SIAM REVIEW
Vol. 44, No. 2, pp. 169–224

Optical Wavefront Reconstruction: Theory and Numerical Methods*

D. Russell Luke†
James V. Burke‡
Richard G. Lyon§

Abstract. Optical wavefront reconstruction algorithms played a central role in the effort to identify gross manufacturing errors in NASA's Hubble Space Telescope (HST). NASA's success with reconstruction algorithms on the HST has led to an effort to develop software that can aid and in some cases replace complicated, expensive, and error-prone hardware. Among the many applications is HST's replacement, the Next Generation Space Telescope (NGST).

This work details the theory of optical wavefront reconstruction, reviews some numerical methods for this problem, and presents a novel numerical technique that we call extended least squares. We compare the performance of these numerical methods for potential inclusion in prototype NGST optical wavefront reconstruction software. We begin with a tutorial on Rayleigh–Sommerfeld diffraction theory.

Key words. geometric optics, phase retrieval, nonconvex programming, least squares

AMS subject classifications. 78A45, 49N45, 90C90, 93E24

PII. S0036144501390754

1. Introduction.

The history of science is filled with misfortunes that have been transformed into scientific triumphs. This article describes some of the scientific progress in numerical methods for wavefront reconstruction that contributed to the eventual and remarkable successes of NASA's Hubble Space Telescope (HST). Shortly after launch on April 24, 1990, it was discovered that the primary mirror of the HST suffered from a large spherical aberration [21]. Several teams of researchers were dispatched to apply a variety of image processing techniques to the flight data of stellar images in order to identify the aberration and aid in the design of corrective optics. Burrows [20] and Lyon, Miller, and Grusczak [79] applied parametric techniques; Fienup [44] applied gradient-based algorithms; Fienup et al. [45] and Roddier and Roddier [102] applied nonparametric projection techniques; Redding et al. [100]

*Received by the editors June 12, 2001; accepted for publication (in revised form) September 5, 2001; published electronically May 1, 2002.
http://www.siam.org/journals/sirev/44-2/39075.html

†Institute for Numerical and Applied Mathematics, Universität Göttingen, Lotzestr. 16-18, D-37083 Göttingen, Germany (luke@math.uni-goettingen.de). This author's work was supported by NASA grant NGT5-66.

‡Department of Mathematics, University of Washington, Seattle, WA 98195 (burke@math.washington.edu). This author's work was supported by NSF grant DMS-9971852.

§NASA/Goddard Space Flight Center, Greenbelt, MD 20771 (lyon@jansky.gsfc.nasa.gov).

and Meinel, Meinel, and Schulte [82] applied ray tracing and diffraction propagation techniques; and Barrett and Sandler [8] applied neural network techniques. The results of all groups were used in conjunction with archival HST manufacturing records to pinpoint the size and source of the error. It wasn't until 1993 that corrective optics were installed. In the meantime, researchers continued with efforts to model the telescope with enough precision to recover unaberrated images through postprocessing. In addition to the gross manufacturing errors, researchers were able to identify aberrations due to the polish marks on the primary and secondary mirrors. Again, wavefront reconstruction techniques played an important role in this effort [69]. During this time much was learned about reconstruction algorithms. An important lesson learned from the HST is that relatively simple software can aid and in some cases replace complicated and sensitive optical systems. In 2012 the replacement for the HST, the Next Generation Space Telescope (NGST), will be folded into the nose of a rocket and launched into geosynchronous orbit, far beyond the reach of astronauts. Wavefront reconstruction algorithms will play a central role in maintaining alignment on the NGST [81, 106].

Optical wavefront reconstruction is an inverse problem that arises in many applications in physics and engineering. Numerical algorithms for solving this problem have been employed in crystallography, microscopy, optical design, and adaptive optics for three decades. The history of the problem goes back much further. The celebrated algorithm of Gerchberg and Saxton [48] demonstrated that practical *numerical* solutions to the two-dimensional problems were possible. Since the introduction of the Gerchberg–Saxton algorithm, numerous variations have been studied [16, 31, 111, 36, 43, 45, 72, 78, 85, 88, 97, 128, 5, 90, 116, 49, 24]. The success of these algorithms on the HST together with techniques for *simultaneous* phase retrieval and deconvolution developed for use with land-based astronomical observations [23, 50, 68, 76, 94, 95, 96, 120, 123, 122, 105] has led to the development of software that, in conjunction with simple optical systems, can achieve the same resolution as complicated, expensive, and error-prone optical systems [77, 98, 99, 74, 73, 70].

