluating the gradient steps for Eq. (3.17) yields

$$I_k = \beta_k M_k, \quad (3.42)$$

by Eqs. (51) in [43] and there are no other terms despite the many appearances of $I_k$ in the formula for $\beta_k$ and $\gamma_k$ for the reasons we describe now. The construction of
Eqs. (3.62) and (3.63) guarantee that \( \partial E_k / \partial \beta_k = 0 \) and \( \partial E_k / \partial \gamma_k = 0 \), so \( \beta_k = 0 \) and \( \gamma_k = 0 \), respectively. Consequently, we need not consider gradient steps arising due to \( M_k \)'s dependence on \( I_k \) through \( \beta_k \) and \( \gamma_k \) since these are zero by the chain rule. This simple gradient evaluation was made possible by casting the bias-independent metric as Eq. (3.18) rather than the result of [36] for reasons outlined in Section 3.9.

By Eq. (53) in [43], the gradient step for Eq. (3.16) is

\[
\overline{G}_k = 2G_k \circ \overline{I}_k. \tag{3.43}
\]

Next, we have the steps unique to calculating the gradient for \( \alpha \) using the MMDFT. For simplicity, define

\[
h_k = \Omega^T \ast g_k, \tag{3.44}
\]

such that Eq. (3.15) can be written instead as

\[
G_k = h_k \ast \Omega. \tag{3.45}
\]

By Eq. (50) in [43], the gradient steps due to the new Eq. (3.45) are

\[
\overline{h}_k = \overline{G}_k \ast \Omega^H,
\]

\[
\overline{\Omega}^{(G_k)} = h_k^H \ast \overline{G}_k, \tag{3.46}
\]

where \( ^H \) indicates the conjugate-transpose and \( \overline{\Omega}^{(G_k)} \) is the contribution to \( \overline{\Omega} \) due to explicit the appearance of \( \Omega \) in Eq. (3.45) using the notation introduced in Section
3.7. The same matrix-gradient evaluation applied to Eq. (3.44) yields

\[
\overline{\Theta}^{(h_k)} = \overline{h}_k \ast g_k^H, \\
\overline{g}_k = \overline{\Theta} \ast \overline{h}_k, 
\]

(3.47)

where \( \overline{\Theta}^{(h_k)} \) is the contribution to the gradient due to \( \overline{\Theta} \) appearing in Eq. (3.44). Since the values of \( \Theta \) explicitly appear in \( \overline{\Theta} \), but-rearranged, the gradient contribution due to these values appearing in \( h_k \) is

\[
\overline{\Theta}^{(h_k)} = \left[ \overline{\Theta}^{(\omega)} \right]^T = g_k^* \ast \overline{h}_k^T. 
\]

(3.48)

\( \Theta \) appears twice in the model of each PSF and the individual contributions must be accumulated element-wise for all PSFs to find the overall gradient with respect to \( \Theta \),

\[
\overline{\Theta} = \sum_{k=k_{\text{min}}}^{k_{\text{max}}} \overline{\Theta}^{(G_k)} + \overline{\Theta}^{(h_k)}. 
\]

(3.49)

By Eq. (57) in [43], the gradient step associated with Eq. (3.13) is

\[
\overline{\omega} = \mathcal{F} \{ \overline{\Theta} \circ \overline{\Theta}^* \}. 
\]

(3.50)

Accumulating the gradient due to appearances of \( \alpha \) in Eq. (3.12) gives

\[
\overline{\alpha} = -2\pi \sum_{n=0}^{N-1} (n - n_0) \sum_{m=0}^{M-1} (m - m_0) \overline{\omega}[n,m]. 
\]

(3.51)

This concludes the gradient steps specifically involved for the MMDF since \( \overline{g}_k \), which is needed by all future steps, was derived in Eq. (3.47).
Next we turn our attention to evaluating the gradients for $s_k$, $t_k$ and $\theta_k$. Assuming $B_k$ in Eq. (3.6) is real, applying Eqs. (49) and (67) in [43] to Eq. (3.6) yields

$$B_k = \Re \{ f_k^* \circ g_k \},$$

(3.52)

which is needed for calculating the gradients for $s_k$, $t_k$ and $\theta_k$ using the results of Section 3.7.

The steps necessary to find the gradient with respect to $d_l$ are as follows. Beginning with Eq. (3.6),

$$f_k = B_k \circ g_k.$$  

(3.53)

Applying Eqs. (49) and (67) in [43] to Eq. (3.5) gives

$$\overline{A_F^{(f_k)}} = \Re \{ \exp (-i\phi_k) \circ \overline{f_k} \},$$

(3.54)

where $\overline{A_F^{(f_k)}}$ is the contribution to $\overline{A_F}$ due to $A_F$ appearing in Eq. (3.5) for the $k$th PSF.

Accumulating these contributions element-wise with

$$\overline{A_F} = \sum_{k=1}^{K} \overline{A_F^{(f_k)}},$$

(3.55)

and evaluating the thresholding operation in Eq. (3.3) yields

$$\overline{C}[n_r, n_c] = \begin{cases} 
\overline{A_F}[n_r, n_c] & \text{for } C[n_r, n_c] > -1 \\
0 & \text{otherwise.}
\end{cases}$$

(3.56)
Then evaluating Eq. (3.4) finds

$$
\overline{d}_i = \sum_{n_r=0}^{N-1} \sum_{n_c=0}^{N-1} (Z_i \circ \overline{C})[n_r, n_c],
$$

(3.57)

by Eq. (58) in [43].

Finally, we turn to evaluating the gradients with respect to the phase coefficients $a_{2,k}$, $a_{3,k}$ and $a_k$. By Eqs. (47) and (57) in [43], the gradient steps corresponding to Eq. (3.5) are

$$
\overline{\phi}(f_k) = \Im \{ f_k^* \circ f_k \},
$$

(3.58)

where $\overline{\phi}(f_k)$ is the contribution due to the appearance of $\phi$ in $f_k$ of the $k$th PSF. In the shared linear phase case, $\phi_k$ has the same value in each PSF and thus the corresponding gradient contributions can be accumulated element-wise to find

$$
\overline{\Phi} = \sum_{k=1}^{K} \overline{\phi}(f_k),
$$

(3.59)

and then from evaluating Eq. (3.2) we see that

$$
\overline{a}_j = \sum_{n_r=0}^{N-1} \sum_{n_c=0}^{N-1} (Z_j \circ \overline{\phi})[n_r, n_c],
$$

(3.60)

for $j = 2, \ldots, J$ by Eq. (58) in [43]. In the unshared linear phase case, the linearity of Eq. (3.1) means that $\overline{a}_j$ can still be calculated by Eq. (3.60) but just for $j = 4, \ldots, J$. The gradients for the linear phase terms $\overline{a}_{2,k}$ and $\overline{a}_{3,k}$ must be calculated from the individual contributions $\overline{\phi}_k$ according to

$$
\overline{a}_{j,k} = \sum_{n_r=0}^{N-1} \sum_{n_c=0}^{N-1} (Z_j \circ \overline{\phi}_k)[n_r, n_c].
$$

(3.61)
3.9 Coefficients of the Bias Independent Metric

The derivation of the bias independent metric in [36] does not include explicit formula for both the coefficients $\beta_k$ and $\gamma_k$. We derive these coefficients in different terms than one would find by substituting the equations of [36] into one another.

The value of $\beta_k$ that puts $E_k$ in Eq. (3.18) at the extremum has the property that

$$0 = \frac{\partial E_k}{\partial \beta_k} = \sum_{m_r, m_c} w_k \circ (\beta_k I_k + \gamma_k 1 - D_k) \circ 2I_k$$

$$= 2\beta_k \langle I_k, I_k \rangle + 2\gamma_k \langle I_k \rangle - 2 \langle D_k, I_k \rangle,$$

(3.62)

where $\langle X, Y \rangle = \sum_{m_r, m_c} w_k \circ X \circ Y$ and $\langle X \rangle = \sum_{m_r, m_c} w_k \circ X$. The value that puts $\gamma_k$ at extremum is, by a similar calculation,

$$0 = \frac{\partial E_k}{\partial \gamma_k} = 2\beta_k \langle I_k \rangle + 2\gamma_k \hat{\omega}_k - 2 \langle D_k \rangle,$$

(3.63)

where $\hat{\omega}_k = \sum_{m_r, m_c} w_k$. Linear Eqs. (3.62) and (3.63) have the solution

$$\beta_k = \frac{\hat{\omega}_k \langle D_k, I_k \rangle - \langle I_k \rangle \langle D_k \rangle}{\hat{\omega}_k \langle I_k, I_k \rangle - \langle I_k \rangle^2},$$

$$\gamma_k = \frac{\langle I_k, I_k \rangle \langle D_k \rangle - \langle I_k \rangle \langle D_k, I_k \rangle}{\hat{\omega}_k \langle I_k, I_k \rangle - \langle I_k \rangle^2}.$$  

(3.64)

Each evaluation of $E_k$ for unique parameter values requires calculating $\langle I_k \rangle$, $\langle D_k, I_k \rangle$ and $\langle I_k, I_k \rangle$. The values of $\langle D_k \rangle$ and $\hat{\omega}_k$ only need to be evaluated once for particular $D_k$ and $\omega_k$. 
References


4

Efficient Prescription Retrieval through Reverse-Mode Algorithmic Differentiation

This thesis has thus far considered phase retrieval as a means for estimating the phase aberrations of an optical system. Although these aberrations can be of inherent interest, often the ultimate interest is the mechanical state of the optical system that induced those aberrations. We will refer to this unknown mechanical state as the prescription of the optical system to be consistent with the terminology of lens design. This prescription will be understood in terms of conventional lens design quantities like the radius of curvature for a surface or the thickness of an element. We refer to these quantities as the unknown prescription parameters in systems where they are not known to the desired accuracy.

Recovering prescription information from measured point spread function (PSF) data can be done in various ways, but this chapter considers a nonlinear optimization strategy called prescription retrieval [1]. In Section 4.1, prescription retrieval and its advantages relative to similar estimation methods involving phase retrieval and the
linear optical model are introduced. We also introduce the computational challenges associated with prescription retrieval that have posed a barrier to its use to date. One of these challenges is what we call the local search challenge, and this chapter finds a new solution for this problem utilizing the reverse mode of algorithmic differentiation (RMAD) [2–4]. In Section 4.2, we develop an RMAD gradient formulation for arbitrary scalar metrics that utilize the output of a sequential raytrace. Later, in Section 4.3, we construct computational models for the field on the reference sphere and the diffraction integral. The models from Sections 4.2 and 4.3 are utilized in an error metric in Section 4.4. Finally, in Section 4.5, the results of a prescription retrieval simulation experiment are discussed.

4.1 Promise and Challenge of Prescription Retrieval

Since phase retrieval is by now familiar to the reader, we introduce prescription retrieval in the context of estimating prescription parameters using conventional phase retrieval. The process of interpreting phase retrieval results is greatly simplified if the unknown prescription parameters only act on an optical surface near a pupil plane. In these cases, the retrieved aberrations are readily interpreted as an optical phase delay induced by an error in the shape of the surface. For example, in the Hubble Space Telescope (HST) primary-mirror mishap, most of the spherical aberration was due to the primary mirror which was also the aperture stop. Variants of phase retrieval could then be applied to estimate the aberration [1, 5–9]. For the James Webb Space Telescope (JWST), the connection between phase retrieval results [10, 11] and the prescription
parameters of the telescope has been established [12, 13]. These parameters include the 6 degrees of freedom due to rigid body motion for each primary mirror segment as well as the intra-segment mirror aberrations. However, misalignments of the secondary mirror, which is not in a pupil plane, also contribute optical path difference to the retrieved phase. This complicates the correspondence between the phases estimated in phase retrieval and the actual prescription parameters. Similarly, in the HST case, Redding et al. [1, 8] found it necessary to model the spherical aberration due to Hubble’s secondary and primary mirrors. Their technique, called prescription retrieval, coupled a raytrace model that incorporated aberrations from both surfaces with a diffraction calculation that modeled point spread functions (PSFs). Values for the unknown prescription parameters were adjusted by nonlinear optimization until the model-predicted PSFs best fit the measured PSFs. It is also possible to model the secondary mirror aberrations using conventional phase retrieval with additional paraxial propagations between planes representing the primary and secondary mirrors [14]. However, in this chapter, we focus on prescription retrieval as an optical metrology tool and its advantages outlined below.

As with phase retrieval [15], prescription retrieval can, with an appropriate optimization strategy, be a maximum-likelihood estimator for the unknown prescription parameters [1]. Prescription retrieval also readily adapts to unknowns on multiple surfaces that are not in pupil planes. While phase retrieval can incorporate aberrations in multiple planes [14], the unknowns should act on the field in a plane or spherical surface to and from which there exist easily calculated propagations. For instance,
most unknowns associated with a surface alter the optical path lengths traversed by light in the neighborhood of the surface. These changes in optical path lengths can only be associated with phase changes in the propagating fields if the fields where these phase changes were imparted can be calculated. Also, after being affected by the phase change, the resulting fields must be numerically propagated to the next portion of the optical system. For efficient optimization, such propagations should be paraxial and have tractable discrete sampling requirements. This can prove difficult in some metrology situations. For example, consider a highly-diverging beam incident on an F/1 parabolic surface with a mild additional surface-height defect. Assume that the radius of curvature surface is very different from the radius of curvature of the wavefronts incident upon and leaving the surface. It is difficult to find a plane or spherical curve on which the beam leaving the surface can be represented without an extremely fine sampling of the field. Evaluating diffraction calculation on arbitrary surfaces has generated various propagation methods [16–20] but we are unaware of their use in the context of iterative phase retrieval. In contrast, raytracing a fast parabolic surface is relatively straightforward and thus prescription retrieval is a potentially viable strategy for inferring unknown prescription parameters from PSF data. The estimation of freeform surfaces with high degrees of surface departure has already been shown with a process known as reverse-raytracing [21]. The deterministic reconstruction of a single surface performed by reverse-raytracing is, however, fundamentally different than prescription retrieval which seeks to estimate unknowns on multiple surfaces using an optimization strategy.
With the exception of [22, 23], phase retrieval usually avoids modeling variations in the PSFs due to being at different points in the field of view or considering other object angles. This is because these other points have different generalized exit pupil functions, and phase retrieval in optical metrology is usually about inferring the field of one exit pupil. The anisoplanatic behavior of an optical system does not aid such phase retrieval estimation without a detailed model that constrains different exit pupils to behave jointly in a certain fashion, as in Chapter 6 of Ref. [22]. The optical prescription of a system is a model for relating different exit pupils. Thus, prescription retrieval is ideally positioned to integrate PSF measurements acquired from different field angles with dissimilar, but intrinsically related, exit pupils.

The final advantage we discuss for prescription retrieval is its ability to handle nonlinearity in the connection between prescription parameters and the observed data. Consider an abstract metrology problem with an unknown column vector of prescription parameters $\theta$ having $N_\theta$ elements. $N_\theta$ may range from just a few unknowns [1] to over a hundred. Similarly, let the data, $D'$, be a vector of observed PSF intensity samples with $N_{D'}$ elements. These PSFs may have been acquired at multiple places in the field of view, in multiple wavelengths or with various diversity arrangements such as focus or transverse-translation diversity as describe in Section 1.2. Therefore, $N_{D'}$ may range from dozens of pixels for one compact PSF to a few million pixels assembled from tens of spread-out PSFs acquired at different wavelengths, field points and diversities. By intensive raytracing and diffractive calculations, we can find a mathematical model, $I(\theta)$, that attempts to match $D'$, although $I$ has a highly nonlinear relationship
with $\theta$. This nonlinearity arises from: (i) the nonlinear dependence of optical path lengths on prescription parameters, (ii) calculating the complex field in the exit pupil due to optical path lengths, and (iii) the dependence of intensity in the image plane on the field in the image plane. In practice, finding the $\theta$ that yields the result of $I(\theta)$ similar to $D'$ has proved sufficiently challenging that linearizing approximations have been made to solve prescription-retrieval-like problems. For instance, one can find prescription parameters using linear optical model [24–26] (LOM) methods. The LOM is a common unifying simplification of the nonlinearity (i) above that assists joint mechanical, thermal and optical engineering on a reference design for an optical system. Instead of a model that predicts PSF intensity like $I$, the LOM method uses a raytraced calculation, $W(\theta)$, that predicts distinct optical path differences or Zernike coefficients of the wavefront aberrations for all measured field angles, wavelengths and diversities. For this method, nonlinearity sources (ii) and (iii) above are addressed using parametric phase retrieval techniques [14, 27, 28] or iterative transform phase retrieval methods [10, 11] to estimate Zernike coefficients of the exit pupil phase or actual optical path differences. Let the result of this phase retrieval solution be $R(D')$ and choose a nominal prescription, $\theta_0$, about which the optical behavior is to be linearized. By the method of finite differences [26] or sophisticated differential raytracing [24, 25] one evaluates the sensitivity matrix

$$S(\theta_0)[p, q] = \frac{\partial W[p]}{\partial \theta[q]} (\theta_0),$$

(4.1)
which is the Jacobian matrix of $W$, evaluated at $\theta_0$. Then, the nominal exit-pupil wavefront,

$$R_0 = W(\theta_0), \quad (4.2)$$

can be found by the raytracing computation $W$. The first-order approximation linking PSFs to prescription parameters is

$$R(D') - R_0 \approx S \ast (\theta - \theta_0), \quad (4.3)$$

where $\ast$ indicates matrix multiplication. Consequently, one may attempt to solve the linear equation

$$R(D') - R_0 + S \ast \theta_0 \approx S \ast \theta, \quad (4.4)$$

for the unknown $\theta$ since everything on the left and $S$ are known.

Jurling [23] showed that it was possible to integrate solution of the LOM model in Eq. (4.4) in a multi-field phase retrieval solution. This was similar to earlier work by Bolcar [22] which assumed wavefront aberrations at various field points could be related by a double-Zernike expansion, although this did not attempt to resolve the challenges of the prescription retrieval problem addressed by the LOM. In Jurling’s LOM method, phase retrieval optimization directly estimates prescription parameters $\theta$ assuming that Eq. (4.4) is an equality. This reduces the errors associated with finding $R(D')$ for a single field point as it constrains the errors of phase retrieval to be consistent with the LOM at all field points. Bolcar [22] observed a similar noise-
reduction effect using the double-Zernike expansion wavefront model. A limitation of the LOM is that Eq. (4.1) is evaluated at just the nominal prescription, and a similar limitation applies to the double-Zernike expansion model. If the true unknown prescription is far from the nominal guess, Eq. (4.3) is inaccurate and using it to estimate $\theta$ will yield additional errors. In contrast, a prescription retrieval solution, if it is found, would not encounter such challenges as it would incorporate the nonlinear effects (i)-(iii), listed above, directly. Another advantage of prescription retrieval is that it can exploit changes in pupil shape due field angle and prescription that might not be represented in $\mathcal{R}(D')$.