While computational wavefront reconstruction algorithms have been successfully applied for many years, most of the fundamental mathematical questions about the behavior and properties of projection algorithms and related techniques remain open, in particular questions regarding existence, uniqueness, and convergence. The theoretical results often cited for projection algorithms do not apply to the problem of phase retrieval since the required hypotheses are not satisfied. This work details the theory of projection techniques for the wavefront reconstruction problem and related gradient-based techniques. For ease of discussion, the problem is formulated in the continuum. Results for the discrete case follow easily from these results. The numerical theory for wavefront reconstruction is divided into two different approaches. The first approach, which we call *geometric*, is based on the geometric properties of sets in a Hilbert space and involves the projection onto these sets [16, 32, 43, 48, 72, 85, 131, 129, 31]. The second approach, which we call *analytic*, is based on the analytic properties of smooth objective functions that are to be minimized [6, 7, 35, 55, 54, 65, 71, 117, 43, 49]. Due to the ease with which they are implemented, geometric approaches, known in the optics community as *iterative transform* or *projection* methods, are common. However, convergence is still an open question except in very special cases [31, 25]. Analytic methods, on the other hand, while providing more stable and theoretically sound algorithms, are often complicated and computationally expensive. It is shown below that first-order analytic methods associated with a particular error metric are closely related to iterative transform algorithms. For the wavefront reconstruction problem,

however, analytic methods have several advantages over geometric approaches. The most important of these is robustness. In addition, first-order analytic methods can be readily extended to second-order methods for accelerated convergence.

The similarity between iterative transform algorithms and line search methods applied to a particular error metric has been known for some time [6, 43, 48]. A precise analysis of this correspondence, however, has proven elusive. The source of the difficulty is the nonconvexity of the underlying sets and the nonsmoothness of the error metric. This work details the connection between geometric and analytic methods for the phase retrieval problem. We compare the numerical performance of projection algorithms to standard line search algorithms applied to a perturbed least squares error metric. In addition, we investigate a novel approach, which we call *extended least squares*, together with limited memory and trust region techniques for stabilizing and accelerating first-order analytic algorithms.

2. Literature Review. The problem of wavefront reconstruction is a special case of the more general inverse problem of *phase retrieval*. The phase retrieval problem arises in such diverse fields as microscopy [83, 47, 118, 119, 61, 37], holography [42, 115], crystallography [84], neutron radiography [3], optical design [39], adaptive optics, and astronomy. Earlier reviews of the phase problem can be found in [62, 110]. Millane [84] provided an excellent review of the phase problem in X-ray crystallography. The physical setting is discussed in some detail in the following section. The abstract problem is stated as follows: Given $a : \mathbb{R}^2 \to \mathbb{R}_+$ and $b : \mathbb{R}^2 \to \mathbb{R}_+$, find $u : \mathbb{R}^2 \to \mathbb{C}$ satisfying $|u| = a$ and $|u^\wedge| = b$. Here $\mathbb{R}_+$ denotes the positive orthant, $\cdot^\wedge$ denotes the Fourier transform, and the modulus is the *pointwise Euclidean* magnitude. Simply stated, the problem is to find the phase of a complex-valued function given its pointwise amplitude and the pointwise amplitude of its Fourier transform, hence the name *phase retrieval*.

Until the 1970s the problem of phase retrieval was thought to be hopeless for a number of reasons. In a letter to A. A. Michelson, Lord Rayleigh stated that the continuous phase retrieval problem in interferometry was in general not possible without a priori information on the symmetry of the data [113]. In one dimension it was shown that the discrete problem has a multitude of solutions. Indeed, for a signal that is represented by n terms of the Fourier series expansion, there are as many as 2^{n-1} possible solutions to the problem [1, 2]. Wolf was among the first to suggest that these obstacles might not be insurmountable [126]. Kano and Wolf followed this claim with an analytic reconstruction of the temporal complex coherence function of blackbody radiation [67]. Their reconstruction was not numerical in nature but depended, rather, on the analytic properties of the continuous Fourier transform. Further efforts were made to broaden the applicability of these results [103]. At the same time Walther and O'Neill provided some hope for the possibility of meaningful solutions in the discrete case and, in some relevant cases, uniqueness [91, 125]. As Dialetis and Wolf later pointed out, however, the applicability of the theory for the continuous case was limited [34]. Nevertheless, a number of researchers proposed the addition of constraints to narrow the number of potential solutions for the one-dimensional problem [37, 59, 61, 92, 97, 118, 119, 127].