Given the above list of the advantages of prescription retrieval, one might wonder why it has seen so little use [1]. Even if we generalize prescription retrieval to also include methods that use Shack-Hartmann wavefront sensing data rather than PSFs, we found just two more usages [29, 30]. We believe this lack of usage to be principally due to the computational complexity of finding solutions when there are many unknown parameters. This complexity has two components: (i) the stagnation problem and (ii) the local search problem. Both are challenging in prescription retrieval and will be discussed below. We define these problems in terms of the minimization of a sum of squared-differences error metric,

$$
E(\theta) = \sum_{u=0}^{N_\text{pix} - 1} (I(\theta)[u] - D'[u])^2, \tag{4.5}
$$

where $u$ is the index of a pixel of measured PSF data. While this chapter focuses on pixel data derived from PSFs, $D'$ could also be wavefront measurements expressed in
Zernike polynomial coefficients acquired using some other wavefront sensing method like phase retrieval or a Shack-Hartmann sensor. Prescription retrieval is an algorithm that finds

$$\hat{\theta} = \arg \min_{\theta} E(\theta).$$

(Prescription (4.6))

Prescription retrieval can use a method like nonlinear least squares, conjugate gradient or a quasi-Newton method [3] to find the $\theta$ that minimizes $E(\theta)$.

The stagnation problem is that, after starting at an initial estimate for $\theta_0$, the optimizer finds a local minimum $\hat{\theta}$ where $\nabla E$ is zero but where $E(\hat{\theta})$ is not as small as $E$ evaluated at the true prescription parameters. This occurs because $E$ may have many local minima in addition to the global minimum. In this case, $\hat{\theta}$ may be very unlike the true prescription parameters of the system. We distinguish the stagnation problem from the prescription problem actually being ill-posed, however. The problem is ill-posed if there is no unique global minimum of $E$ near the true prescription parameters and the PSF data is insufficiently descriptive to specify a unique correct solution. Returning to the stagnation problem, it has much in common with the stagnation and convergence issues in parametric phase retrieval [31, 32]. Similarly, there is likely no general solution to finding the global minimum in prescription retrieval. We have noted that randomizing the starting guess greatly aids phase retrieval involving pupil phase models that are parameterized by small numbers of Zernike polynomial coefficients [33]. To the extent that prescription retrieval is similar to a low-order phase parameterization of the exit pupil field, this insight may be transferable to
prescription retrieval. For phase retrieval problems that recover a large number of unknown complex field values, random starting points have proven extremely useful [28]. In [30, 34, 35], simulated annealing is applied to a parameterized phase retrieval problem to solve the convergence issues associated with local minima of an error metric, and this approach may be applicable to the stagnation problem of prescription retrieval. It does not, however, have provisions for a gradient-based search, and this may limit its utility beyond locating quality starting guesses for the methods listed above that utilize gradients. Other methods which combine randomization and gradient search such as basin hopping [36] or stochastic global search methods [37] may be useful. One may also apply an optical designer's intuition about the effect of perturbing prescription parameters to aid the solution process. For instance, to first order, most deviations in prescription will yield a change in defocus. Consequently, preconditioning of the prescription retrieval problem may be necessary if the nominal prescription differs markedly in defocus from that of the measured data. For instance, before applying optimization one could perturb one or more surfaces in the nominal prescription until the prescription has a defocus similar to the measured data. In our work, we have chosen to perturb the surface for which the curvature tolerances are most sensitive under the assumption that this surface would not have to move very far. We have found this necessary to improve the chances of successful retrieval in some cases. Similarly, if one uses a stochastic method, it is likely profitable to move randomly within a constrained subspace of all prescriptions having a defocus similar
to the measured data. Beyond these insights, we do not propose new solutions to the stagnation problem.

We define the local search problem to be that many repeated calculations of $\nabla E$ becomes too computationally expensive and requires an excessive amount of processing time. This problem varies with (i) the accuracy requirements, (ii) the unknown prescription parameters, (iii) the optical system, (iv) the available computational resources, and (v) the magnitude of the stagnation problem. For instance, if issues (i)-(iii) lead to ill-posed prescription retrieval using a small number of PSFs, additional PSFs may need to be included in the error metric that require additional processing time. Alternatively, if the stagnation problem is severe, many optimizations with different starting points may need to be employed, multiplying the overall expense of the calculation.

Evaluating $\nabla E$ by the method of finite differences, as is commonly done in lens design optimization, can be prohibitively computationally expensive. Suppose that calculating $I(\theta)$ requires tracing $N_W$ individual rays. Evaluating $\nabla E$ at $\theta$ once using single-ended finite differences [3] will require $N_W (N_\theta + 1)$ raytraces. Doing this for $N_V$ iterations of a nonlinear optimization method like conjugate gradient will require a total of $N_V N_W (N_\theta + 1)$ raytraces. Suppose a hypothetical prescription retrieval problem where it takes 4 seconds to trace $N_W$ rays and calculate $I(\theta)$. Further, suppose that there are $N_\theta = 100$ unknowns and that it takes $N_V = 200$ iterations to find the solution. Such a hypothetical problem using finite-difference gradients in the optimization would take more than 22 hours. In this chapter, though, we wish to
solve prescription retrieval problems with orders of magnitude more unknowns that are even less suited to finite-differences.

A more sophisticated approach to computing a gradient might use the results of a differential raytrace \([24, 38, 39]\) to evaluate Eq. (4.1) as quickly as practicable. One could also take the computational process that generates \(\mathbf{W}\) and apply the forward mode of algorithmic differentiation \([3]\) to get a similar result. Then evaluate

\[
\nabla_{\theta[q]} E(\theta) = \sum_{p=0}^{N_W-1} \frac{\partial E}{\partial \mathbf{W}[p]} \frac{\partial \mathbf{W}[p]}{\partial \theta[q]}(\theta)
\]

(4.7)

where \(\partial E/\partial \mathbf{W}[p]\) is the complicated gradient calculation that is common to optimization-based phase retrieval problems \([4, 14]\). While it is conceivable that one could compute Eq. (4.7), we contend that this, and the method of finite differences, are inefficient ways to evaluate the gradient of a scalar function because their overall computational effort is linear in \(N_\theta\). Since differentiation methods with complexities that are sub-linear in the number of unknowns are guaranteed to exist by the reverse-mode of automated differentiation \([2, 3]\), we know that there are better ways to calculate the gradient for prescription retrieval. To see that this is likely, suppose that we compute \(E\) using a deterministic computer program doing a certain number of elementary operations. Each of those elementary operations has a gradient that could be expressed in about three additional elementary operations \([40]\). Since the chain rule could be used to join all of these elementary gradient operations into a new program that finds the gradient, there is an upper bound on the computational effort of the gradient. This concept is known as the cheap derivative principle and the program
that one constructs to find a gradient with such modest complexity is the result of applying the reverse mode of algorithmic differentiation (RMAD) [2, 3]. The computational complexity of evaluating the gradient by RMAD can be at most five times the complexity of evaluating the error metric itself [3]. For instance, if it requires a computational effort of tracing $N_W$ rays to evaluate $E$, the RMAD gradient can take an effort no greater than $5N_W$. Critically, this bound is independent of the number of unknowns. Consequently, an RMAD gradient evaluation has a considerable advantage over the method of finite differences and Eq. (4.7) when $N_\theta$ becomes large since the complexity of those methods is approximately linear in $N_\theta$. RMAD techniques began appearing in various fields in the late 1960s [41]. However, the application to phase retrieval did not come until Jurling [4] noted that the laborious symbolic differentiation methods applied to accelerate gradient calculation such as [14, 42, 43] could be simplified and understood using the RMAD. For the remainder of this chapter, we apply these results to the local search problem in prescription retrieval.

### 4.2 Raytrace Calculation and Gradient

In this section, we will apply the reverse mode of algorithmic differentiation to the classic problem of raytracing a sequential optical system. We adopt the notation of [3, 4] where $\bar{x} = \partial E / \partial x$ is referred to as the adjoint variable for $x$ for the reason described below. There will be one adjoint variable for essentially every quantity
defined in this chapter. The RMAD process involves taking the steps of the calculation

\[ R = W(\theta) \]  

(4.8)

and finding a gradient calculation \( \overline{W} \) such that if \( \overline{R}[p] = \partial E / \partial R[p] \),

\[ \overline{\theta} = \overline{W}(\overline{R}) \]  

(4.9)

where \( \overline{\theta}[p] = \partial E / \partial \theta[p] \). \( \overline{\theta} \) and \( \overline{R} \) are known as adjoint variables [3] because, in linear algebra terms, Eq. (4.9) can be phrased as the multiplication of the adjoint of the Jacobian matrix of \( W \) with a column vector realization of \( \overline{R} \). It is not necessary to explicitly calculate the Jacobian matrix, however, if the gradient of the error metric is all that is desired [3]. For simplicity, gradient calculations involving the adjoint variables will be grouped into smaller computations like ray transfer and refraction.

First, we pause to recognize the long history of sophisticated differentiation methods in optical design. Feder [38] described the differential raytrace, an efficient method for calculating \( \partial W[p] / \partial \theta[q] \) for arbitrary parameter \( \theta[q] \). This efficiency was echoed in the generalized raytrace formalism of [39], which could handle arbitrary freeform surfaces, and extended further to higher order partial differentials in [44]. The differential raytrace has been found useful for accelerating tolerancing [45] and for optimization using error metrics involving MTF [46–48]. In MTF optimization, a gradient involving a mixture of analytic derivative formulas, using differential raytrace efforts, and finite-differenced gradients can be combined to optimize a complicated lens design merit function [47, 48]. In general, though, error metrics or merit functions with
fast fully-analytic gradients have found few applications in lens design. The analytic gradient methods described in [49, 50] formulate metrics in ways analogous to Eq. (4.7) and are likely suboptimal relative to an RMAD formulation.

Our RMAD formulation is also not without faults. For instance, $\mathbf{W}$ requires the storage of many more intermediate quantities found during the computation of $\mathbf{W}$ as is generally true of reverse mode accumulation methods [3]. Also, we do not claim that the RMAD formulation is the best possible gradient formulation, just that it is better than those gradient calculations that are clearly linear in $N_\theta$ for the prescription retrieval problem. The RMAD gradient method may not be entirely novel, either. The propagation of gradients represented by the function in Eq. (4.9) is also reminiscent of the single-step matrices of classic differential raytracing [38, 39] since those are also applications of the chain rule. Indeed, many aspects of our RMAD formulation may be equivalent to or inferior to the equivalent differentials found in the classic methods and we hope that such oversights can be corrected in future work. For instance, the differential raytrace methods [24, 38, 39, 44] make particularly effective use of the implicit functions solved by the translation and refraction calculations for arbitrary surfaces. In this work, we only utilize the implicit function theorem (IFT) [40] to find the gradient steps for the intersection of a ray with an aspherical or freeform surface. Future work could apply this theorem to the refraction calculation. Also, the IFT could be used with the ray intersection with a spherical surface to find a different gradient formula than the naive application of RMAD provided in Section 4.2.2. Still, we believe that the process of accumulating dependencies due to a single variable
appearing in multiple places has no direct equivalent in differential raytracing. This is particularly important in prescription retrieval where raytrace quantities from multiple surfaces, like the stop and reference sphere surface, become part of the error metric. The single-step matrices, as defined in [38, 39], do not have the notion of accumulation [3, 4] that is central to RMAD. This is required to combine gradient contributions due to the use of variables from multiple surfaces in the error metric.

The procedures for tracing a general skew-ray through a sequential optical system have changed little in recent years and appear in accessible literature [39, 51–54]. Our intent is not to re-derive these equations but to reference them in a way that allows easy application of RMAD to spherical surfaces and certain freeform surfaces. When possible, we follow the notation of [51, 53] but adjust it to include the notion of many rays propagating simultaneously as it is relevant to structure of the RMAD gradient calculation.

For simplicity, assume that the optical system can be described by a sequence of surfaces with ascending indexes beginning with 0. We define the first (spherical) surface to be centered on the point target being imaged and label it the object-space reference sphere. Later in the sequence of surfaces, there will be a stop surface with index \( s \) whose impact is discussed in Section 4.3.1. The last surface, with index \( r \), will be designated the image-space reference sphere and it is to be centered on some point in image space where the modeled PSF is to be calculated. The synthesis of PSFs given ray intersections on the reference sphere is described in Section 4.3. Thus
described, this geometry is the consistent with the conventions for calculation of the wavefront aberration and evaluation of the diffraction integral [54–56].

It is, however, conventional to center the image-space reference sphere on the point designated by the real chief or principle ray [53] crossing the image plane. This is useful for lens design as it fixes the wavefront aberration to be evaluated about a point near where many of the rays are crossing the image plane. In prescription retrieval, however, we know that most of the rays cross the image plane near the centroid of the measured PSF. When we refer to the PSF, we mean the 2-dimensional intensity function on the image plane and not an alternate quantity like the 3-dimensional PSF in the neighborhood of the detector. Provided the centroid point and an alternate chief ray intersection point are within an isoplanatic patch of the optical system, they will predict wavefront aberrations which differ only by the displacement theorem [54]. To ensure this, we require that the regions of the PSF with significant intensity be contained within an isoplanatic patch of the system. This reasonable restriction would have been made anyway to use the diffraction integral [54–56] to calculate predicted intensity values over the entire PSF. This is equivalent to restricting the algorithm to the field points of an optical system for which the sine condition [57, 58] holds. By centering the image-space reference sphere on a point that is known prior to optimization, we avoid the need to differentiate a chief ray aiming computation, thereby considerably simplifying an arduous calculation.

Let \((x_j, y_j, z_j)\) be the coordinate of the intersection of a ray with the \(j\)th surface. We follow the usual right-handed coordinate system with \(z\) being along the optical
axis if none of the optical surfaces are rotated [52, 53]. Also, since much of the
raytrace calculation involves 3-dimensional coordinate arithmetic, we also treat the
three spatial coordinates x, y and z simultaneously. This allows us to limit the number
of mathematical operations that must be enumerated in the formula but also provides
for easy “vectorization” when implementing the formula in a high level programming
language such as Fortran, Matlab or Python. Let \( x_j[\rho, \kappa] \) be a 3 by \( K \) array consisting
of column vectors \( [x \ y \ z]^T \) representing the intersection of the \( \kappa \)th ray on the \( j \)th
surface in the local coordinate system of the \( j \)th surface. The index \( \rho \) indicates which
of the three coordinates is specified, where \( \rho = 0 \) refers to the first row of \( x_j \): a vector
of the \( x \) ray intersection coordinates on the \( j \)th surface. \( \rho = 1 \) indicates the second
row which is a vector of the \( y \) coordinates for the same intersections, and similarly for
the \( z \) coordinates for \( \rho = 2 \).

In this section and Section 4.3, many quantities such as \( x_j \) will be introduced which
are unique to the evaluation of a particular intensity at a given wavelength, field point,
polarization or other condition. Sometimes these conditions must be evaluated in
multiple cases to predict the intensity of a single PSF using a programming construct
like a loop. In this case, we refer to these additional dimensions of evaluation as
loop dimensions. For instance, a polychromatic PSF can be modeled by looping
over wavelength dimension and summing the resulting intensities while applying a
weighting to account for variations in the polychromatic spectrum. Each sample of the
polychromatic intensity requires its own raytrace and diffraction integral due to varia-
tions in index of refraction and diffraction effects with wavelength. For a complicated
PSF model, there may be many such loop dimensions like polychromatic spectrum, field angle, polarization or other conditions being applied simultaneously to estimate just one PSF. For generality and compactness of notation, we do not explicitly index a particular instance of a variable like $x_j$ with regards to the state of these loop dimensions.

Returning to the description of the raytrace, let $X$, $Y$, and $Z$ be the directional cosines of a ray with respect to the local coordinate vectors of a surface after intersection with that surface as in [53]. We define $X_j[\rho, \kappa]$ as an array of column vectors $[X \ Y \ Z]^T$ representing the directional cosine of the $\kappa$th ray after intersection with the $j$th surface in the local coordinate system of the $j$th surface. Since this array is of directional cosines, it is required that $\sum_{\rho=0}^{2} X_j^2[\rho, \kappa] = 1$ for all $j$ and $\kappa$.

The primary inputs to this raytrace calculation are the prescription parameters listed in Table 4.1. The conic constant $k_j$ for aspheric surfaces and the polynomial coefficients $a_j$ for freeform surfaces are described in Section 4.2.3, while the specific meaning and orientation of the surface decentrations and rotations are discussed in Section 4.2.1. The rotation angles have been primed to make them distinct from unrelated variables in Chapter 3. Another input to the raytrace are $x_0$ and $X_0$ which represent a fan of rays crossing in object-space where the point target is known to be located. If the object is at a finite conjugate, the $0$th surface may intersect the object point provided $x_0[\rho, j] = 0$.

One output of the raytrace calculation is the optical path distance $o[\kappa]$ measured from the object-space reference sphere to the image-space reference sphere for the
Table 4.1: Prescription parameters of the raytrace for the $j$th surface.

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Quantity</th>
</tr>
</thead>
<tbody>
<tr>
<td>$c_j$</td>
<td>Surface curvature (reciprocal of the radius of curvature)</td>
</tr>
<tr>
<td>$k_j$</td>
<td>Conic constant of surface (for aspheric surfaces)</td>
</tr>
<tr>
<td>$a_j$</td>
<td>Array of polynomial coefficients of the surface sag equation for a freeform surface</td>
</tr>
<tr>
<td>$t_j$</td>
<td>Thickness of the media after the surface to the next surface</td>
</tr>
<tr>
<td>$n_j$</td>
<td>Index of refraction of the media after the surface</td>
</tr>
<tr>
<td>$\Xi_j$</td>
<td>Decentration of surface vertex in $x$ direction</td>
</tr>
<tr>
<td>$\Upsilon_j$</td>
<td>Decentration of surface vertex in $y$ direction</td>
</tr>
<tr>
<td>$\alpha_j'$</td>
<td>Left-handed rotation of surface about $+x$ axis</td>
</tr>
<tr>
<td>$\beta_j'$</td>
<td>Left-handed rotation of surface about $+y$ axis</td>
</tr>
<tr>
<td>$\gamma_j'$</td>
<td>Right-handed rotation of surface about $+z$ axis</td>
</tr>
</tbody>
</table>

$x$th ray. Alternately, the optical path distance can be measured from the object point to the image-space reference sphere. However, one must also know where on the image-space reference sphere this optical path distance was measured so the raytracer must also output $x_r$. To this optical path distance, we must also associate an amplitude which we will calculate from the ray’s intersection with the stop surface, so $x_s$ is also required. These steps may be useful for models that depend on angle of incidence, such as polarization, that one may choose to base an error metric upon.

For generality, we formulate the gradient of the raytrace such that any result of the raytrace computation can be used in later models on which the error metric depends.
Further, the gradient of the error metric with respect to all the parameters listed in Table 4.1 will be provided. Evaluating the raytrace involves the following steps:

1. Set the current surface index $j$ to 1 and $a[\kappa] = 0$.

2. **Coordinate Transformation**: Shift and rotate the coordinates $x_{j-1}$ and $X_{j-1}$ according to $\Xi_j$, $\Upsilon_j$, $\alpha'_j$, $\beta'_j$, and $\gamma'_j$ specified for the $j$th surface to find the ray intersections, $x'_{j-1}$, and directional cosines, $X'_{j-1}$, for the $(j-1)$th surface in the local coordinate system of the $j$th surface. Details of this step are provided in Section 4.2.1.

3. **Ray Transfer**: Given $x'_{j-1}$, $X'_{j-1}$, and $c_j$, calculate $x_j$ and $E_{1,j}[\kappa]$, the cosine of the angle of incidence of the $\kappa$th ray with the $j$th surface. Also, find the physical path distance $L_j[\kappa]$ along the $\kappa$th ray between the $(j-1)$th and $j$th surfaces. Details of this step are provided in Section 4.2.2, for spherical surfaces. For aspherical or freeform surfaces, the calculation is described in Section 4.2.3, and additionally involves calculating $N_j$, an array of the directional cosine of the surface normal at the intersection of each ray with the surface.

4. **Accumulating Optical Path Distance**: Add to the total optical path distance $L_j$ times the index of refraction $n_{j-1}$ as described in Section 4.2.4.

5. If $j = r$, the image-space reference sphere has been intersected and the necessary outputs calculated. Stop evaluation of the raytrace and continue to the diffraction integral.

6. **Refraction**: Given $x_j$, $X'_{j-1}$, $E_{1,j}[\kappa]$, $c_j$, $n_j$, $n_{j-1}$ and, in the case of an aspherical or freeform surface, $N_j$, calculate $X_j$ as described in Section 4.2.5.

7. Add 1 to $j$ and go to Step 2 to continue the rays on to the next surface.
The preceding steps are inspired by [59] which differ slightly from descriptions [52, 53] that do not account for translations and rotations between surfaces.

Prior to the evaluation of the raytrace gradient, any models that contribute to the error metric and which use raytrace values must have already had their gradient steps executed. These other gradient steps must have already been executed so that the gradient contributions due to the dependence of these models on outputs of the raytrace can be accumulated to the proper adjoint values before beginning the raytrace gradient. For prescription retrieval, the models that depend on the raytrace are described in Sections (4.3)-(4.4). The gradient steps of these models yield $\delta$, $\overline{x_r}$, and $\overline{x_s}$ so these adjoint variables will be initialized prior to the gradient of the raytrace. However, this discussion of the raytrace gradient does not limit itself to the specific prescription retrieval model in this chapter. For error metrics with different dependencies on the output of the raytrace computation, this text cannot list which adjoint variables have been initialized by prior gradient steps and which have not. It is only required that any outputs of the raytrace that become part of the error metric have their associated adjoint variable initialized to the correct value prior to the raytrace gradient. To notate this generality of the raytrace gradient, we employ a special accumulation operator that explicitly initializes any adjoint variables not previously initialized at the time they are required. This is discussed further in Section 4.2.1.

A high-level view of the gradient computation shows that it takes a similar form to the above list of steps but executed in reverse [2, 4]. It begins the substantial calcula-
tions at the gradient of Step 4 above which is just before the termination of the forward evaluation. The steps for the gradient evaluation of the raytrace are as follows:

1. Set the current surface index \( j \) to \( r \).

2. **Accumulating Optical Path Distance:** Using \( \vec{\sigma} \), calculate \( \vec{L}_j \) and accumulate a contribution to \( \vec{n}_j \) as described in Section 4.2.4.

3. **Ray Transfer:** Given \( \vec{x}_j, \vec{L}_j, \) and \( \vec{E}_{1,j} \), calculate \( \vec{x}'_{j-1} \) and accumulate contributions to \( \vec{X}'_{j-1} \) and \( \vec{c}_j \). Details of this step are provided in Section 4.2.2 for spherical surfaces. For aspherical or freeform surfaces, the transfer gradient also requires \( \vec{N}_j \) as described in Section 4.2.3. The arrays \( \vec{x}_j, \vec{L}_j, \vec{E}_{1,j}, \) and \( \vec{N}_j \) may be removed from memory.

4. **Coordinate Transformation:** Accumulate contributions to \( \vec{x}_{j-1} \) and \( \vec{X}_{j-1} \) given \( \vec{x}'_{j-1} \) and \( \vec{X}'_{j-1} \). Accumulate contributions to \( \vec{E}_{j}, \vec{Y}_j, \vec{I}_j, \alpha'_j, \beta'_j, \) and \( \gamma'_j \), if necessary, as described in Section 4.2.1. The arrays \( \vec{x}'_{j-1} \) and \( \vec{X}'_{j-1} \) may be removed from memory.

5. Subtract 1 from \( j \). If \( j \) has become 0, stop, because \( x_0 \) and \( X_0 \) do not depend on varying parameters that require gradient contributions.

6. **Refraction:** Given \( \vec{X}_j \), found in gradient Step 4, calculate \( \vec{X}'_{j-1} \), and accumulate contributions to \( \vec{x}_j, \vec{E}_{1,j}, \vec{c}_j, \vec{n}_j, \vec{n}_{j-1} \) and, for aspherical or freeform surfaces, \( \vec{N}_j \). This is described in Section 4.2.5 and, at the conclusion, the array \( \vec{X}_j \) may be removed from memory.

7. Return to Step 2 and continue.
The complexity of the RMAD gradient for a raytrace is apparent from the above steps. Adjoint variables for various surfaces are regularly being accumulated to and, if needed to save memory, being eliminated. Also, the surface index $j$ changes during the existence of many of the variables so care must be taken to understand which values are used where. For example, the $\overline{X'}_{j-1}$ used in Step 4 was actually accumulated to in an earlier iteration of Step 6 and, later, added to in Step 3.

### 4.2.1 Coordinate Transformation

In the coordinate transformation step, Step 2, ray intersection coordinates $x_{j-1}$ and $X_{j-1}$ in the local coordinate system of the $(j-1)$th surface are translated into coordinates $x'_{j-1}$ and $X'_{j-1}$ relative to the $j$th surface.

The vertex of the $j$th surface will be displaced by the vector $\Lambda_j = [\Xi_j \Upsilon_j \ t_j]^T$ from the last surface’s vertex. The ray coordinates change in the opposite direction as they become represented in the coordinate system of the new vertex location, so let

$$x''_{j-1}[\rho, \kappa] = x_{j-1}[\rho, \kappa] - \Lambda_j[\rho], \quad (4.10)$$

be the coordinates of the ray intersections with the $(j-1)$th surface in the local coordinate system of the $j$th surface. We ignore the distinction that would be made if “z-decentration” variable was introduced as in [59] because, from a parameter estimation standpoint, it would be indistinguishable from the surface thickness $t_j$.

Next, the surface is rotated about the $+x$ axis by $-\alpha'_{j}$ radians, then about the $+y$ axis by $-\beta'_{j}$ radians and then about the $+z$ axis by $\gamma'_{j}$ radians. Relative to the
vertex position defined by \( \Lambda_j \), this yields two left-handed rotations followed by a right-handed rotation consistent with the CodeV\textsuperscript{TM} software. The \( x''_{j-1} \) coordinates can be expressed in the coordinates of the rotated surface by rotating in the opposite directions as the surface. Define

\[
A_j = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos \alpha_j' & \sin \alpha_j' \\ 0 & -\sin \alpha_j' & \cos \alpha_j' \end{bmatrix}, \quad B_j = \begin{bmatrix} \cos \beta_j' & 0 & \sin \beta_j' \\ 0 & 1 & 0 \\ -\sin \beta_j' & 0 & \cos \beta_j' \end{bmatrix}, \quad C_j = \begin{bmatrix} \cos \gamma_j' & -\sin \gamma_j' & 0 \\ \sin \gamma_j' & \cos \gamma_j' & 0 \\ 0 & 0 & 1 \end{bmatrix}.
\]

(4.11)

The composition of these rotations is

\[
R_j = A_j \ast B_j \ast C_j,
\]

(4.12)

where the \( \ast \) indicates matrix-multiplication. This allows to calculate the ray intersection coordinates and directional cosines using simple matrix multiplication applied to the coordinate and direction arrays,

\[
\begin{align*}
x_j' & = R_j \ast x''_{j-1}, \\
X_j' & = R_j \ast X_{j-1}.
\end{align*}
\]

(4.13)

The gradient calculation receives \( \overrightarrow{x'}_{j-1} \) and \( \overrightarrow{X'}_{j-1} \). By Eq. (50) in [4], the gradient steps for intercept coordinates and directional cosines due to Eq. (4.13) is

\[
\begin{align*}
\overrightarrow{x''}_{j-1} & = R_j^T \ast \overrightarrow{x'}_{j-1}, \\
\overrightarrow{X}_{j-1} & = R_j^T \ast \overrightarrow{X'}_{j-1}.
\end{align*}
\]

(4.14)
where \(+=\) is an accumulate operator like that defined in [4]. In this chapter, the accumulation operator adds the value of the right operand to the value of the left operand and stores it in place of the left, if the left has been defined. If the left operand has no previous definition or value, the accumulation operator defines the value on the right to be copied and stored with the name of the left operand. We use the accumulator operator for all gradient operations in Section 4.2 to account for error metrics that depend on arbitrary outputs of the raytrace. Since the gradient of a general error metric could have gradient contributions due to any result of the raytrace, the left operand in a raytrace gradient equation may or may not already be defined prior to evaluation of the raytrace gradient. This process of accumulation is fundamental to the evaluation of RMAD gradients [3, 4].

By Eq. (47) in [4] applied to Eq. (4.10),

\[
\begin{align*}
\overline{x''_{j-1}} + &= R_j^T \overline{x'_{j-1}}, \\
\overline{x_{j-1}} + &= \overline{x''_{j-1}},
\end{align*}
\]  

(4.15)

If none of the decentrations, \(t_j\), or the surface rotations are varying, evaluation of the gradient steps for coordinate transformation for the \(j\)th surface ceases as the required gradients have been found. If decentrations or \(t_j\) are varying, Eq. (61) in [4] applied to Eq. (4.10) yields

\[
\overline{\Lambda}_j[\rho] + = - \sum_{\kappa=0}^{K-1} \overline{x''_{j-1}[\rho, \kappa]},
\]  

(4.16)

and values to be accumulated to \(\overline{\Xi}_j, \overline{\Upsilon}_j\) and/or \(\overline{t}_j\) can be unpacked from \(\overline{\Lambda}_j\).
If surface rotations are varying, one must return to the gradient steps for Eq. (4.13) and apply Eq. (50) in [4] to yield two contributions to the term $\bar{R}_j$,

$$\bar{R}_j = \chi'_{j-1} \ast \chi''_{j-1} + \chi'_{j-1} \ast X'_{j-1}. \tag{4.17}$$

Repeatedly applying the Eq. (50) matrix differentiation rule to Eq. (4.12) yields

$$\bar{A}_j = \bar{R}_j \ast C_j^T \ast B_j^T,$$
$$\bar{B}_j = A_j^T \ast \bar{R}_j \ast C_j^T,$$
$$\bar{C}_j = B_j^T \ast A_j^T \ast \bar{R}_j. \tag{4.18}$$

Gradients for a matrices like $A_j$ were evaluated in Section 3.7 using Eqs. (3.38) and (3.39). By a similar evaluation, the gradient steps for Eq. (4.11) are

$$\bar{\alpha}_j = -\sin \alpha'_j \ast (\bar{A}_j[1, 1] + \bar{A}_j[2, 2]) + \cos \alpha'_j (\bar{A}_j[1, 2] - \bar{A}_j[2, 1]),$$
$$\bar{\beta}_j = -\sin \beta'_j \ast (\bar{B}_j[0, 0] + \bar{B}_j[2, 2]) + \cos \beta'_j (\bar{B}_j[0, 2] - \bar{B}_j[2, 0]),$$
$$\bar{\gamma}_j = -\sin \gamma'_j \ast (\bar{C}_j[0, 0] + \bar{C}_j[1, 1]) + \cos \gamma'_j (\bar{C}_j[1, 0] - \bar{C}_j[0, 1]). \tag{4.19}$$

### 4.2.2 Ray Transfer for a Spherical Surface

In this section, we describe the calculation and gradient of the ray-intersection between a ray starting at $\chi'_{j-1}$ and having directional cosine $X'_{j-1}$ with a spherical surface having curvature $c_j$. 
In the notation of [53], the first step in the ray transfer process is Eq. (A.2c) in [53] with

\[ e = t_j Z - (xX + yY + zZ). \]  

(4.20)

As noted above, for notational clarity and computational efficiency, we use arrays containing all dimensions of the coordinates for all of the ray intersections in a single array. Our equivalent to Eq. (A.2c) in [53] is

\[ e_j[\kappa] = -\sum_{\rho=0}^{2} x'_{j-1}[\rho, \kappa] X'_{j-1}[\rho, \kappa], \] 

(4.21)

since \( t_j \) was already subtracted in Eq. (4.10). By similar translation, Eq. (A2.d) in [53] is

\[ m_j[\kappa] = x'_{j-1}[2, \kappa] + e_j[\kappa] X'_{j-1}[2, \kappa], \] 

(4.22)

where \( m_j[\kappa] \) is more commonly called \( M_{1z} \) but for clarity we neglect the additional subscripts. By Eq. (A.2e) in [53],

\[ M_j[\kappa] = \sum_{\rho=0}^{2} x'_{j-1}^{2}[\rho, \kappa] - e_j^{2}[\kappa] \] 

(4.23)

which is more commonly referred to as \( M_1^2 \) in [53]. Next, we depart from the normal formula and define the additional auxiliary quantities

\[ t_j[\kappa] = c_j M_j[\kappa] - 2m_j[\kappa], \] 

(4.24)

\[ \nu_j[\kappa] = X'_{j-1}^{2}[2, \kappa] - c_j t_j[\kappa], \] 

(4.25)
such that the cosine of the angle of incidence with the surface for each ray is

\[ E_{1,j}[\kappa] = \sqrt{\nu_j[\kappa]}, \quad (4.26) \]

by Eq. (A.2f) in [53]. We defined \( t_j[\kappa] \) because its content appears twice in the ray transfer and therefore its equivalent in the gradient will be accumulated. Similarly, we defined \( \nu_j[\kappa] \) to simplify the gradient calculation due to Eq. (4.26)'s nonlinearity. Define another auxiliary quantity,

\[ \zeta_j[\kappa] = X'_{j-1}[2, \kappa] + \nu_j[\kappa] \quad (4.27) \]

so that the physical path length is calculated by Eq. (A.2g) is

\[ L_j[\kappa] = e_j[\kappa] + t_j[\kappa] \zeta_j^{-1}[\kappa]. \quad (4.28) \]

Finally,

\[ x_j[\rho, \kappa] = x'_{j-1}[\rho, \kappa] + L_j[\kappa] X'_{j-1}[\rho, \kappa], \quad (4.29) \]

by Eq. (A.2h) in [53].

Next, we find the steps of the gradient for ray-transfer whose inputs are \( \overline{x_j}, \overline{L_j} \) and \( \overline{E_{1,j}} \). Applying Eq. (47) from [4] to find the first gradient steps associated with Eq. (4.29):  

\[ \overline{x'_{j-1}} += \overline{x_j}. \quad (4.30) \]
The contribution to $L_j$ due to $L_j$ in Eq. (4.29) is

$$L_j[\kappa] = \sum_{\rho=0}^{2} X_{j-1}'[\rho, \kappa] \bar{x}_j[\rho, \kappa]. \quad (4.31)$$

To find the right hand side of Eq. (4.31), one must apply Eqs. (47) and (51) of [4] to Eq. (4.29) for each ray. Since $X_{j-1}'$ was previously defined by the refraction step, accumulate

$$X_{j-1}'[\rho, \kappa] = L_j[\kappa] \bar{x}_j[\rho, \kappa]. \quad (4.32)$$

Evaluating the gradient for Eq. (4.28) yields two results using previously introduced gradient types and a more difficult result involving the power rule Eq. (52) in [4].

$$\bar{e}_j = L_j,$$

$$\bar{t}_j[\kappa] = L_j[\kappa] \zeta_j^{-1}[\kappa],$$

$$\bar{\zeta}_j[\kappa] = - \zeta_j^{-2}[\kappa] t_j[\kappa] L_j[\kappa]. \quad (4.33)$$

By Eq. (4.27) then,

$$X_{j-1}'[2, \kappa] = \bar{\zeta}_j[\kappa],$$

$$E_{1,j}[\kappa] = \bar{\zeta}_j[\kappa]. \quad (4.34)$$

Applying the power rule to Eq. (4.26) yields

$$\nu_j[\kappa] = \frac{1}{2 \sqrt{\nu_j[\kappa]}} E_{1,j}[\kappa]. \quad (4.35)$$
and it is also required for the gradient steps of Eq. (4.25),

\[ X_{j-1}^{\prime}[2, \kappa] + = 2X_{j-1}^{\prime}[2, \kappa] \overline{v}_j[\kappa], \]

\[ \overline{c}_j + = - \sum_{\kappa=0}^{K-1} t_j[\kappa] \overline{v}_j[\kappa], \]

\[ \overline{t}_j[\kappa] + = - c_j \overline{v}_j[\kappa]. \]  \hspace{1cm} (4.36)

The gradient steps for Eq. (4.24),

\[ \overline{c}_j + = \sum_{\kappa=0}^{K-1} M_j[\kappa] \overline{t}_j[\kappa], \]

\[ \overline{M}_j[\kappa] + = c_j \overline{t}_j[\kappa], \]

\[ \overline{m}_j[\kappa] + = - 2\overline{t}_j[\kappa]. \]  \hspace{1cm} (4.37)

and those for Eq. (4.23) also require the power rule with,

\[ \overline{x}_{j-1}^{\prime}[\rho, \kappa] + = 2x_{j-1}^{\prime}[\rho, \kappa] \overline{M}_j[\kappa], \]

\[ \overline{e}_j[\kappa] + = - 2e_j[\kappa] \overline{M}_j[\kappa]. \]  \hspace{1cm} (4.38)

From Eq. (4.22), we have,

\[ \overline{x}_{j-1}^{\prime}[2, \kappa] + = \overline{m}_j[\kappa], \]

\[ \overline{e}_j[\kappa] + = X_{j-1}^{\prime}[2, \kappa] \overline{m}_j[\kappa], \]

\[ \overline{X}_{j-1}^{\prime}[2, \kappa] + = e_j[\kappa] \overline{m}_j[\kappa]. \]  \hspace{1cm} (4.39)
Finally, for the gradient of Eq. (4.21),

\[
\begin{align*}
\vec{x}'_{j-1}[\rho, \kappa] &= -X'_{j-1}[\rho, \kappa]\bar{e}_j[\kappa], \\
\bar{X}'_{j-1}[\rho, \kappa] &= -x'_{j-1}[\rho, \kappa]\bar{e}_j[\kappa].
\end{align*}
\]

(4.40)

4.2.3 Ray Transfer for an Aspherical or Freeform Surface

The shape of the optical surface is defined by the points \((x, y, z)\) satisfying the equation

\[0 = z - s(x, y), \]

(4.41)

where \(s(x, y)\) is known as the sag of the surface at \((x, y)\). Ray transfer involves finding a point \((x, y, z)\) on each ray satisfying Eq. (4.41) and then finding other parameters such as the surface normal vector and the cosine of the angle of incidence on the surface.