As early as 1972 a practical algorithm was proposed for numerical solutions to the seemingly more difficult two-dimensional problem. In their famous paper, Gerchberg and Saxton [48], independently of previous mathematical results for projections onto convex sets, proposed a simple algorithm for solving phase retrieval problems in two dimensions. In [72] the algorithm was recognized as a projection algorithm. Projection

algorithms in convex settings have been well understood since the early 1960s [17, 53, 56, 112, 124, 131, 129, 132]. The phase retrieval problem, however, involves *nonconvex* sets. For this reason, the convergence properties of the Gerchberg–Saxton algorithm and its variants are not completely understood.

In the majority of relevant cases numerical experience demonstrated that projection-type algorithms converged to correct solutions [40, 41]. It was suggested in [18] that this seeming robustness of numerical methods is due to the factorability (or lack thereof) of related polynomials. Indeed, the solution to the two-dimensional phase retrieval problem for a discrete signal that can be represented by a finite Fourier series expansion, that is, for a *band-limited* image, if it exists, is almost always unique up to rotations by 180 degrees, linear shifts, and multiplication by a unit magnitude complex constant. The proof and details of this result can be found in [58]. While this result is of fundamental importance, it does not apply to many of the algorithms used for phase retrieval, in particular in the presence of noise. Thus, while the uniqueness result above remains valid for band-limited signals, it says nothing about the uniqueness of *approximate* solutions in the event that a true solution does not exist, that is, when the feasible set is empty. In the convex setting, when the constraint sets onto which the projections are computed do not intersect, convergence of projection algorithms is an open question [10, 12, 9, 30, 53, 130]. Much less is known about the nonconvex setting where many applications lie [28, 29, 24, 107, 60].

In 1982 Fienup [43] generalized the Gerchberg–Saxton algorithm and analyzed many of its properties, showing, in particular, that the directions of the projections in the generalized Gerchberg–Saxton algorithm are formally similar to directions of steepest descent for a squared set distance metric. We show in section 5.1 that this connection to directions of steepest descent is complicated by the fact that the metric is not everywhere differentiable. In 1985 Barakat and Newsam [6, 7] developed an approach similar to the gradient descent analogy suggested in [43]. They modeled their analysis on the projection theory for convex sets. A well-known fact from convex analysis is that the gradient of the squared distance to a convex set is *equivalent* to the direction toward the projection onto the set. To extend this property to the nonconvex sets, Barakat and Newsam required the projection operators to be single-valued; however, there is no known example of a nonconvex set for which the projection operator is single-valued.[1] Indeed, we show that the projections in the case of phase retrieval are multivalued. We show precisely how the multivaluedness of the projections is related to the nonsmoothness of the squared set distance metric.

In section 5.2 a smooth error metric is proposed and bounds are derived for the distance between the gradient of the smooth metric and the directions toward the projections. While projection methods often work well in practice, fundamental mathematical questions concerning their convergence remain unresolved. What are often referred to as convergence results for projection algorithms are statements that the error between iterations will not increase [48, 72]. In general, projection algorithms may not converge to the intersection of nonconvex sets. See [72] and [31] for discussions.

The plan of the paper is as follows. As is true with any inverse problem, great care must be taken in the formulation of the forward problem. In section 3 we derive the mathematical model for diffraction imaging and formulate the inverse problem

[1] The issue of nonuniqueness of the projection operator is not to be confused with the uniqueness of the phase problem. The results of [58] are not affected by the multivaluedness of the projection operators.

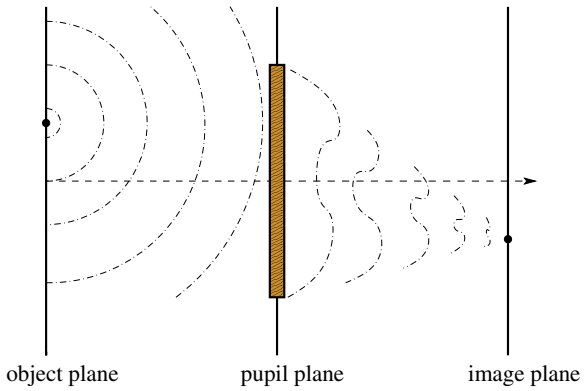

Fig. 1 *Model optical system.*

associated with phase retrieval. In the same section, the abstract optimization problem associated with wavefront reconstruction is formulated. The notation that is used throughout the paper is introduced in section 3.1.6. Readers familiar with this theory can skip to section 4, referring back to sections 3.1.6 and 3.2 for notation. Section 4 details projection algorithms, among which are the well-known iterative transform techniques, reviewed in section 4.2. In section 5 we formulate the specific optimization problem to which iterative transform algorithms are applied. In section 5.1 a least squares measure is formulated. This measure is shown to be nonsmooth, and its relationship to projections is discussed. To avoid theoretical technicalities and numerical instability associated with nonsmoothness, a smooth perturbation to the least squares error metric is proposed in section 5.2. The relation of this perturbed measure to the projections is summarized in Theorem 5.1. In section 5.3 we apply a recent extension to the least squares measure that allows adaptive weighting of the errors between measurements [13]. The convergence of a line search algorithm to first-order optimality conditions for smooth measures is proved in Theorem 6.1 of section 6. The application of limited memory techniques with trust regions is studied in section 6.2. The corresponding algorithm is given by Algorithm 6.2. Numerical results are detailed in section 7.

3. Optical Imaging.

3.1. The Forward Imaging Model. The physical setting we consider here is that of a *monochromatic, time-harmonic electromagnetic field in a homogeneous, isotropic medium with no charges or currents*. This is depicted as a wave propagating away from some source to the left of the *pupil plane* in Figure 1. By Maxwell's equations, at a given frequency $\omega \in \mathbb{R}_+$, the spatial components of the electric and magnetic fields can be represented as the real part of complex-valued functions $U_\omega : \mathbb{R}^3 \to \mathbb{C}$ satisfying the Helmholtz equation describing the spatial distribution of energy in an expanding wave:

$$(\triangle + k^2 n^2) U_\omega(\boldsymbol{x}) = 0. \tag{1}$$

Here $\triangle$ denotes the Laplacian, $n \in \mathbb{R}_+$ is the index of refraction of the medium, and $k \in \mathbb{R}_+$ is the wave number. The wave number is related to the frequency since ω/k is the speed of light. Another quantity that arises is the *wavelength* λ defined as

$\lambda = 2\pi/k$. For convenience, let $n = 1$. In all that follows, the fields are assumed to be monochromatic (i.e., single frequency); thus we drop the ω subscript from U_ω.

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

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

> *light falling on the aperture* [$\mathbb{A}$] *propagates as if every* [*surface*] *element* [dS] *emitted a spherical wave the amplitude and phase of which are given by that of the incident wave* [U] [109].

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

3.1.1. Rayleigh–Sommerfeld Diffraction. We now provide a terse summary of Rayleigh–Sommerfeld diffraction theory. More detailed developments can be found in [52, 15, 109]. Let Ω be a closed volume in $\mathbb{R}^3$ whose boundary is the orientable closed surface $\mathbb{S}$ and let $\vec{n}$ denote the unit *inward* normal to Ω. Let U and $\tilde{U}$ be twice continuously differentiable scalar fields mapping Ω and $\mathbb{S}$. By Green's theorem,[2]

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

[2]Green's theorem is usually stated in terms of the unit *outward* normal. In optics, for the derivation of Rayleigh–Sommerfeld diffraction, the unit inward normal is generally used.

where $\frac{\partial}{\partial \vec{n}}$ denotes the derivative in the direction of the unit inward normal at $\mathbb{S}$. If both U and $\tilde{U}$ satisfy the Helmholtz equation (1), then

$$-\int_{\mathbb{S}} \tilde{U}\frac{\partial U}{\partial \vec{n}} - U\frac{\partial \tilde{U}}{\partial \vec{n}} dS = 0. \tag{2}$$

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

$$G_0(\boldsymbol{x};\boldsymbol{\xi}) = \frac{\exp(ik|\boldsymbol{x}-\boldsymbol{\xi}|)}{|\boldsymbol{x}-\boldsymbol{\xi}|}, \quad \boldsymbol{x} \neq \boldsymbol{\xi}, \tag{3}$$