For simplicity and generality, we treat ray transfer as two distinct calculations. The first calculation, outlined in Section 4.2.3.1, evaluates \(s(x, y)\) and first derivatives \(\frac{\partial s}{\partial x}\) and \(\frac{\partial s}{\partial y}\) of the surface for a given \((x, y)\) from each ray. We will refer to \(\frac{\partial s}{\partial x}\) and \(\frac{\partial s}{\partial y}\) as the surface slope to avoid confusion with derivatives associated with the RMAD gradient. The sag and slope of the surface depend on the prescription parameters \(c_j, k_j,\) and \(a_j\). The second calculation of ray transfer, ray intersection, is described in Section 4.2.3.2 and uses the Newton-Raphson method to find the point on each ray that satisfies Eq. (4.41). Solution of this constraint requires multiple executions of the sag function in an iterative root-finding process. In contrast, such
iterations are not necessary for the gradient since the implicit function theorem (IFT) [40] can be employed to simplify the gradient steps.

The distinction between the sag/slope calculation and the ray intersection is useful for two reasons. First, parameterizations of freeform surfaces other than the model employed in Section 4.2.3.1 are easily accommodated by replacing the functions with the computations for a different surface parameterization. Second, while intermediate RMAD steps in the sag calculation are easily debugged by comparison with the method of finite differences, it is more difficult to isolate errors in the intermediate RMAD steps of the ray intersection. Thus, it is desirable to isolate the ray intersection gradient from the sag and slope gradient to isolate programming errors.

4.2.3.1 Sag and Slope of a Freeform Surface

In this section, we describe the sag computation for a freeform surface whose sag

\[
s(x, y) = \frac{c_j(x^2 + y^2)}{1 + \sqrt{1 - (1 + k_j)c_j^2(x^2 + y^2)}} + \sum_{p=0}^{P} \sum_{q=0}^{P} H_j[p, q]x^p y^q \tag{4.42}
\]

is the sum of the sag of a conic section asphere and the sag of a polynomial surface whose coefficients are a power series in Cartesian \((x, y)\) coordinates. The slope of the surface is

\[
\frac{\partial s}{\partial x}(x, y) = \frac{c_jx}{\sqrt{1 - (1 + k_j)c_j^2(x^2 + y^2)}} + \sum_{p=0}^{P-1} \sum_{q=0}^{P} (p + 1)H_j[p + 1, q]x^p y^q,
\]

\[
\frac{\partial s}{\partial y}(x, y) = \frac{cjy}{\sqrt{1 - (1 + k_j)c_j^2(x^2 + y^2)}} + \sum_{p=0}^{P} \sum_{q=0}^{P-1} (q + 1)H_j[p, q + 1]x^p y^q, \tag{4.43}
\]
and, like the sag, the slope has a term from the conic section and a term from the polynomial surface.

Direct optimization of values in $H_j[p, q]$ would likely be poorly regularized since the polynomial basis is neither orthogonal nor usefully normalized over the extent of the surface. Instead, we optimize over $O$ coefficients $a_j$ of an alternate basis set and compute

$$H_j[p, q] = \sum_{o=0}^{O-1} a_j[o] b_j[o, p, q], \quad (4.44)$$

where $b[o, p, q]$ specifies the coefficient of $x^p y^q$ for $o$th basis vector. If the basis is the Zernike polynomials [60], the coefficients $b[o, p, q]$ can be evaluated using the algorithm in [61] or extracted from Table 2.1. In this case, $P$ is the highest radial-degree [60] present on the Zernike basis. Also, the values in $H_j[p, q]$ and $b_j[o, p, q]$ with $p+q \geq P$ will be zero. The Forbes polynomial freeform parameterization can also be expressed in a Cartesian sum like Eq. (4.42) [62], so there exists some $b_j[o, p, q]$ for which $a_j[o]$ are Forbes freeform polynomial coefficients. Normalization constants necessary to convert coordinates $(x, y)$ to normalized coordinates on the unit circle can be applied by scaling values of $b_j[o, p, q]$ appropriately.

The prescription parameters $c_j, k_j$ and $a_j$ and an for estimate $x_j[\rho, \kappa]$ constitute the input to the sag and slope calculations. The sag calculation begins by finding $x^2 + y^2$ for each ray, so define

$$h_j[\kappa] = \sum_{\rho=0}^{1} x_j^2[\rho, \kappa], \quad (4.45)$$
making the quantity inside the root of Eq. (4.42)

\[ V_j[\kappa] = 1 - (1 + k_j)c_j^2 h_j[\kappa]. \]  

(4.46)

Rays for which \( V_j[\kappa] < 0 \) are in regions where the aspheric portion of the sag is undefined. These rays should be treated as missing the surface unless an alternate convention for the surface shape than Eq. (4.42) is adopted. For the rays which intersect the surface at a point where the surface is defined, let

\[ v_j[\kappa] = \sqrt{V_j[\kappa]}, \]  

(4.47)

such that the sag for each ray is

\[ s_j[\kappa] = \frac{c_j h_j[\kappa]}{1 + v_j[\kappa]} + \sum_{p=0}^{P} \sum_{q=0}^{P} H_j[p, q]x_j^p[0, \kappa]x_j^q[1, \kappa]. \]  

(4.48)

This section requires notating the first and second partial derivatives of the polynomial expansion of Eq. (4.48). For clarity and efficient computation, define the 5 by \( P \) by \( P \) tensor

\[ G_j[\rho, p, q] = \begin{cases} 
(p + 1)H_j[p + 1, q] & \text{for } \rho = 0 \text{ and } p < P \\
(q + 1)H_j[p, q + 1] & \text{for } \rho = 1 \text{ and } q < P \\
(p + 1)(p + 2)H_j[p + 2, q] & \text{for } \rho = 2 \text{ and } p < P - 1 \\
(q + 1)(q + 2)H_j[p, q + 2] & \text{for } \rho = 3 \text{ and } q < P - 1 \\
(p + 1)(q + 1)H_j[p + 1, q + 1] & \text{for } \rho = 4, p < P \text{ and } q < P \\
0 & \text{otherwise.} 
\end{cases} \]  

(4.49)
For \( \rho = 0 \) and \( \rho = 1 \), the values of \( G_j[\rho, p, q] \) correspond to the polynomial coefficients of the polynomials corresponding to partial derivatives \( \partial / \partial x \) and \( \partial / \partial y \), respectively, of the polynomial of Eq. (4.48). The remaining cases, \( \rho = 2 \), \( \rho = 3 \) and \( \rho = 4 \), correspond to the second-order partial derivatives \( \partial^2 / \partial^2 x \), \( \partial^2 / \partial^2 y \) and \( \partial^2 / \partial x \partial y \), respectively. Using this tensor, an array compactly expressing the slope in Eq. (4.43) is

\[
U_j[\rho, \kappa] = \frac{c_jx_j[\rho, \kappa]}{v_j[\kappa]} + \sum_{p=0}^{P} \sum_{q=0}^{P} G_j[\rho, p, q]x_j^p[0, \kappa]x_j^q[1, \kappa] \quad \text{for } \rho = 0 \text{ or } 1. \quad (4.50)
\]

Next, we consider the RMAD gradient steps for the above. Let \( \bar{x}_j^{(U)} \) be the contribution to the gradient of the error metric due to the dependence of \( U_j \) on \( x_j \) as in the notation of Section 3.7. The ray intersection gradient below calculates \( \bar{s}_j \) from \( \bar{x}_j^{(U)} \), but \( \bar{x}_j^{(U)} \) must be calculated by the sag and slope model from \( \bar{u}_j \). Consequently, there must be two distinct RMAD gradient calculations for the above sag and slope computation. The first, the slope gradient, calculates \( \bar{x}_j^{(U)} \) from \( \bar{U}_j \) and also accumulates contributions associated with \( U_j \)'s dependence on \( c_j \), \( k_j \) and \( a_j \). The second RMAD gradient calculation, the sag gradient, finds the contribution to the error metric gradient due to \( s_j \)'s depending on \( x_j \), \( c_j \), \( k_j \), and \( a_j \).
The slope gradient begins with the RMAD steps associated with conic term in Eq. (4.50),

$$
\overline{c}_j = \sum_{\kappa=0}^{K-1} \sum_{\rho=0}^{1} \frac{U_j[\rho, \kappa]}{v_j[\kappa]}
$$

$$
\overline{v}_j^{(u_j)}[\kappa] = - c_j \sum_{\rho=0}^{1} \frac{x_j[\rho, \kappa]U_j[\rho, \kappa]}{v^2[\kappa]}
$$

$$
\overline{x}_j^{(u_j)}[\rho, \kappa] = \frac{c_j U_j[\rho, \kappa]}{v_j[\kappa]} \text{ for } \rho = 0 \text{ and } 1,
$$

(4.51)

where the gradient contribution due to $U_j$’s dependence on $v_j$ is uniquely notated to distinguish it from $s_j$’s dependence on $v_j$ which appears in the sag gradient. From Eq. (4.47)

$$
\overline{V}_j^{(u_j)}[\kappa] = \frac{\overline{v}_j^{(u_j)}[\kappa]}{2v[\kappa]},
$$

(4.52)

and by Eq. (4.46),

$$
\overline{k}_j = - c_j^2 \sum_{\kappa=0}^{K-1} h_j[\kappa] \overline{V}_j^{(u_j)}[\kappa]
$$

$$
\overline{c}_j = - 2c_j(1 + k_j) \sum_{\kappa=0}^{K-1} h_j[\kappa] \overline{V}_j^{(u_j)}[\kappa]
$$

$$
\overline{h}_j^{(u_j)}[\kappa] = - (1 + k_j)c_j^2 \overline{V}_j^{(u_j)}[\kappa].
$$

(4.53)

The step for Eq. (4.45) is

$$
\overline{x}_j^{(u_j)}[\rho, \kappa] = 2x_j[\rho, \kappa] \overline{h}_j^{(u_j)}[\kappa] \text{ for } \rho = 0 \text{ or } 1.
$$

(4.54)
Returning to the RMAD steps associated with the polynomial term in Eq. (4.50),

\[
G_j[\rho, p, q] = \sum_{\kappa=0}^{K-1} U_j[\rho, \kappa] x_j^p[0, \kappa] x_j^q[1, \kappa] \text{ for } \rho = 0 \text{ or } 1,
\]

(4.55)

and, for simplicity, we defer further accumulations of \( G_j[\rho, p, q] \) until the sag gradient.

Since the polynomial coefficients for \( \partial^2/\partial^2 x \) and \( \partial^2/\partial^2 y \) are already in \( G_j \),

\[
\bar{x}_j^{(U)}[\rho, \kappa] = \sum_{p=0}^{P} \sum_{q=0}^{P} G_j[\rho + 2, p, q] x_j^p[0, \kappa] x_j^q[1, \kappa] \text{ for } \rho = 0 \text{ or } 1.
\]

(4.56)

Accumulating the dependence of the \( x \)-slope’s dependence on the \( y \) coordinate and \( y \)-slope’s dependence on the \( x \) coordinate requires the \( \partial^2/\partial x \partial y \) coefficients:

\[
\bar{x}_j^{(U)}[0, \kappa] = \sum_{p=0}^{P} \sum_{q=0}^{P} G_j[4, p, q] x_j^p[0, \kappa] x_j^q[1, \kappa],
\]

\[
\bar{x}_j^{(U)}[1, \kappa] = \sum_{p=0}^{P} \sum_{q=0}^{P} G_j[4, p, q] x_j^p[0, \kappa] x_j^q[1, \kappa].
\]

(4.57)

Incidentally, if \( P \) is large, it may be computationally expedient to simultaneously compute the polynomial sums in Eqs. (4.48), (4.50), (4.56) and (4.57) during the forward execution of the sag and slope model. \( G_j[\rho, p, q] \) is structured so that all polynomial sums can be found in one tensor multiplication operation from an optimized library such as [63].

There is no need to employ RMAD on Eq. (4.48) for the purpose of finding \( \bar{x}_j^{(s)}[\rho, \kappa] \). To see this, consider that the first row of \( \bar{x}_j^{(s)} \) is given by

\[
\frac{\partial E}{\partial \bar{x}} = \frac{\partial s}{\partial \bar{x}} \frac{\partial E}{\partial s'}
\]

(4.58)
and the second row is given by

$$\frac{\partial E}{\partial y} = \frac{\partial s}{\partial y} \frac{\partial E}{\partial s'}, \quad (4.59)$$

and $\partial s/\partial x$ and $\partial s/\partial y$ are the first and second rows, respectively, of $U_j$ due to its definition in Eq. (4.50). Thus, by Eqs. (4.58)-(4.59),

$$\bar{x}_j^{(s)}[\rho, \kappa] = U_j[\rho, \kappa] \bar{s}_j[\kappa]. \quad (4.60)$$

Similarly, evaluation of $\bar{h}_j^{(s)}$ can be skipped since it depends only on $x_j$ and those contributions were accumulated in Eq. (4.60). The remaining RMAD steps associated with conic term in Eq. (4.48) are

$$\bar{c}_j + = \sum_{\kappa=0}^{K-1} h_j[\kappa] \bar{s}_j[\kappa] = c_j^{-1} \sum_{\kappa=0}^{K-1} s_j[\kappa] \bar{s}_j[\kappa],$$

$$\bar{v}_j^{(s)}[\kappa] + = - c_j h_j[\kappa] \bar{s}_j[\kappa] \frac{1}{(1 + v_j[\kappa])^2} = \frac{s_j[\kappa] \bar{s}_j[\kappa]}{1 + v_j[\kappa]} \quad (4.61)$$

We return to evaluation of the second term in Eq. (4.48) below, but now we continue evaluating the remaining steps due to the first term. Analogous to Eqs. (4.52) and (4.53),

$$\bar{V}_j^{(s)}[\kappa] = \frac{\bar{v}_j^{(s)}[\kappa]}{2v[\kappa]},$$

$$\bar{k}_j + = - c_j^2 \sum_{\kappa=0}^{K-1} h_j[\kappa] \bar{V}_j^{(s)}[\kappa],$$

$$\bar{c}_j + = - 2c_j(1 + k_j) \sum_{\kappa=0}^{K-1} h_j[\kappa] \bar{V}_j^{(s)}[\kappa]. \quad (4.62)$$
The RMAD steps associated with the polynomial term in Eq. (4.48) begin with

$$H_j[p, q] = \sum_{\kappa=0}^{K-1} s_j[\kappa] x_j^p[0, \kappa] x_j^q[1, \kappa].$$  \hspace{1cm} (4.63)

Accumulating the gradient contributions \(G_j[p, p, q]\) left over from the slope gradient yields

$$H_j[p, q] += (p - 1)G_j[0, p - 1, q] \text{ for } p \geq 1,$$

$$H_j[p, q] += (q - 1)G_j[1, p, q - 1] \text{ for } q \geq 1,$$  \hspace{1cm} (4.64)

by the gradient steps associated with Eq. (4.49). Finally, the gradient step associated with Eq. (4.44) is

$$\vec{a}_j[0] += \sum_{p=0}^{P} \sum_{q=0}^{P} b_j[0, p, q]H_j[p, q].$$  \hspace{1cm} (4.65)

### 4.2.3.2 Iterative Ray Intersection with a Freeform Surface

The ray intersection calculation finds \(x_j\) where rays cross the surface model defined in Section 4.2.3.1. As with the spherical intersection, \(L_j\) and \(E_{1,j}\) must be evaluated. The unit normal of the surface, \(N_j\), must also be computed.

Algorithm 4.1 lists one way of finding \(x_j\) and \(L_j\) using Newton-Raphson iteration as in [51, 53]. Lines 1 and 2 intersect the ray with the plane defined by \(z = 0\) to find an initial estimate for \(x_j\). Line 5 evaluates the inner product between the normal vector of the surface, \((-\partial s/\partial x, -\partial s/\partial y, 1)\), and the directional cosine of each ray. Line 6 finds the step distance that, to first order, should eliminate the error between the sag and \(z\) coordinate for each intersection point. The potential intersection points
Algorithm 4.1 Iterative estimation of ray intersection: Finds points $x_j$ on the surface and on each ray as well as the distance $L_j$ along the rays from last intersection points $x'_{j-1}$.

1: $L_j[κ] ← -x'_{j-1}[2, κ]/X'_{j-1}[2, κ]$ \( \triangleright \) Distance from $x'_{j-1}$ to $z = 0$ plane along ray.
2: $x_j[ρ, κ] ← x'_{j-1}[ρ, κ] + L_j[κ]X'_{j-1}[ρ, κ]$ \( \triangleright \) Intersection point with $z = 0$ plane.
3: for $p = 0$ to $5$ do
4: Evaluate $s_j$ and $U_j$ using Eq. (4.48) and Eq. (4.50), respectively.
5: $E_{0,j}[κ] ← X'_{j-1}[2, κ] - \sum_{ρ=0}^{1} U_j[ρ, κ]X_j[ρ, κ]$
6: $G_{0,j}[κ] ← \{ s_j[κ] - x_j[2, κ] \} / E_{0,j}[κ]$ \( \triangleright \) Analogous to Eq. (A.3g) in [53]
7: $L_j[κ] += G_{0,j}[κ]$
8: $x_j[ρ, κ] += G_{0,j}[κ]X'_{j-1}[ρ, κ]$
9: end for
10: return $x_j$, $L_j$ and $E_{0,j}$

are constrained to lie on each ray by lines 2 and 8. If, at the conclusion of the loop, some values of $G_0$ are not vanishingly small, these rays either missed the surface or failed to converge to a valid intersection. If the latter, an alternate ray intersection estimation method may be needed. Due to the use of the IFT, it is of no consequence to the gradient computation how the intersection point of the ray with the surface is actually accomplished. If Algorithm 4.1 does not converge satisfactorily, one is free to use an alternate technique that finds $x_j$, $L_j$ and $E_{0,j}$.