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

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

Thus, as in (2),

$$-\int_{\mathbb{S}+\mathbb{S}_\epsilon} \frac{\exp(ik|\boldsymbol{x}-\boldsymbol{\xi}|)}{|\boldsymbol{x}-\boldsymbol{\xi}|}\frac{\partial U}{\partial \vec{n}} - U\frac{\partial}{\partial \vec{n}}\frac{\exp(ik|\boldsymbol{x}-\boldsymbol{\xi}|)}{|\boldsymbol{x}-\boldsymbol{\xi}|} dS = 0. \tag{4}$$

The integral theorem of Helmholtz and Kirchhoff [15, 109] uses (4) to establish the identity

$$\begin{aligned} U(\boldsymbol{\xi}) &= \lim_{\epsilon\to 0} \frac{-1}{4\pi}\int_{\mathbb{S}_\epsilon} \frac{\exp(ik|\boldsymbol{x}-\boldsymbol{\xi}|)}{|\boldsymbol{x}-\boldsymbol{\xi}|}\frac{\partial U}{\partial \vec{n}} - U\frac{\partial}{\partial \vec{n}}\frac{\exp(ik|\boldsymbol{x}-\boldsymbol{\xi}|)}{|\boldsymbol{x}-\boldsymbol{\xi}|} dS \\ &= \frac{1}{4\pi}\int_{\mathbb{S}} \frac{\exp(ik|\boldsymbol{x}-\boldsymbol{\xi}|)}{|\boldsymbol{x}-\boldsymbol{\xi}|}\frac{\partial U}{\partial \vec{n}} - U\frac{\partial}{\partial \vec{n}}\frac{\exp(ik|\boldsymbol{x}-\boldsymbol{\xi}|)}{|\boldsymbol{x}-\boldsymbol{\xi}|} dS. \end{aligned} \tag{5}$$

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

Rayleigh–Sommerfeld diffraction theory is derived by considering a specific volume Ω and surface $\mathbb{S}$ (see Figure 2) together with a particular Green's function G. Let the surface $\mathbb{S}$ be the arbitrarily large half-sphere composed of the hemisphere $\mathbb{E}$ and the disk $\mathbb{D}$ contained in the plane $\mathbb{T}$. The disk $\mathbb{D}$ consists of an annulus $\mathbb{A}'$ with a small opening $\mathbb{A}$. Let $\boldsymbol{x}'$ be an element of the open half-space determined by the plane $\mathbb{T}$ and having empty intersection with Ω. Let $\mathbb{I} \subset \Omega$ be a screen parallel to $\mathbb{T}$ and whose distance from $\mathbb{T}$ equals that of $\boldsymbol{x}'$ to $\mathbb{T}$. The problem is to determine the field U on $\mathbb{I}$ under the assumption that the field propagates only through $\mathbb{A}$.

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

$$G(\boldsymbol{x};\boldsymbol{x}',\boldsymbol{\xi}) \equiv G_0(\boldsymbol{x};\boldsymbol{x}') - G_0(\boldsymbol{x};\boldsymbol{\xi}), \tag{6}$$

where G_0 is defined in (3) and $|\boldsymbol{x}-\boldsymbol{x}'| = |\boldsymbol{x}-\boldsymbol{\xi}|$ for all $\boldsymbol{x} \in \mathbb{T}$. The field G is the Green's function for a half-space with Dirichlet boundary conditions; that is, it satisfies the following conditions:

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

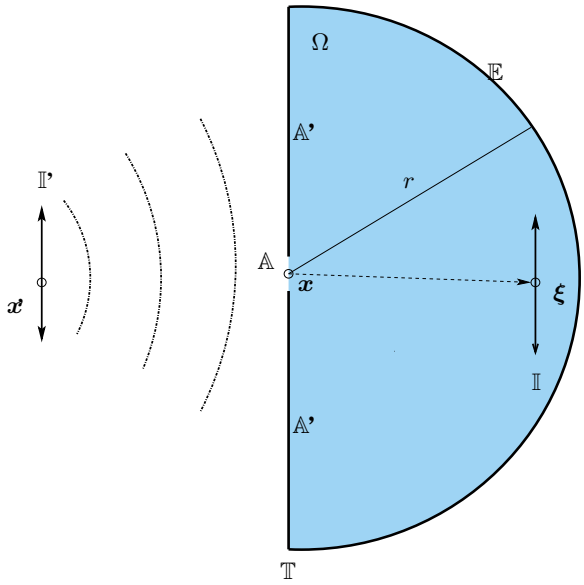

Fig. 2 *Rayleigh–Sommerfeld diffraction.*

The field G satisfies the conditions required for substitution into (5) in place of G_0, yielding

$$U(\boldsymbol{\xi}) = \frac{1}{4\pi}\int_{\mathbb{S}} G\frac{\partial U}{\partial \vec{n}} - U\frac{\partial G}{\partial \vec{n}}dS. \tag{7}$$