The scale factor that makes the normal vectors unit length is

$$N_{1,j}[κ] = \left( 1 + \sum_{ρ=0}^{1} U_j^2[ρ, κ] \right)^{-\frac{1}{2}}, \quad (4.66)$$
and the unit-normal vectors are

\[
N_{0,j}[\rho, \kappa] = \begin{cases} 
-N_{1,j}[\kappa]U_j[\rho, \kappa] & \text{for } \rho = 0 \text{ or } 1 \\
N_{1,j}[\kappa] & \text{for } \rho = 2.
\end{cases}
\]  

(4.67)

The cosine of the angle of incidence with the surface can be evaluated by taking the inner product (over the \(\rho\) dimension) of \(N_{0,j}\) and \(X'_{j-1}\). This was already done in Line 5 of Algorithm 4.1 but without unit normalization. Consequently it is expedient to reuse that result and calculate:

\[
E_{1,j}[\kappa] = N_{1,j}[\kappa]E_{0,j}[\kappa].
\]  

(4.68)

This concludes the ray intersection computation.

Next, we find the gradient of the above ray intersection computation. Although it is possible to explicitly differentiate every iterate of the loop in Algorithm 4.1 using RMAD, such a gradient would require storing six copies of intermediate values of \(E_{0,j}, G_{0,j}\) and \(U_j\) for each of the six iterations of the loop. Further, the sag and slope gradients would need to be executed six times each. However, since the result of this loop satisfies the constraint Eq. (4.41) for each ray, the IFT can be employed as shown below to avoid the effort of differentiating each loop iterate.
First, if a ray with directional cosine \((X', Y', Z')\) intersected the previous surface at \((x', y', z')\), and \(L\) is an estimate for the path length to the intersection,

\[
x = x' + LX',
\]

\[
y = y' + LY',
\]

\[
z = z' + LZ',
\]

is an estimate the intersection point. Substituting this estimate into Eq. (4.41) yields

\[
\Psi(L; x', y', z', X', Y', Z') = z - s(x, y)
\]

\[
= z' + LZ' - s(x' + LX', y' + LY'),
\]

the departure of the intersection estimate from the surface along the \(z\) axis. In vector form, Eq. (4.70) is

\[
\Psi_j(L_j; x'_{j-1}, X'_{j-1})[\kappa] = \{x'_{j-1}[2, \kappa] + L_j[\kappa]X'_{j-1}[2, \kappa]\} -
\]

\[
S\{x'_{j-1}[\rho, \kappa] + L_j[\kappa]X'_{j-1}[\rho, \kappa]\} \text{ for } \rho = 0 \text{ or } 1
\]

where \(S\) is the sag function that generates \(s_j\) given the first two rows of an estimate for \(x_j\). The result of the function \(\Psi_j\), which we notate \(\psi_j\), is zero for successfully intersected rays. Next, assume that \(\Psi_j\) is Lipschitz [40] and continuously differentiable in all variables. Also, let \(\left(d\Psi_j[\kappa]/dL_j[\kappa]\right)^{-1}\) exist and be continuous such that \(\Psi_j\) is a well-behaved implicit mapping [40]. These assumptions are likely valid for most polynomial surfaces modeled by Eq. (4.42) under the restriction each ray intersects the surface only once. If rays can intersect the surface multiple times or if the surface
is discontinuous like a segmented mirror, this assumption may need to be revisited. By the implicit function theorem, \( \Psi_j = 0 \) defines an implicit function \( L_j(x'_{j-1}, X'_{j-1}) \) such that

\[
\Psi_j(L_j(x'_{j-1}, X'_{j-1}); x'_{j-1}, X'_{j-1}) = 0. \tag{4.72}
\]

Further, although a simple formula for the computation \( L_j \) is unavailable, its RMAD gradient steps can be found in Algorithm 3.1 of [40]. Assuming \( L_j \) is the result of the function \( L_j \), find

\[
\overline{\psi}_j[\kappa] = -\frac{\overline{L}_j[\kappa]}{\overline{\Delta L}_j[\kappa]} = -\frac{\overline{L}_j[\kappa]}{X'_{j-1}[2, \kappa] - \sum_{\rho=0}^{1} U_j[\rho, \kappa] X_j[\rho, \kappa]}
= -\frac{\overline{L}_j[\kappa]}{E_{0,j}[\kappa]} \tag{4.73}
\]

and then compute the RMAD steps of Eq. (4.71) using \( \overline{\psi}_j \) as if \( \psi_j \) was the result of an execution of \( \Psi_j \). Note that the operator in Eq. (4.73) is an equality because \( \psi_j \) is not one of the outputs of the raytrace calculation and its adjoint should not exist prior to Eq. (4.73).

Utilizing the RMAD steps for \( L_j \) to find the gradient steps for the ray intersection is different than applying RMAD to the actual computation. To utilize the IFT, one differentiates a notional version of the ray intersection process where (i) \( L_j \) is computed from \( x'_{j-1} \) and \( X'_{j-1} \) using \( L_j \), and (ii) all other outputs are calculated from \( L_j, x'_{j-1}, \) and \( X'_{j-1} \). The steps for this notional ray transfer calculation are listed in Algorithm 4.2 and it is this computation which we differentiate using RMAD rather than the true ray intersection outlined earlier in the section.
Algorithm 4.2 Notional freeform ray transfer computation to be differentiated using RMAD. Assumes existence of \( \mathcal{L}_j \) and finds outputs equivalent to the actual ray transfer computation.

\[
\begin{align*}
1: & \quad \mathcal{L}_j \leftarrow \mathcal{L}_j(x'_{j-1}, X'_{j-1}) \\
2: & \quad x_j[\rho, \kappa] \leftarrow x'_{j-1}[\rho, \kappa] + L_j[\kappa]X'_{j-1}[\rho, \kappa] \quad \triangleright \text{Same as Eq. (4.29).} \\
3: & \quad \text{Compute } U_j \text{ from } x_j \text{ using Eq. (4.50).} \\
4: & \quad N_{2,j}[\kappa] \leftarrow 1 + \sum_{\rho=0}^{1} U_j^2[\rho, \kappa] \quad \triangleright \text{Defined for simplicity of gradient.} \\
5: & \quad N_{1,j}[\kappa] \leftarrow \left\{ N_{2,j}[\kappa] \right\}^{-\frac{1}{2}} \quad \triangleright \text{Same as Eq. (4.66)} \\
6: & \quad N_{0,j}[\rho, \kappa] \leftarrow \begin{cases} -N_{1,j}[\kappa]U_j[\rho, \kappa] & \text{for } \rho = 0 \text{ or } 1 \\ N_{1,j}[\kappa] & \text{for } \rho = 2. \end{cases} \quad \triangleright \text{Same as Eq. (4.67)} \\
7: & \quad E_{1,j}[\kappa] \leftarrow \sum_{\rho=0}^{2} N_{0,j}[\rho, \kappa]X'_{j-1}[\rho, \kappa] \quad \triangleright E_{0,j}[\kappa] \text{ was not calculated in the notional algorithm so this alternate definition is employed.} \\
8: & \quad \text{return } x_j, L_j, N_{0,j} \text{ and } E_{1,j}.
\end{align*}
\]

The RMAD gradient for ray transfer to a freeform surface begins with prior gradient calculations finding values for \( \mathcal{X}'_j, \mathcal{L}_j, \mathcal{N}_{0,j} \) and \( \mathcal{E}_{1,j} \). Then the gradient steps associated with line 7 of Algorithm 4.2 are:

\[
\begin{align*}
\mathcal{N}_{0,j}[\rho, \kappa] & \quad + = X'_{j-1}[\rho, \kappa]E_{1,j}[\kappa], \\
\mathcal{X}'_{j-1}[\rho, \kappa] & \quad + = N_{0,j}[\rho, \kappa]E_{1,j}[\kappa],
\end{align*}
\quad (4.74)
\]

and the steps for line 6 are:

\[
\begin{align*}
\mathcal{N}_{1,j}[\kappa] & \quad + = \mathcal{N}_{0,j}[2, \kappa] - \sum_{\rho=0}^{1} U_j[\rho, \kappa]\mathcal{N}_{0,j}[\rho, \kappa] \\
\mathcal{U}_j[\rho, \kappa] & \quad + = -N_{1,j}[\kappa]\mathcal{N}_{0,j}[\rho, \kappa] \quad \text{for } \rho = 0 \text{ or } 1.
\end{align*}
\quad (4.75)
Evaluating the gradient steps for lines 5 and 4 yields

$$\overline{N}_{2j}[\kappa] += -\frac{1}{2} \left( N_{2j}[\kappa] \right)^{-\frac{3}{2}} \overline{N}_{1j}[\kappa],$$

$$\overline{u}_j[\rho, \kappa] += 2u_j[\rho, \kappa] \overline{N}_{2j}[\kappa] \quad \text{for } \rho = 0 \text{ or } 1. \quad (4.76)$$

Next, by line 3, $\overline{u}_j$ is used in the slope gradient of Section 4.2.3.1 to find $\overline{x}_j^{(u)}$, which can be further accumulated with

$$\overline{x}_j[\rho, \kappa] += \overline{x}_j^{(u)}[\rho, \kappa] \quad \text{for } \rho = 0 \text{ or } 1. \quad (4.77)$$

Then, the gradient steps for line 2 of Algorithm 4.2 are identical to Eqs. (4.30)-(4.32) from the spherical surface ray transfer gradient. Next, as described above, the gradient steps for line 1 require using Eq. (4.73) to calculate $\overline{\psi}_j$ given the $\overline{L}_j$ found using Eq. (4.31). Next, one calculates the several RMAD gradient steps associated Eq. (4.71). These begin with

$$\overline{x}_{j-1}[2, \kappa] += \overline{\psi}_j[\kappa],$$

$$\overline{X}_{j-1}[2, \kappa] += L_j[\kappa] \overline{\psi}_j[\kappa]. \quad (4.78)$$

Then the sag gradient is executed assuming $\overline{s}_j = \overline{\psi}_j$ yielding the $\overline{x}_j^{(s)}$ needed by the next stage of ray intersection gradient. Finally, the gradient steps associated with the argument of $S$ in Eq. (4.71) are evaluated:

$$\overline{x}_{j-1}[\rho, \kappa] += \overline{x}_j^{(s)}[\rho, \kappa] \quad \text{for } \rho = 0 \text{ or } 1,$$

$$\overline{X}_{j-1}[\rho, \kappa] += L_j[\kappa] \overline{x}_j^{(s)}[\rho, \kappa] \quad \text{for } \rho = 0 \text{ or } 1. \quad (4.79)$$
Since $x'_{j-1}$ and $X'_{j-1}$ have been accumulated above and the surface sag and slope gradient computations accumulated $\overrightarrow{c}_j$, $\overrightarrow{k}_j$ and $\overrightarrow{a}_j$, this concludes the RMAD gradient steps for ray transfer to an aspheric or freeform surface.

### 4.2.4 Accumulating Optical Path Distance

To the overall optical path length, we add the physical path length multiplied by the index of refraction of previous surface,

$$o[\kappa] +/- n_{j-1}L_j[\kappa].$$ (4.80)

The gradient steps are similarly simple,

$$\overline{n}_{j-1} +/- \sum_{\kappa=0}^{k-1} L_j[\kappa] \overrightarrow{\delta}[\kappa],$$

$$\overline{L}_j[\kappa] +/- n_{j-1}\overrightarrow{\delta}[\kappa].$$ (4.81)

### 4.2.5 Refraction

The refraction step applies the vector version of the Snell’s law given the cosine of the angle of incidence on the surface, $E_{1,j}[\kappa]$, the intersection point $x_j$ and the directional cosine of the incident ray $X'_{j-1}$.

Motivated by [51], we define the scalar

$$\mu_j = \frac{n_{j-1}}{n_j},$$ (4.82)
but for convenience we also define the following auxiliary arrays

\[ \chi_j[\kappa] = 1 - E^2_{1,j}[\kappa], \]  
\[ \eta_j[\kappa] = 1 - \mu_j^2 \chi_j[\kappa], \]  

(4.83)  
(4.84)

to find the cosine of the angle of refraction,

\[ E'_{1,j}[\kappa] = \sqrt{\eta_j[\kappa]}, \]  

(4.85)

by Eq. (A.2k) in [53]. Next, define

\[ \Sigma_j[\kappa] = E'_{1,j}[\kappa] - \mu_j E_{1,j}[\kappa]. \]  

(4.86)

Assuming refraction at a spherical surface, Eqs. (A.2m)-(A.2o) in [53] can be condensed into just two steps,

\[ X_j[\rho, \kappa] + \mu_j X'_{j-1}[\rho, \kappa] - c_j \Sigma_j[\kappa] x_j[\rho, \kappa], \]  
\[ X_j[2, \kappa] + \Sigma_j[\kappa], \]  

(4.87)

completing the refraction calculation. Alternately, for an asphere or freeform surface we rewrite Eqs. (A.3p)-(A.3r) in [53] as

\[ X_j[\rho, \kappa] = \mu_j X'_{j-1}[\rho, \kappa] + \Sigma_j[\kappa] N_j[\rho, \kappa]. \]  

(4.88)
The gradient for refraction involving an aspherical or freeform surface begins with the steps associated with Eq. (4.88),

\[
\bar{\mu}_j = \sum_{\rho=0}^{K-1} \sum_{\kappa=0}^{\rho} X'_{j-1}[\rho, \kappa] \bar{X}_j[\rho, \kappa],
\]

\[
X'_{j-1}[\rho, \kappa] = \mu_j \bar{X}_j[\rho, \kappa],
\]

(4.89)

and

\[
\bar{\Sigma}_j[\kappa] = \sum_{\rho=0}^{2} N_j[\rho, \kappa] \bar{X}_j[\rho, \kappa],
\]

\[
N_j[\rho, \kappa] = \Sigma_j[\kappa] \bar{X}_j[\rho, \kappa].
\]

(4.90)

Alternately, for refraction at a spherical surface using Eq. (4.87), the gradient steps are Eq. (4.89) and

\[
\bar{c}_j = - \sum_{\rho=0}^{K-1} \sum_{\kappa=0}^{\rho} \Sigma_j[\kappa] x_j[\rho, \kappa] \bar{X}_j[\rho, \kappa],
\]

\[
\bar{\Sigma}_j[\kappa] = x_j[2, \kappa] - c_j \sum_{\rho=0}^{2} x_j[\rho, \kappa] \bar{X}_j[\rho, \kappa],
\]

\[
\bar{x}_j[\rho, \kappa] = - c_j \Sigma_j[\kappa] \bar{X}_j[\rho, \kappa].
\]

(4.91)
The remainder of the refraction gradient steps are independent of whether the surface was spherical, aspherical or freeform. Applying the gradient rules to Eq. (4.86) yields

\[ E'_{1,j} = \Sigma_j, \]

\[ \mu_j = - \sum_{\kappa=0}^{K-1} E_{1,j}[\kappa] \Sigma_j[\kappa], \]

\[ \bar{E}_{1,j}[\kappa] = - \mu_j \Sigma_j[\kappa], \]  \hspace{1cm} (4.92)

and given Eq. (4.85),

\[ \eta_j[\kappa] = \frac{1}{2 \sqrt{\eta_j[\kappa]}} E'_{1,j}[\kappa]. \]  \hspace{1cm} (4.93)

The derivative contributions due to the auxiliary variable \( \eta_j \) in Eq. (4.84) is

\[ \bar{\chi}_j[\kappa] = - \mu_j^2 \eta_j[\kappa], \]

\[ \bar{\mu}_j = - \sum_{\kappa=0}^{K-1} \chi_j[\kappa] \eta_j[\kappa], \]  \hspace{1cm} (4.94)

where \( \bar{\mu}_j^2 \) is understood as the gradient of the error metric with respect to the explicit dependence on \( \mu_j^2 \) in Eq. (4.84). This contribution to the gradient terms is accumulated with

\[ \bar{\mu}_j = 2 \mu_j \mu_j^2. \]  \hspace{1cm} (4.95)

The gradient steps for Eq. (4.83) are

\[ \bar{E}_{1,j}[\kappa] = -2E_{1,j}[\kappa] \bar{\chi}_j[\kappa], \]  \hspace{1cm} (4.96)
while, finally, the gradient for Eq. (4.82) is

\[
\bar{n}_{j-1} + = \frac{\mu_j}{n_j} \\
\bar{n}_j + = - \frac{n_{j-1}}{n^2_j} \mu_j.
\] (4.97)

4.3 The Diffraction Integral

In this section, we describe a model for converting raytrace data into a differentiable model for the detected PSF beginning with the computation of the PSF intensity.

Although we do not trace rays to it for prescription retrieval, it will be convenient to discuss a flat image plane surface that is next after the reference sphere and that has surface index \( I = r + 1 \). In Section 4.2, we required that the reference sphere’s center be located at the centroid of the detected PSF. To accomplish this, we assume that the operator has found decentralization \( \Xi_r \) and \( \Upsilon_r \) that center the reference sphere on the PSF centroid. Further, let the reference sphere have radius \( R \) and that \( t_r = R \) such that if \( \Xi_I = \Upsilon_I = 0 \), the image plane would also be centered on the centroid of the PSF data. Finally, we require that the image plane be parallel to the \( x-y \) plane of the reference sphere’s coordinate system by \( \alpha'_I = \beta'_I = \gamma'_I = 0 \). We do not require the image plane to be parallel to parts of the optical system as rotations may be applied to surfaces prior to surface \( I \).

If \((x_r, y_r, z_r)\) is a local coordinate of a point on the reference sphere and \( \lambda \) the wavelength of light in the media after the last physical surface, we assume a field
on the reference sphere due to a wavefront amplitude function $A_W$ and wavefront aberration function $W$. These concepts are ubiquitous in aberration theory [54–56, 58, 64], but we discuss specific models for $A_W$ and $W$ later.

The diffraction integral restricts the extent of the non-zero areas of the wavefront amplitude. In particular, the distance

$$d(x_r, y_r, z_r, x_I, y_I) = (x_r - x_I)^2 + (y_r - y_I)^2 + (z_r - R)^2,$$  \hspace{1cm} (4.99)

between any point $(x_I, y_I)$ on the image having significantly non-zero amplitude and any points $(x_r, y_r, z_r)$ having significantly non-zero wavefront amplitude must be sufficiently close to $R$ that [55, 56]

$$\frac{R - d(x_r, y_r, z_r, x_I, y_I)}{2R} \approx 0.$$  \hspace{1cm} (4.100)

If the location and shape of the exit pupil relative to the measured PSF is approximately known, the degree to which Eq. (4.100) might be violated can be assessed.

Assuming Eq. (4.100), the Kirchhoff approximation [54], and a positive time dependent solution to the wave equation, the field in the image plane is [55]

$$G'(x_I, y_I) = \frac{-i \exp \left[ -i \varepsilon(x_I, y_I) - i \frac{2\pi R}{\lambda} \right]}{\lambda R} \int_{x_I, y_I = -\infty}^{\infty} f(x_r, y_r) \exp \left[ i \frac{2\pi}{\lambda R} \left( x_I x_r + y_I y_r \right) \right] \frac{dx_r \, dy_r}{u},$$  \hspace{1cm} (4.101)
where
\[ \varepsilon(x_I, y_I) = \frac{2\pi}{\lambda} \left( \frac{x_I^2 + y_I^2}{2R} \right), \] (4.102)
and \( u \) is the cosine between the z axis and the normal of a differential area element at \((x_r, y_r)\) on the reference sphere. In [55], the variation in \( u \) across the wavefront amplitude is later approximated by an average value based on the angle of incidence on the image plane. The value of \( \varepsilon \) is important if any further coherent or partially coherent calculations are done with \( G' \) [56]. However, accurate modeling of the phase of \( G' \) is neglected here since only the PSF intensity \( |G'|^2 \) is measured. The complex factor \(-i \exp[-i\varepsilon(x_I, y_I) - i2\pi R/\lambda]\) also has no effect on our understanding of \( f \) so we will neglect this term as well. Substituting Eq. (4.98) into Eq. (4.101) and letting \( B(x_r, y_r) = A_W(x_r, y_r)/(\lambda Ru) \),
\[ G(x_I, y_I) = \int_{x_y, y_y = -\infty}^{\infty} B(x_r, y_r) \exp\left[ i\frac{2\pi}{\lambda} W(x_r, y_r) \right] \exp\left[ i2\pi \left( \frac{x_I x_r + y_I y_r}{\lambda R} \right) \right] dx_r dy_r. \] (4.103)

For general wavefront aberrations, Eq. (4.103), an inverse Fourier transform, does not have an analytic expression and must be evaluated using one of many numerical methods like those found in [55, 65–69]. Many methods address the related problem of calculating the autocorrelation of \( f \), an unscaled version of the optical transfer function (OTF), which could then be Fourier transformed to yield the PSF. Gaussian quadrature summation (GQS) [55, 65, 66, 68] techniques require a moderate number of samples of \( W \) and its spatial derivatives, in a geometry that varies with the exact
method and pupil shape. Samples of $W(x_r, y_r)$ can be obtained by tracing rays through the necessary points on the reference sphere. Alternately, some methods fit a Zernike polynomial model for $W(x_r, y_r)$ using a sparse sample of rays [66] with a method such as found in [70]. Fast evaluation of the fitted polynomial can be done at the necessary sample locations. This method carries with it the risk that a polynomial of limited order might not represent the true wavefront aberrations and automated checks should be made [66]. For simplicity of the gradient calculation, this formulation of prescription retrieval will not fit a Zernike polynomial function to the wavefront aberration.

The fast Fourier transform (FFT) mitigates much of the work involved to compute PSF intensities on a uniformly-sampled grid of pixels [69, 71]. It comes at the expense of requiring a particular grid of samples of $W(x_r, y_r)$ at a particular spacing, though this can be mitigated with the Winograd Fourier transform [69], the chirp-Z transform [72, 73] or the matrix multiply discrete Fourier transform (MMDFT) introduced in Section 3.2. For any of the uniformly-sampled grid methods, the sampling will be denser than that of the GQS techniques for equivalent accuracy in the lowest spatial frequencies of the PSF [65]. The GQS methods, which use wavefront derivative information, may be advantageous for large PSFs for this and other reasons noted in [65]. For all methods, as the magnitude of the wavefront aberration increases, the density of samples on the reference sphere must increase to accurately represent the lowest spatial frequencies present in the PSF. If the reference sphere sampling is insufficient, periodic repetitions of the PSF will encroach on the borders of the
calculated field in a form of aliasing. These other repetitions are the error associated with approximating Eq. (4.103) using a discrete sum of uniformly spaced samples of the reference sphere. As a practical matter, propagations in phase retrieval are greatly aided by highly optimized libraries for computing the FFT [74] and matrix multiplication. For this reason, and because they are familiar to the methods of phase retrieval, discrete Fourier transform propagation methods are adopted for this prescription retrieval method.

Another concern in calculating the OTF or PSF of an optical system is the radiometric implications of a particular sampling of the image and object reference spheres. For instance, if the rays evenly sample a grid on the image-space reference sphere, they may not evenly sample the object-space reference sphere due to pupil aberrations and deviations from the sine condition. The integrand of Eq. (4.103) would then need to include a multiplicative factor accounting for the difference in the differential area on the image sphere represented by each ray and the corresponding differential area on the object sphere [75]. This factor depends on \((x_r, y_r)\) and would difficult to distinguish from \(B(x_r, y_r)\). PSFs from physical systems are likely to include additional amplitude effects that may be unknowable at the time of the measurement, such as (i) variations of Fresnel reflection/transmission, (ii) manufacturing defects such as scratches near a pupil plane, and (iii) the entrance pupil illumination effects of a resolved coherently illuminated target. Consequently, we treat \(B(x_r, y_r)\) as a quantity to be estimated by prescription retrieval since this includes both unknowns, which must be retrieved, and some knowns that we do not wish to model such as
the radiometry factor and the dependence on $u$. In phase retrieval, it is possible to estimate the exit pupil amplitude if there is a good support constraint. However, in prescription retrieval for a general off-axis field point, $B(x_r, y_r)$ is a warped image of the true aperture stop with some additional unknown amplitude modulations. The warping is determined by distortions in the imaging of the aperture stop onto the image reference sphere. This is most notable for off-axis field points where a circular stop maps approximately to an ellipse, unless one works in the canonical coordinates of Hopkins [58, 64]. Since the warping varies with prescription, it is easiest to enforce the support constraint at the physical stop surface as this can be common to all PSFs.

For our prescription retrieval, synthesis of the modeled PSF begins with a raytrace, as described in Section 4.2, that densely sampled the image-space reference sphere. Next, a discrete model for $B(x_r, y_r)$ is interpolated from the estimated aperture stop transmission function according to Section 4.3.1. This allows the enforcement of a hard-edged support constraint in the stop surface plane for every field. The reference sphere amplitude contribution from each ray is then multiplied with a complex phasor based on the ray’s optical path distance as described in Section 4.3.2. The resulting contribution to the integrand in Eq. (4.103) is summed using either an exact summation outlined in Section 4.3.3 or a faster approximation described in Section 4.3.4 that assumes uniformly-spaced samples on the image reference sphere.
4.3.1 Discrete Model for the Reference Sphere Amplitude

In this section, we describe the interpolation of an \( N' \) by \( N' \) sample array \( A \), the transmission function of the aperture stop, in order to find the amplitude associated with the each ray. Let \( A \) intersect the optical axis at array index \([n'_0, n'_0]\) and let each sample be spaced by a physical distance of \( \Delta_s \) in both directions. For compatibility with the interpolation scheme in Section 3.7, we require the number of rays traced \( K = N^2 \), for some integer \( N \).

Given the coordinates \( x_s \),

\[
n'[\rho, \kappa] = \begin{cases} 
\frac{x_{[0,\kappa]}}{\Delta_s} + n'_0 & \text{for } \rho = 0 \\
\frac{x_{[1,\kappa]}}{\Delta_s} + n'_0 & \text{for } \rho = 1
\end{cases}
\]  

(4.104)

is an array of columns corresponding to coordinates relative to samples of \( A \) for each ray. These coordinates will not lie on integer indexes of \( A \), so bilinear interpolation will be used to find the amplitude for each ray. We use Eq. (3.28) to calculate an interpolated output given \( A \), the input array to be resampled, and the coordinate array \( n' \). The result of that bilinear interpolation is an \( N \) by \( N \) two-dimensional array of amplitudes. This result will be reinterpreted as an \( N^2 \) long one-dimensional array, \( B \), that is linearly indexed in row-major order and provided to the field model described in the next Section. While \( A \) is assumed to be sampled on a uniform grid, the samples of \( B \) are not necessarily arranged in any particular spatial distribution in the stop plane.

Note that Eq. (4.104) has no dependence on the third row of \( x_s \) which corresponds to the \( z \)-intercept coordinate of the ray with the aperture stop surface. If \( c_s \neq 0 \), it is as
if there is a transmission function imposed along a circular curve whose orthographic projection on to the $x$-$y$ axis is $A$.

The gradient of the reference sphere amplitude model receives an array which is reinterpreted as an $N$ by $N$ two-dimensional array $\mathbf{B}$. If the transmission function of the aperture stop is varying, the gradient contribution $\mathbf{A}$ can be found using Algorithm 3.1 in Section 3.7. If any part of the prescription parameters on surfaces prior to the stop are varying, the gradient contribution due to the amplitude model’s dependence on $\mathbf{x}_s$ must be calculated. Given $\mathbf{B}$, the arrays $\mathbf{F}\rightarrow$ and $\mathbf{F}\downarrow$ can be calculated from Eqs. (3.32) and (3.31). Then,

$$
\mathbf{n}[\rho,\kappa] = \begin{cases} 
\mathbf{F}\rightarrow[\kappa] & \text{for } \rho = 0 \\
\mathbf{F}\downarrow[\kappa] & \text{for } \rho = 1,
\end{cases} \quad (4.105)
$$

where $\mathbf{F}\rightarrow$ and $\mathbf{F}\downarrow$ are interpreted as one dimensional arrays. Finally the gradient step for Eq. (4.104) is

$$
\mathbf{x}_s[\rho,\kappa] = \begin{cases} 
\mathbf{x}^{(0)}[\rho,\kappa] & \text{for } \rho = 0 \\
\frac{\mathbf{x}^{(1)}[\rho,\kappa]}{\Delta_s} & \text{for } \rho = 1 \\
0 & \text{for } \rho = 2.
\end{cases} \quad (4.106)
$$

### 4.3.2 Reference Sphere Field

Let $\mathbf{o}'$ be the optical path distance of some reference ray, like a chief ray, between the object reference sphere and the image reference sphere. The wavefront aberration is
\[ W(x_r, y_r) = o' - o(x_r, y_r), \] (4.107)

where \( o(x_r, y_r) \) is the optical path distance traveled by a ray from one reference sphere to the next and arriving on the image reference sphere at \((x_r, y_r)\). Since

\[
\exp \left[ i \frac{2\pi}{\lambda} W(x_r, y_r) \right] = \exp \left[ i \frac{2\pi}{\lambda} o' \right] \exp \left[ -i \frac{2\pi}{\lambda} o(x_r, y_r) \right] \tag{4.108}
\]

and we neglect constant phases in Eq. (4.103), we define

\[
f[\kappa] = B[\kappa] \exp \left( -i \frac{2\pi}{\lambda} o[\kappa] \right) \tag{4.109}
\]

as the discrete model for each ray’s contribution to the field on the reference sphere. We assume that \( A \) and \( B \) are real valued.

The gradient contributions due to \( f \) depending on the amplitude model requires applying Eqs. (49) and (67) in [4] to Eq. (4.109) to find

\[
\overline{B} = \Re \left\{ \exp \left( i \frac{2\pi}{\lambda} o[\kappa] \right) \overline{f[\kappa]} \right\}, \tag{4.110}
\]

where the realness of \( B \) was assumed. By Eqs. (47) and (57) in [4], the other gradient step corresponding to Eq. (4.109) is

\[
\overline{o[\kappa]} = -\frac{2\pi}{\lambda} \Im \left\{ f^*[\kappa] \overline{f[\kappa]} \right\}. \tag{4.111}
\]
4.3.3 Nonuniform Discrete Diffraction Integral

Let the $M \times M$ array $G[m_r, m_c]$ represent the field in the image plane with samples spaced by a physical distance $\Delta_I$ in both directions. The physical coordinates of the sample at index $[m_r, m_c]$ are

$$(x_l, y_l) = (m_c \Delta_I - m', m_r \Delta_I - m')$$

(4.112)

if $G$ intersects the optical axis at index $[m', m']$. Recalling that $x_r$ contains coordinates $(x_r, y_r)$ of the ray intersection for each ray,

$$G[m_r, m_c] = \sum_{\kappa} f[\kappa] \exp \left[ i 2 \pi \psi (m_c x_r[0, \kappa] + m_r x_r[1, \kappa]) \right],$$

(4.113)

is a discrete summation approximating Eq. (4.103) where $\psi = \Delta_I / (\lambda R)$ and a complex linear phase involving $m'$ has been neglected since it does not affect the PSF intensity. We refer to Eq. (4.113) as the nonuniform diffraction integral since no assumption has been made about the distribution of $x_r[0, \kappa]$ and $x_r[1, \kappa]$. It is a naive summation of a discrete Fourier transform and computationally intensive if implemented as written. However, since the output is uniformly spaced, it can be approximated using the type-1 non-uniform fast Fourier transform [76] to arbitrary accuracy.

By Section 8 in [4], one of the gradient steps associated with Eq. (4.113) is

$$\overline{f}[\kappa] = \sum_{m_r=0}^{M} \sum_{m_c=0}^{M} \overline{G}[m_r, m_c] \exp \left[ -i 2 \pi \psi (m_c x_r[0, \kappa] + m_r x_r[1, \kappa]) \right].$$

(4.114)

This step could be computed by a type-2 non-uniform fast Fourier transform [76]. However, since $\overline{x}_r$ must also be evaluated, and that gradient also requires a particular
Algorithm 4.3 Calculation of \( \tilde{f} \) and \( \tilde{x}_r \) from \( \tilde{G} \).

1: for \( \kappa = 0 \) to \( K - 1 \) do

2: \[ \Theta[m_r, m_c] \leftarrow 2\pi\psi(m_rx_r[0, \kappa] + m_rx_r[1, \kappa]) \]

3: \[ \Psi[m_r, m_c] \leftarrow \exp(i\Theta[m_r, m_c]) \]

4: \[ \bar{f}[\kappa] \leftarrow \sum_{m_r=0}^{M} \sum_{m_c=0}^{M} \Psi^*[m_r, m_c] \bar{G}[m_r, m_c] \] \( \triangleright \) As in Eq. (4.114).

5: \[ \bar{\Psi}[m_r, m_c] \leftarrow f^*[\kappa] \bar{G}[m_r, m_c] \] \( \triangleright \) The contribution due to \( \Psi \) appearing in Eq. (4.113).

6: \[ \bar{\Theta}[m_r, m_c] \leftarrow \mathcal{S}(\Psi[m_r, m_c] \bar{\Psi}[m_r, m_c]) \] \( \triangleright \) By Eq. (57) in [4] applied to line 3.

7: \[ \tilde{x}_r[0, \kappa] + = 2\pi\psi \sum_{m_r=0}^{M} m_r \sum_{m_c=0}^{M} \bar{\Theta}[m_r, m_c] \] \( \triangleright \) By Eq. (61) in [4] applied to line 2.

8: \[ \tilde{x}_r[1, \kappa] + = 2\pi\psi \sum_{m_r=0}^{M} m_r \sum_{m_c=0}^{M} \bar{\Theta}[m_r, m_c] \]

9: end for

10: return \( \tilde{f} \) and \( \tilde{x}_r \)

complex exponential, it is likely more efficient to accumulate \( \tilde{f} \) and \( \tilde{x}_r \) simultaneously. In Algorithm 4.3 below, we list one way of jointly accumulating \( \tilde{f} \) and \( \tilde{x}_r \), although there may be more optimal algorithms.

The principal advantage of the nonuniform discrete diffraction integral, Eq. (4.113), is that it accepts an ensemble of rays that arbitrarily sample the image reference sphere. This is useful in this prescription retrieval technique because we have not restricted the rays to intersect the reference sphere at particular points. At least, we wish to avoid such ray aiming for every single evaluation of the error metric for both computational speed and simplicity. This avoids differentiating an additional ray-aiming process.
4.3.4 Uniformly-Sampled Approximate Diffraction Integral

Since the nonuniform diffraction integral involves a Fourier transform on a spatial distribution that is not sampled on a uniform grid, it cannot directly utilize computational methods like the FFT, the chirp-Z transform (CZTDFT) [72], or the matrix-multiply DFT (MMDFT) discussed in Section 3.2. In this section, we develop an approximate diffraction integral that can utilize more efficient methods that require uniform sampling on the image reference sphere.

For systems obeying the sine condition, there exists a parameterization of the object and image reference spheres involving Hopkin’s canonical coordinates [58]. In these canonical coordinates, rays crossing the object point that produce an evenly-spaced grid of ray intersections on one reference sphere will do so simultaneously on the other. Thus, there exists a set of rays $x_0$ and $X_0$ that yield a nearly-evenly spaced grid of ray intersections on the image reference sphere. This is because an evenly sampled grid on the image sphere could be generated, expressed in canonical coordinates and then those canonical coordinates could be expressed in regular object reference sphere coordinates to find $x_0$. However, a more practical means would be to use an iterative ray-aiming process to find an $x_0$ that yields a regularly spaced grid of intersections $x_r$.

Supposing that a nominal prescription for the system is available, a set of rays that evenly sample the image reference sphere can be found. We further suppose that, during the process of optimizing, this grid of ray intersections will likely only undergo modest transformation as the modeled prescription fits the actual prescription. Consid-
ering only optical systems for which these transformations are essentially translation and scaling, a discrete Fourier transform, which requires uniform sampling such as the CZT or MMDFT, may be employed to approximate Eq. (4.113). Translation of the grid may be neglected since, by the discrete Fourier shift theorem, that yields a linear complex phase in the image plane which does not affect intensity.

As outlined above, assume that the \( x_r \) intercept coordinates in the first row of \( x_r \) and the \( y_r \) intercept coordinates in the second row of \( x_r \) are approximately an evenly spaced grid in \((x_r, y_r)\). Since a factor of \( R^{-1} \) appear in the Fourier-kernel of Eq. (4.113), this grid corresponds to a uniform sampling of the directional cosines \((x_r/R, y_r/R)\) of rays incident on the image plane. We can then reinterpret \( f[\kappa] \) as a two-dimensional complex array \( f[\kappa][n_r, n_c] \) with indices \( n_r \) and \( n_c \), each ranging from 0 to \( N-1 \), corresponding to the amplitudes at points on this uniform grid in directional-cosine space. We approximate Eq. (4.113) as

\[
G[m_r, m_c] \approx \sum_{n_r=0}^{N-1} \sum_{n_c=0}^{N-1} f[\kappa][n_r, n_c] \exp \left[ -i2\pi \left( \alpha_c m_c n_c + \alpha_r m_r n_r \right) \right],
\]

(4.115)

for some scale factors \( \alpha_c \) and \( \alpha_r \) discussed below. A negative sign was introduced to the exponent of Eq. (4.115) to make it consistent with the CZTDFT and MMDFT definitions which are usually written as forward Fourier transforms. We allow the scale to differ between the row and column dimensions in cases where there is an extreme difference in numerical aperture about the cross sections of the beam that happen to align with the \( x \) and \( y \) axes. This could happen for field points on the \( x \) or \( y \) axis where the sagittal and tangential axes of a highly off-axis beam usefully align
with x and y axes. In this case, we might expect that during optimization the sagital
and tangential scaling might prefer to vary independently of one another.