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

$$U = 0 \text{ on } \mathbb{A}'; \tag{8}$$

$$|\boldsymbol{x} - \boldsymbol{\xi}|\left(\frac{\partial U}{\partial \vec{n}} - ikU\right) \to 0 \text{ as } |\boldsymbol{x} - \boldsymbol{\xi}| \to \infty. \tag{9}$$

Condition (8) states that the screen is a nearly "perfect conductor"; (9) is the *Rayleigh–Sommerfeld radiation condition*. In the limit as the radius of the hemisphere $\mathbb{E}$ goes to infinity, (6) and (7), together with the radiation conditions and (8), yield

$$U(\boldsymbol{\xi}) = \frac{1}{4\pi}\int_{\mathbb{A}} -U\frac{\partial G}{\partial \vec{n}}dS. \tag{10}$$

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

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

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

$$\begin{aligned}\frac{\partial G}{\partial \vec{n}} &= 2\frac{\exp(ik|\boldsymbol{\xi} - \boldsymbol{x}|)}{|\boldsymbol{\xi} - \boldsymbol{x}|}\left(ik - \frac{1}{|\boldsymbol{\xi} - \boldsymbol{x}|}\right)\alpha(\vec{n}, \boldsymbol{\xi} - \boldsymbol{x}) \\ &\approx 2ki\frac{\exp(ik|\boldsymbol{\xi} - \boldsymbol{x}|)}{|\boldsymbol{\xi} - \boldsymbol{x}|}\alpha(\vec{n}, \boldsymbol{\xi} - \boldsymbol{x}). \end{aligned} \tag{11}$$

Substituting (11) into (10) yields the following mathematical formulation of Huygens's principle:

$$U(\boldsymbol{\xi}) \approx \int_{\mathbb{A}} U(\boldsymbol{x}) h(\boldsymbol{\xi}; \boldsymbol{x}) dS, \tag{12}$$

where

$$h(\boldsymbol{\xi}; \boldsymbol{x}) \equiv \frac{\exp(ik|\boldsymbol{\xi} - \boldsymbol{x}|)}{i\lambda|\boldsymbol{\xi} - \boldsymbol{x}|} \alpha(\vec{n}, \boldsymbol{\xi} - \boldsymbol{x})$$

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

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

$$|\boldsymbol{\xi} - \boldsymbol{x}| \approx \xi_3 \left[1 + \frac{1}{2\xi_3^2}(\xi_1 - x_1)^2 + \frac{1}{2\xi_3^2}(\xi_2 - x_2)^2\right]. \tag{13}$$

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

$$h_{Fre}(\boldsymbol{\xi}; \boldsymbol{x}) = \frac{\exp(ik\xi_3)}{i\lambda\xi_3} \exp\left(\frac{ik}{2\xi_3}\left((\xi_1 - x_1)^2 + (\xi_2 - x_2)^2\right)\right). \tag{14}$$

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

$$\left[\frac{\partial}{\partial \xi_3} - \frac{i}{2k}\triangle_t - ik\right] h_{Fre} = 0, \tag{15}$$

where $\triangle_t$ is the Laplacian in the $\xi_1\xi_2$-plane, i.e., $\triangle_t = \frac{\partial^2}{\partial \xi_1^2} + \frac{\partial^2}{\partial \xi_2^2}$. By substituting h_{Fre} into (12), we obtain the Fresnel diffraction field

$$U_{Fre}(\boldsymbol{\xi}) = \int_{\mathbb{A}} U(\boldsymbol{x}) h_{Fre}(\boldsymbol{\xi}; \boldsymbol{x})\, dx_1\, dx_2. \tag{16}$$

This field also satisfies (15).