We estimate the scale parameters $\alpha_c$ and $\alpha_r$ from the ray intersections $x_r$. Let the
first and second rows of $x_r$ be reinterpreted as the two-dimensional arrays $x[n_r, n_c]$ and
$y[n_r, n_c]$, respectively. Assuming that the first column of $x$ corresponds to the column
of rays in the grid with the least x coordinates of their row, the average of these x
coordinates is

$$x_- = N^{-1} \sum_{n_r=0}^{N-1} x[n_r, 0], \quad (4.116)$$

while the average of the x coordinates for the column with the greatest value is

$$x_+ = N^{-1} \sum_{n_r=0}^{N-1} x[n_r, N-1], \quad (4.117)$$

and the average space between samples along the x axis is $(x_+ - x_-)/(N - 1)$. Then, let

$$\alpha_c = -\psi \frac{x_+ - x_-}{(N - 1)}$$

$$= -\psi \frac{\sum_{n_r=0}^{N-1} x[n_r, N-1] - x[n_r, 0]}{N(N - 1)} \quad (4.118)$$

and similarly,

$$\alpha_r = -\psi \frac{\sum_{n_c=0}^{N-1} y[N-1, n_c] - y[0, n_c]}{N(N - 1)}. \quad (4.119)$$

Having evaluated $\alpha_r$ and $\alpha_c$, the CZTDFT outlined in [72] may be applied to $f$ to
estimate $G$. Alternately, the MMDFT may be applied although the MMDFT described
in Section 3.2 was for a scale factor being equal along both axes. If one calculates \( \Omega_r \) from Eq. (3.13) assuming \( \alpha = \alpha_r \), and \( \Omega_c \) from Eq. (3.13) assuming \( \alpha = \alpha_c \), then

\[
G \approx \Omega_r^T \star f_\square \star \Omega_c,
\]

(4.120)

where \( \star \) indicates matrix-multiplication.

The gradient computation begins with \( \overline{G} \) being calculated by an earlier portion of the gradient computation. In the case of the CZTDFD, \( f_\square, \alpha_c \) and \( \alpha_r \) are calculated by the method in [72]. If the MMDFT was used for the calculation, repeated application of Eq. (50) in [4] to Eq. (4.120) shows that

\[
\overline{f_\square} = \Omega_r^* \star \overline{G} \star \Omega_c^H,
\]

\[
\overline{\Omega_c} = f_\square^{\dagger} \star \Omega_c^* \star \overline{G},
\]

\[
\overline{\Omega_r} = f_\square^{\dagger} \star \Omega_c^* \star \overline{G}^T,
\]

(4.121)

where the \( ^* \) represents elementwise complex conjugation and \( ^H \) represents the complex conjugation and transposition of the array. A similar evaluation was performed in Section 3.8. Then \( \overline{\alpha_c} \) and \( \overline{\alpha_r} \) can be calculated from \( \overline{\Omega_c} \) and \( \overline{\Omega_r} \), respectively, using Eqs. (3.50) and (3.51).

Once computed from the CZTDFD or MMDFT, \( \overline{f_\square} \) can be interpreted as a one dimensional array to yield the \( \overline{f} \) required in Section 4.3.2. The gradient step for Eq. (4.118) consists of adding the value \( \psi \overline{\alpha_c}/[N(N - 1)] \) to \( \overline{x_r}[0, \kappa] \) for all rays \( \kappa \) corresponding to the column of rays with the smallest \( x \) reference sphere intersections, and the negative of that value added to the column of rays having the largest.
Similarly, $\psi r/[N(N - 1)]$ is added to $x_r[1, \kappa]$ for all rays $\kappa$ corresponding to the row of rays with the smallest $y$ intersections, and the negative of that value is added to the row of rays having the largest.

### 4.4 Error Metric for Prescription Retrieval

In this section, we describe a bias and gain invariance error metric [77]. For Sections 4.2 and 4.3, we suppressed the dependence of all quantities on particular values of the loop dimensions mentioned in Section 4.2.

Supposing that the model for each PSF requires just one raytrace, there is no need to employ the loop dimension concept, and the bias and gain invariant metric used in Section 3.2 suffices. Letting $G$ from Section 4.3 be subscripted with the index of the $k$th PSF, the intensity of the $k$th PSF can be found with

$$I_k = |G_k|^2,$$  \hspace{1cm} (4.122)

where the absolute value and squaring operations are performed element-wise. Our model for the intensity measured by the detector, including an unknown detector scalar gain $\beta_k$ and bias $\gamma_k$, is

$$M_k = \beta_k I_k + \gamma_k 1, \hspace{1cm} (4.123)$$

where $1$ is an array the same shape $I_k$ but with the value 1. Let $D_k$ be samples of intensity of the $k$th measured PSF, and let $w_k$ be a array of weightings on the detector coordinates $[m_r, m_c]$ that are zero for samples in $D_k$ containing bad pixels. A sum-
of-squared-differences error metric between the model and measured data for the $k$th PSF is

$$E_k = \sum_{m_r, m_c} w_k \circ (M_k - D_k)^2,$$  \hspace{1cm} (4.124)

where the $\beta_k$ and $\gamma_k$ that make it bias- and gain-invariant are listed in Eq. (3.64). Summing over the $E_k$ for $K$ individual PSFs yields a mean-squared error metric

$$E = \sum_{k=0}^{K} E_k.$$  \hspace{1cm} (4.125)

The gradient steps for this simple metric are listed in Eqs. (3.41) - (3.43).

**4.5 Simulated Prescription Retrieval Example**

To demonstrate the efficacy of the prescription retrieval with RMAD gradients, we implemented software that could execute the models and RMAD gradients described in the above sections. Our long term goal is to perform prescription retrieval with complicated non-radially symmetric systems like the James Webb Space Telescope introduced in Section 1.3. However, for simplicity, in this example we consider prescription retrieval of a simpler system analogous to the Hubble Space Telescope: a hypothetical 1 m F/21 Ritchey-Crétien telescope having the nominal prescription listed in Table 4.2. The design begins with a stop surface followed by a hyperbolic primary mirror that is almost a paraboloid. Next is a hyperbolic secondary mirror, and after that a refractive triplet to mitigate field curvature. The refractive triplet is
Table 4.2: Nominal prescription for 1 m F/21 Ritchey-Crétien telescope with refractive field-flattener. Object is at infinity and the full-field of view is 0.4 degrees.

<table>
<thead>
<tr>
<th>$j$</th>
<th>Name</th>
<th>Radius $c_{j}^{-1}$ (mm)</th>
<th>Conic const. $k_{j}$</th>
<th>Thick. $t_{j}$ (mm)</th>
<th>Material</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>Stop</td>
<td>-</td>
<td>-</td>
<td>1840</td>
<td>Air</td>
</tr>
<tr>
<td>1</td>
<td>Primary</td>
<td>-4630</td>
<td>-1.0747</td>
<td>-1840</td>
<td>Mirror to air</td>
</tr>
<tr>
<td>2</td>
<td>Secondary</td>
<td>-1214.15</td>
<td>-3.2197</td>
<td>1860</td>
<td>Mirror to air</td>
</tr>
<tr>
<td>3</td>
<td></td>
<td>-313.7</td>
<td>-</td>
<td>13.0</td>
<td>N-FK51A</td>
</tr>
<tr>
<td>4</td>
<td></td>
<td>236.6</td>
<td>-</td>
<td>17.5</td>
<td>N-KZFS11</td>
</tr>
<tr>
<td>5</td>
<td></td>
<td>-14000</td>
<td>-</td>
<td>7</td>
<td>Air</td>
</tr>
<tr>
<td>6</td>
<td></td>
<td>768</td>
<td>-</td>
<td>15</td>
<td>N-SF1</td>
</tr>
<tr>
<td>7</td>
<td></td>
<td>430.45</td>
<td>-</td>
<td>587.453</td>
<td>Air</td>
</tr>
<tr>
<td>8</td>
<td>Image</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

nominally composed of all spherical surfaces. Finally, there is an image plane which, for a 0.4 degree full field of view, is 147 by 147 mm wide.

A drawing of this optical system generated using the CodeV\textsuperscript{TM} software is shown in Fig. 4.1. Not represented in the drawing is an obscuration in the stop plane due to the secondary which is 230 mm in diameter. This obscuration, the stop and the shadows of struts holding the secondary are part of the nominal stop amplitude shown in Fig. 4.3 (a) and presumed present in the system. A transverse-ray aberration plot, also produced with CodeV\textsuperscript{TM}, is shown in Fig. 4.2. This system is limited by lateral chromatic aberration induced by the refractive field flattener. Still, when unperturbed from the nominal state, the wavefront error of this optical system was under 0.025 waves root-mean squared (RMS) for all field angles at the visible wavelengths from
Figure 4.1: Drawing of Ritchey-Crétien telescope simulated.

Figure 4.2: Transverse ray aberration plot of Ritchey-Crétien telescope employed to demonstrate prescription retrieval.
Figure 4.3: Stop amplitude model $A$ for a) the nominal system and b) the simulated perturbed system having additional features unknown to prescription retrieval.

486 to 656 nm. In this nominal state, the system was essentially diffraction limited. The detector at the image plane was assumed to have 5 $\mu$m wide pixels so that the intensity distribution on the image plane is slightly more than Nyquist sampled in intensity for 486 nm light.

The simulated prescription retrieval problem involved both the primary and secondary departing from the nominal hyperboloidal sag by a small amount. These errors were assumed to be smoothly varying and well modeled by coefficients of Zernike polynomials [60]. In the case of the primary mirror, these coefficients included 9th radial-degree and lesser terms, excluding the constant or piston term. For reference, the first 11 Zernike polynomial terms expressed in Cartesian coordinates, and their radial degrees, appear in Table 2.1. To generate the simulated data, coefficients were drawn from a zero-mean Gaussian distribution whose variance decreased with both increasing radial degree and for non-radially symmetric terms. These Zernike coeffi-
cients for the primary surface error relative to the nominal are plotted in Fig. 4.4 (a) and a map of the actual sag is shown in Fig. 4.5 (a). The root mean squared (RMS) sag departure for the primary was 33 nm with a peak to valley (PTV) value of 282 nm. For the secondary, these coefficients included 7th radial-degree terms and lesser but not the 2nd radial degree terms and lower. The 2nd radial degree and lower terms include defocus, astigmatism and the tip and tilt linear terms. In the preparation of this example, we were unable to accurately allocate defocus and astigmatism uniquely to the primary or secondary mirror surfaces given simulated PSF data using prescription retrieval. The division of these terms between the primary and secondary may fall into the null space of the measurement made by the PSFs. Consequently, we assumed the 1st radial degree terms and 2nd radial degree terms on the secondary were defined to be zero as they would just be confused with the corresponding terms on the primary. The particular random Zernike coefficients simulated on the secondary are plotted in Fig. 4.4 (b) and a map of the actual sag is shown in Fig. 4.5 (b). The root mean squared (RMS) sag departure for the secondary was 17 nm with a PTV value of 154 nm.

To make a more formidable challenge for prescription retrieval, we considered more aspects of the actual prescription to be unknown. We assumed that the final surface of the refractive triplet, surface 7, also deviated from its nominally spherical shape by an unknown sag term. For this surface error, we included all coefficients of 7th radial degree but the 1st degree linear terms and the constant term. Like the situation with the secondary mirror, we found that it was difficult to simultaneously estimate the linear tip and tilt terms of the 7th surface and the same for the primary.
Figure 4.4: Coefficients of the Zernike polynomial terms in the simulated data for prescription retrieval in (a) the primary surface, (b) the secondary surface and (c) the 7th surface. Polynomial terms are indexed by $j$ in the Noll ordering [60] and normalized over the unit circle of each surface’s clear aperture.
Figure 4.5: Surface sag deformation simulated in the PSF data for prescription retrieval in (a) primary surface, (b) secondary surface and (c) surface 7. Surface sag error estimated by prescription retrieval for the (d) primary surface, (e) secondary surface and (f) surface 7. Error in prescription retrieval result relative to the sag deformation simulated in the PSF data for (g) primary surface, (h) secondary surface and (i) surface 7. Each colorbar has a difference range of values, and positive sag indicates the surface is displaced in the positive $z$ direction relative to the nominal hyperbola or spherical surface. Not to scale: the outer circle represents the clear aperture of each surface which is different between primary, secondary and the 7th surface.
The random Zernike coefficients for the 7th surface are plotted in Fig. 4.4 (c) and a map of the actual sag is shown in Fig. 4.5 (c). The total number of unknown prescription parameters related to surface shape constituted 117 unknowns whose true values are those plotted in Fig. 4.4.

Finally, as in phase retrieval, there is the potential for the pupil illumination in an experiment to vary from the nominal amplitude due to various effects. For this example, we assumed that there were artifacts similar to a loose piece of insulation and a slowly varying shading that altered the stop illumination as shown in Fig. 4.3 (b). Consequently, the prescription retrieval algorithm estimated these artifacts from the simulated PSF data. To retrieve these additional amplitude defects, we optimized over pixels of the stop amplitude $A$ that were inside of the outer circular aperture stop. Including these pixels, the total number of prescription parameters to simultaneously estimate was 23,226 unknowns. For comparison, we are unaware of previous prescription retrieval examples that involve simultaneous estimation of more than approximately one hundred unknowns.

Accurately estimating the above-mentioned unknowns required many simulated PSFs spanning focus diversity, field diversity, and wavelength diversity. Variations in focus were simulated by altering $t_7$ from the thickness corresponding to the nominal best focus plane. Unresolved point sources were simulated at infinite conjugates and at various of field angles to yield field diversity in the collection of PSFs. Field diversity involved simulating the unresolved point sources at infinite conjugates but various field angles. Wavelength diversity was employed by varying the monochromatic
Table 4.3: Listing of defocus, field location, and wavelength for each PSF in prescription retrieval grouped by defocus and wavelengths. Defocus is a translation of image plane along the \( z \) axis relative to the nominal image location.

<table>
<thead>
<tr>
<th>PSF indices</th>
<th>Defocus (mm)</th>
<th>Wavelength (nm)</th>
<th>Object angles (degrees)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>–8.0</td>
<td>486.13</td>
<td>(0.0,0.0)</td>
</tr>
<tr>
<td>1-4</td>
<td>–2.5</td>
<td>495.76</td>
<td>(0.141,0.141), (–0.141,0.141), (–0.141,–0.141), (0.141,–0.141)</td>
</tr>
<tr>
<td>5-8</td>
<td>–2.5</td>
<td>486.13</td>
<td>(0.2,0.0), (0.0,0.2), (–0.2,0.0), (0.0,–0.2)</td>
</tr>
<tr>
<td>9-12</td>
<td>–2.0</td>
<td>486.13</td>
<td>(0.2,0.0), (0.0,0.2), (–0.2,0.0), (0.0,–0.2)</td>
</tr>
<tr>
<td>13-16</td>
<td>1.0</td>
<td>486.13</td>
<td>(0.141,0.141), (–0.141,0.141), (–0.141,–0.141), (0.141,–0.141)</td>
</tr>
<tr>
<td>17-20</td>
<td>1.0</td>
<td>486.13</td>
<td>(0.2,0.0), (0.0,0.2), (–0.2,0.0), (0.0,–0.2)</td>
</tr>
<tr>
<td>21-24</td>
<td>1.5</td>
<td>486.13</td>
<td>(0.141,0.141), (–0.141,0.141), (–0.141,–0.141), (0.141,–0.141)</td>
</tr>
<tr>
<td>25-28</td>
<td>2.0</td>
<td>486.13</td>
<td>(0.2,0.0), (0.0,0.2), (–0.2,0.0), (0.0,–0.2)</td>
</tr>
<tr>
<td>29</td>
<td>4.0</td>
<td>486.134</td>
<td>(0.0,0.0)</td>
</tr>
</tbody>
</table>
wavelength used during the raytrace and PSF synthesis. Given the many potential choices of defocus, field angle and wavelength, it was unclear what values of each to choose for the PSFs. The problem of choosing optimal values for diversity parameters in phase retrieval-like problems is an unsolved challenge, although considerable effort has gone into focus diversity such as the work in [15, 32]. For this example, we chose the values listed in Table 4.3 through intuition and a trial and error process that considered the Cramer-Rao lower bound for estimation for various combinations of diversity choices. Despite that effort, these values are almost certainly suboptimal in some sense. Other choices may have the property that they converge faster, mitigate stagnation better or yield more accurate retrieved values. Also, an actual prescription retrieval experiment will involve additional considerations, such as the selection of available point sources, that may require other choices.

To compute the PSF data, a grid of 192 by 192 rays was traced through the optical system. Then reference sphere fields were computed assuming the stop amplitude with artifacts shown in Fig. 4.3 (b). The system obeyed the sine condition well, so the uniform diffraction integral of Section 4.3.4 was utilized to speed computation of the PSFs. Poisson noise was added to each PSF simulating an average peak pixel intensity of $8 \times 10^5$ electrons and an additional Gaussian read noise of 15 electrons. These simulated data PSFs are shown in Fig. 4.6.

The software was configured for the nominal values shown in Table 4.2 with the values for the Zernike coefficients $a_1, a_2$ and $a_7$ initialized to zero. For verification purposes, we chose a limited number of coefficient values and evaluated the gradient
Figure 4.6: Simulated data PSFs. Intensity shown stretched to the one-quarter power.
with respect to these values using the method of finite differences (FD). A step size of the FD was found where the relative error between these derivatives and the output of the RMAD function was less than 0.02%. This suggests that the RMAD function provided the correct gradient for the error metric evaluated at the point where all Zernike coefficients were zero.

To avoid local minima, prescription retrieval was done using the following bootstrapping approach with three stages:

1. The error metric was minimized with respect to the coefficients having 3rd and lesser radial degrees on the three surfaces using the L-BFGS-B minimization algorithm [78]. The 1st radial degree terms for the secondary and 7th surface were held constant as was the 2nd radial degree terms for the secondary. The stop amplitude values were also held constant. This required 52 evaluations of the error metric and gradient with respect to 26 parameters.

2. The error metric was minimized with respect to the coefficients having 5th and lesser radial degrees holding constant the 1st and 2nd radial degree terms mentioned in Step 1. The stop amplitude was also not allowed to vary. This required 16 evaluations of the error metric and gradient with respect to 44 parameters.

3. The error metric was minimized with respect to all coefficients modeled, excluding the 1st and 2nd radial degree terms mentioned in Step 1. Also, values of the stop amplitude $A$ inside of the circular stop were allowed to vary so as to decrease the metric. This required 197 evaluations of the error metric and the gradient with respect to 23,226 parameters.
At the conclusion of the optimization, the retrieved Zernike sag coefficients were almost identical to the coefficients used to simulate the data. All of the coefficients plotted in Fig. 4.4 were found to within 0.12 nm or better. The retrieved sag for the primary, secondary and 7th surface are shown in Fig. 4.5 (d), (e), and (f) respectively. To visualize the errors in prescription retrieval, the retrieved sag was subtracted from the actual simulated sag for each surface and shown in Fig. 4.5 (g)-(i). The RMS wavefront sensing error for the primary and secondary mirrors was 0.1 nm while the error for the final refractive surface 0.3 nm. It is unlikely that such accurate results could be obtained in an absolute sense with a real optical system due to modeling errors and tolerances associated with the actual physical system. However, it demonstrates that the RMAD gradient was accurate enough to guide the optimizer to locate what is essentially the global minimum. Prescription retrieval also retrieved the stop amplitude shown in Fig. 4.7. This retrieval has the hard-edged and soft-edged artifacts of the pupil illumination shown in Fig. 4.3 that was used to create the simulated data.
Having verified the accuracy of the output of prescription retrieval above, we next consider the speed of the prescription retrieval using RMAD gradients. The above optimizations required just 80 minutes of computation time using the resources of 1 core of an Intel i7-4900MQ processor operating at 2.8 GHz. The computation of the error metric required 10.4 seconds while computing the gradient using the RMAD method took 5.0 seconds. If we assume this time is proportional to a some imprecise notion of computational complexity, this indicates that the gradient is about 0.5 times as complex to calculate as the error metric itself. Consequently, the RMAD method has the speed advantage over computing the gradient using the method of FD regardless of the number of unknown variables. For instance, using the number of iterations and the number of unknowns varying per stage listed above, we project that a similar optimization involving gradients using single-sided FD [3] would take approximately 551 days. This is because the method of FD must calculate the error metric at least once for every unknown parameter being optimized, during each iteration of the metric. In contrast, calculating the RMAD gradient for all unknowns required just an additional 5.0 seconds, provided the error metric was also being evaluated. In the context of the overall algorithm, the makes prescription retrieval with the RMAD gradient method more than 9,918 times faster, for this example. For reference, we evaluated the time to compute a single evaluation of the RMAD gradient for various combinations of unknowns and listed them in Table 4.4.

The 0.5 ratio of gradient time to error metric time is interesting because our earlier simulations, which only involved spherical surfaces, found that that gradient took
Table 4.4: Listing of time to compute gradient with respect to various combinations of prescription parameters. Values with * are extrapolated from the time it took to evaluate the error metric

<table>
<thead>
<tr>
<th>Parameters Varying</th>
<th># of Unknowns</th>
<th>RMAD time (s)</th>
<th>FD time (s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$a_1$ only</td>
<td>54</td>
<td>4.11</td>
<td>561.6</td>
</tr>
<tr>
<td>$a_1$ and $a_2$ only</td>
<td>84</td>
<td>4.15</td>
<td>872.4</td>
</tr>
<tr>
<td>$a_1$, $a_2$ and $a_7$ only</td>
<td>117</td>
<td>4.25</td>
<td>1216*</td>
</tr>
<tr>
<td>$a_1$, $a_2$, $a_7$, and stop amplitude</td>
<td>23,226</td>
<td>5.02</td>
<td>$2.4 \times 10^5*$</td>
</tr>
</tbody>
</table>

1.4 times as long as the error metric itself [79]. The difference can be explained by the considerable expense of the repeated evaluations of the sag formula in Line 4 of Algorithm 4.1 as required for freeform surfaces. This makes evaluation of the error metric for systems involving freeform surfaces particularly time consuming. The gradient of the freeform ray intersection utilizes the implicit function theorem to avoid a similar expense in the gradient computation. Consequently, the RMAD gradient is particularly efficient for prescriptions involving aspheric or freeform surfaces. The benefit to systems involving all-spherical surfaces is still considerable but not as great. If we assume that execution time is a good indication of computational complexity, both the 0.5 ratio found for the freeform example and the 1.4 ratio found for an all-spherical system are consistent with the cheap gradient theorem discussed in Section 4.1 which indicates this factor should be at most 5. Finally, as shown in Table 4.4, the execution time of the RMAD gradient also grows very slowly as the number of unknowns increases. For example, switching from estimating just the primary Zernike
terms to estimating all surface sag terms and the stop amplitudes increases the gradient computation time by just 22% although the number of unknowns has gone up 430 times. In comparison, the time to compute the FD gradient grows 430 times because the effort required for FD is linear in the number of unknowns.

### 4.6 Conclusion

Prescription retrieval is a process that estimates design parameters found in an optical prescription from measured point spread function (PSF) data. Due to the nonlinear connection between prescription parameters and the data, this search often involves nonlinear optimization. Minimization of an error metric can be used to search for values of the prescription parameters that predict PSFs consistent with the measured PSF data. Such searches are greatly aided by knowledge of the gradient of the error metric with respect to the unknown prescription parameters. If the method of finite differences (FD) is used, the time to find the gradient scales proportionally with the number of unknown parameters, the number of rays to trace for each PSF and the number of PSFs. For large prescription retrieval problems involving many PSFs, many rays, and/or many unknown parameters, calculating the gradient can be very computationally expensive.

In Section 4.2, we addressed this challenge by applying the reverse mode of algorithmic differentiation (RMAD) [4] to the sequential raytrace computation. Due to the cheap gradient theorem [2, 3], the complexity of the RMAD gradient is always going to be less than five times that of evaluating the error metric itself. Since the method of
FD can quickly surpass this upper bound when there are many unknowns, an RMAD gradient is guaranteed to be more optimal in these circumstances [3]. Further, our experiments showed that the RMAD gradient requires considerably less time than five times that of the error metric. Thus, the actual time efficiency is significantly better than is guaranteed by the cheap derivative theorem, as summarized below.

We structured the raytrace gradient calculation so that arbitrary combinations of raytrace outputs, like optical path length or angle of incidence, evaluated at arbitrary surfaces could be used in later computational models contributing to an error metric. This general result can be applied to accelerate computation of gradients of an arbitrary complicated error metric such as a merit function used in lens design. If the later models and error metric have RMAD steps, this yields a fast procedure for finding the gradient of the error metric with respect to prescription parameters.

Particular prescription parameters we found gradient expressions for include the unknown thickness, rotation, and translation of an arbitrary surface. For spherical surfaces, we found the gradient expression for the surface curvature. We also found gradient steps for conic-section aspheres and freeform surfaces composed of Cartesian \( x \) and \( y \) polynomials. The asphere and freeform surfaces require an iterative ray intersection algorithm and thus benefit from gradient steps utilizing the implicit function theorem. These results can be applied to other parameterizations of surfaces simply by exchanging the computations in Section 4.2.3.1 for those of a different surface type.

In Section 4.3, we developed a model for the field on the image space reference sphere suitable for our prescription retrieval technique. This model used bilinear inter-
polation to formulate the reference sphere amplitude in terms of the amplitude in a stop plane. Also, we found discrete diffraction integrals having gradients with respect to the coordinates of each ray on the reference sphere. This was necessary to use those integrals for PSF intensities for our formulation of prescription retrieval. In Section 4.4, these PSF intensities were included in a bias and gain invariant metric suitable for minimization during a prescription retrieval computation.

In Section 4.5, we demonstrated this prescription retrieval technique to find multiple unknown freeform surface departures of a simulated telescope system. Also, the illumination of the stop was unknown and this amplitude was estimated from the PSF data. This large prescription retrieval problem involved 23,226 unknown parameters estimated from the intensity data of 30 PSFs. By exploiting the RMAD gradient method, we were able to solve this prescription retrieval challenge in 80 minutes using 1 core of an Intel i7-4900MQ processor operating at 2.8 GHz. For comparison, we predict that it would take 550 days to solve the same problem using gradients calculated by the method of finite differences. The RMAD method is particularly useful for systems involving freeform surfaces since, due to the time required by the iterative ray-intersection method, evaluation of the error metric is a lengthy process compared the RMAD gradient. In this example, the ratio of time that it took to evaluate the RMAD gradient to the time that it took to evaluate the error metric was 0.5. To the extent that a ratio in the computation times approximates a ratio of the computational complexities, 0.5 is far less than the upper bound of 5.0 guaranteed by the cheap gradient theorem. The RMAD raytrace may also have applications in lens design
for accelerating gradients of merit functions although the differential raytrace concept contributes a similar benefit.

This chapter has not proved when prescription retrieval is well-founded or solved the stagnation challenges associated with prescription retrieval, but the RMAD method greatly mitigates the local search problem outlined in Section 4.1. It is particularly useful when there are many unknown prescription parameters and/or when many rays need to be traced to predict the measured data. We hope that it enables solutions to problems previously considered intractable due to computational expense.

References


Conclusion and Future Work

5.1 Key Contributions

This thesis makes substantial contributions to several areas of optical metrology through phase retrieval and optical modeling. We divide these contributions into a list of general computational methods and a list of algorithms pertaining to specific applications of phase retrieval or prescription retrieval.

5.1.1 Computational Methods

We developed several generally-applicable computational methods which make use of the reverse mode of automated differentiation (RMAD) [1] for efficient calculation when part of a gradient-based nonlinear optimization algorithm. They are generally useful subcomponents of more complicated solutions to specific problems like those listed in Section 5.1.2.
• **The bilinear resampler in Section 3.7:** We developed the RMAD gradient steps for the bilinear resampling of one array to another. These gradients allow for shape-fitting optimizations, such as solution for the unknowns parameterizing a particular geometric warping. Examples of this include the translation and rotation estimation method introduced in Chapter 3. The resampler and its gradients can be easily adapted to compute more arbitrary geometric transformations between arrays. For example, this capability was used to calculate the reference sphere amplitude model in Section 4.3.1. There it provides the valuable service of linking the exit pupil amplitudes for many exit pupils to a single aperture stop amplitude in a way that is geometrically faithful to the underlying raytrace.

• **The parameterized amplitude model in Section 3.2:** In cases where only one or two near-focus PSFs are available, it may be necessary to regularize the amplitude of retrieved pupils for successful phase and amplitude retrieval [2, 3]. We showed that parameterizations using Zernike polynomials can be used to regularize the amplitude retrieved during phase retrieval when used with a thresholding operation to prevent negative amplitude. The technique works very well with smoothly-varying pupil amplitudes [4] such as mild vignetting. We also showed in laboratory experiments that it can model a hard-edged change in illumination. In comparison, earlier techniques like [2, 3] assume and only retrieve solutions with hard-edged transitions and may yield undesirable artifacts when retrieving amplitudes that actually have smoothly-varying or soft-edged changes. Further, since the parameterized model admits analytic gradients, it easily integrates into simultaneous estimation.
involving many different types of variables; one does not need to stop the computation to perform a separate amplitude regularization process as with the methods in Refs. [2, 3].

- **Analytic gradients for the scale of the matrix multiply discrete Fourier transform (MMDFT) in Section 3.2:** Earlier phase retrieval work has utilized the MMDFT for computing discrete Fourier transforms with arbitrary uniform sample spacing [5, 6], and we showed analytic gradients with respect to the scale parameter. Thus, the MMDFT joins the chirp-Z transform (CZTDFT) [7] as a discrete Fourier transform computation with a smoothly-varying scale for which analytic gradients exist. This greatly aides estimation of this scale parameter when it is unknown because one does not have to perform a grid search operation as in [5]. The MMDFT complements the CZTDFT as the MMDFT is faster to compute than the CZTDFT for the small array sizes often used in this thesis [6]. Both Chapters 3 and 4 made extensive use of this scale gradient for the MMDFT.

- **Gradients for a sequential raytrace in Section 4.2:** Differentiation of raytraces has been used for tolerancing optical designs [8, 9], and accelerating particular portions of merit function calculation in lens design [10–12]. Though some attention has been paid to differentiating entire error metrics [13, 14], a general procedure for efficient calculation of gradients of an arbitrary error metric has not been discussed previously. Consequently, in optics, the gradients of error metrics have most often been found by the method of finite differencing (FD) with analytic gradient formula relegated to calculating just a portion of the merit function as in [12]. The compu-
tational cost of FD grows linearly in the number of rays traced and the number of unknown parameters; it becomes time consuming for very large optimization problems. We applied RMAD to find a general procedure for efficiently calculating the gradient of an arbitrary error metric with respect to its dependence on parameters of an optical prescription. Derivatives were found with respect to surface decentration, rotation, thickness and the curvature of spherical surfaces. We also found gradients for the coefficients of Cartesian polynomial-described freeform surfaces such as Zernike or Forbes polynomial surfaces, and also the gradient term for the conic constant describing a classic asphere. We outlined how an arbitrarily parametrized surface could be integrated into the computation with the minimal effort of finding the RMAD steps for the sag and spatial derivatives. The RMAD raytrace procedure addresses the problem of efficiently finding the gradient of an arbitrary scalar error metric or merit function, particularly those for which classic differential raytrace formulii like [11, 15–17] have not already been found. Since the computational cost in time of the generalized raytrace gradient procedure is mostly independent of the number of unknowns, this work greatly reduces the time it takes to calculate derivatives that would otherwise of been computed by FD. Consequently, it has broad application in accelerating optimizations in lens design and prescription retrieval that would otherwise be computationally costly to calculate.
5.1.2 Advances in Theory and Algorithms

We solved significant theoretical and practical challenges in the areas of transverse-translation diversity phase retrieval (TTDPR) and prescription retrieval. These were the following:

- **A Cramer-Rao lower bound for subaperture translation estimation in ptychography applied to optical metrology and/or TTDPR in Chapter 2:** Sometimes the translation of the subaperture or probe translation must be estimated solely from the measured PSF or intensity data with assistance from fiducials in the pupil amplitude. In these cases, it is critical to understand from basic information theory when such estimation is guaranteed to be inaccurate or ill-posed. I found that estimation may be particularly difficult when the motion of the target being imaged must also be estimated from the PSF data and the wavefront error lacks significant aberrations higher than 2nd radial order when described using Zernike polynomials. Previously this was not appreciated, and translation estimation would be utilized in TTDPR without consideration for whether it was capable of useful results.

- **The unknown transverse-translation diversity phase retrieval (UTTDPR) algorithm described in Chapter 3:** In cases where subaperture translation data is unavailable or untrusted, the UTTDPR algorithm infers such translation information from the PSF data. Previous ptychography and TTDPR algorithms [18–23] have required approximate translation information be input to the algorithm for use as starting estimates. In contrast, UTDDPR requires no information about the direction of translation of the subaperture from one PSF to the next. The UTTDPR algorithm
also estimates rotation of the subaperture, the pupil illumination, and even the motion of the target being imaged. Additionally, it estimates the scaling associated with the propagation from pupil to PSF, and the detector bias and gain terms. The above capabilities make UTTDPR an extremely flexible form of ad hoc metrology since the software does not need to be instructed of the most important details of the test.

• The prescription retrieval technique in Chapter 4: Prescription retrieval involving many PSFs or unknowns can require long computation times due to calculation of the gradient of the error metric using FD. We exploited the general raytrace gradient, the MMDFT, and the bilinear resampler to develop a new error metric for prescription retrieval with computationally fast RMAD gradients. This gradient took just half the time of the error metric itself, which is far better than the time required by FD which is linear in the number of unknowns. We demonstrated this prescription retrieval metric on a problem having 23,226 total unknowns that would take 550 days to calculate using FD; instead, minimization was accomplished in 80 minutes. Further, this time increases only very slightly as the number of unknowns is increased. We think that the ability to perform optimizations of prescription retrieval metrics involving large numbers of unknowns and large amounts of PSF data will enable entirely new forms of optical metrology that were previously too time-consuming to consider.
5.2 Future Work

Although this thesis has made major progress on several computational tools for optical metrology, there remains more to research and consider.

5.2.1 Bounds on Estimation of Subaperture Translation

In Chapter 2, we found Cramer-Rao lower bounds for translation estimation in the absence of amplitude fiducials. It may be possible to construct a theory including the beneficial effects of amplitude fiducials. Also, we ignored the effect of translation error in the estimation of the overall shared-phase aberration coefficients $a_j$, and it would be more realistic to incorporate such errors. A theory that estimated the actual errors in shared phase given the joint relationship between phase and translation would be even more valuable than the results of Chapter 2. Similarly, the model for errors in single-plane phase retrieval employed in Chapter 2 assumed unbiased and normally-distributed errors with no covariance. This simplification is necessary to the theory outlined but it could be further justified or replaced with a more accurate statistical model for the errors of phase retrieval.

5.2.2 Unknown Transverse-Translation Diversity Phase Retrieval

As with many problems in phase retrieval, there is no guarantee that the UTTDPR algorithm will converge to a global minimum when jointly estimating subaperture translation, exit pupil phase and exit pupil amplitude. The results in Chapter 2 can be used to predict when subaperture translation is possible for any unbiased algorithm,
but there is not a statement describing when UTTPR is likely to succeed. Future work
could characterize the UTTDPR algorithm’s failure modes and attempt to provide
more guidance on its behavior.

Also, it might be possible to solve the nonlinear equation, Eq. (2.17), for trans-
lation \( q \) and \( r \) given a known overall shared phase aberration. This would be a
novel way of retrieving subaperture translation since it would work entirely with
Zernike coefficients retrieved from individual PSFs. There would then be no need
for optimization involving multiple PSFs in a joint solution to estimate translation,
and this could be potentially useful for UTTDPR.

### 5.2.3 Gradients of a Sequential Raytrace

The RMAD raytrace gradient of Section 4.2 should be carefully compared with modern
works in differential raytracing [16, 17] to see if there are areas where the RMAD
raytrace is deficient in either generality or computational efficiency. For instance, it is
likely possible to simplify the refraction gradient of Section 4.2.5 using the implicit
function theorem.

Further, it may be useful to add the classic differential raytrace variables to the
output of the raytracer and then use RMAD to find gradients of an error metric with
respect to the differential raytrace output. This would allow computations like the
5.2.4 Efficient Prescription Retrieval

We did not attempt to solve the basic stagnation issue described in Section 4.1. Like global optimization in lens design, this is a significant challenge for future work. It would also be very useful to find general principles that would guide selection of which problems are suitable for prescription retrieval and which are not. The classic literature on alignment of optical systems for manufacturing is very relevant to prescription retrieval.

We also did not prove that solutions found by prescription retrieval will be accurate. For instance, in the example of Section 4.5, we assumed that there was zero decen- tration from one element to the next, no rotation of each surface and that the thicknesses were exactly the nominal amount. In practice, there is always some tolerance in the design, and how much error is acceptable before the results of the model fitting are in error is application specific. In any model-fitting exercise there is also the problem of unknown and/or unmodeled aspects of the physical system affecting the quality of the resulting inferences. Characterizing the effects of modeling error on the retrieved prescription parameters is important future work.

The prescription retrieval technique could also benefit from structural improve- ments. For instance, if aiming of the input rays could be integrated into each iteration, this might allow the use of classic FFT-style uniform spacing in the image reference sphere. Accomplishing this will require differentiating the iterative ray-aiming process which will involve another invocation of the implicit function theorem [25] as in Section 4.2.3.2.
References