If the aperture is small compared to the image ($x_1, x_2 \ll \xi_1, \xi_2$, as is the case in diffraction imaging), one can expand the quadratic in the Fresnel kernel (14) and neglect quadratic terms in x_1 and x_2:

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

With this approximation, (14) reduces to

$$h_{Fra}(\boldsymbol{\xi}; \boldsymbol{x}) = \frac{\exp(ik\xi_3)}{i\lambda\xi_3} \exp\left(\frac{ik}{2\xi_3}(\xi_1^2 + \xi_2^2)\right) \exp\left(\frac{ik}{\xi_3}(x_1\xi_1 + x_2\xi_2)\right). \tag{17}$$

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

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

$$U_{Fra}(\boldsymbol{\xi}) = \int_{\mathbb{A}} U(\boldsymbol{x}) h_{Fra}(\boldsymbol{\xi}; \boldsymbol{x})\, dx_1\, dx_2. \tag{18}$$

Close examination of (18) reveals a relationship between the Fraunhofer transform and the Fourier transform. For $u : \mathbb{R}^n \to \mathbb{C}$, let $\wedge$ denote the Fourier transform defined by[3]

$$u^{\wedge}(\boldsymbol{\xi}) \equiv \int_{\mathbb{R}^n} u(\boldsymbol{x}) \exp(-2\pi i \boldsymbol{x} \cdot \boldsymbol{\xi})\, d\boldsymbol{x}. \tag{19}$$

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

$$\mathcal{X}_{\mathbb{A}}(\boldsymbol{x}) \equiv \begin{cases} 1 & \text{for } \boldsymbol{x} \in \mathbb{A}, \\ 0 & \text{for } \boldsymbol{x} \notin \mathbb{A}. \end{cases} \tag{20}$$

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

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

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

$$C(\boldsymbol{\xi}) = \frac{\exp(ik\xi_3)}{i\lambda\xi_3} \exp\left(\frac{ik}{2\xi_3} (\xi_1^2 + \xi_2^2) \right).$$

3.1.2. Diffraction Imaging with a Lens. Based on these integral approximations to the field U on the image plane $\mathbb{I}$, we now derive the associated Green's function of the optical system with a lens. We begin with a brief discussion motivating the mathematical model for a thin lens using the paraxial approximation [15, Chap. 4], [52, Chap. 5].

A lens is modeled from a geometric optics perspective. Under this interpretation a wave propagates along rays orthogonal to its level surface, or in mathematical parlance, along the characteristics of the Helmholtz equation (1). The phase, θ of the complex phasor representation,[4] of a wave describes the geometric shape of the level surface, and thus the orientation of the rays along which the wave travels. A lens is a (thin) piece of glass or some other transparent material with a different index of refraction (depending on the wave number k) than that of the surrounding medium. Physically a lens changes the path [15, Chap. 3] of the wave without altering its amplitude; that is, it changes the geometric path of propagation. This is modeled as a change in the direction of the rays, or equivalently a change in the phase θ of the wave across the lens.

For instance, the direction of propagation of the wave described by the Fresnel kernel h_{Fre} (14) is parabolic with axis of symmetry along the ξ_3 axis.[5] Suppose

[3] Note that this definition is valid only for functions in $L^1 \cap L^2$. In section 3.2 we use the extension of this transform to functions on L^2, the Fourier–Plancherel transform.

[4] The phase function $\theta : \mathbb{R}^3 \to \mathbb{R}$, sometimes called the *eikonal*, satisfies the eikonal equation [15, Chap. 3].

[5] The amplitude of the wave is constant in the $x_1 x_2$-plane.

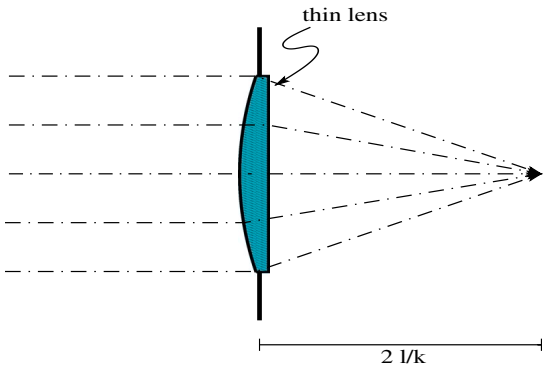

Fig. 3 *Lens model.*

we place a lens, shown in Figure 3, at the pupil plane of our model optical system (Figure 1) with axis of symmetry in the x_3 direction centered at $x_1 = x_2 = 0$. Suppose further that this lens is designed to change the direction of propagation of the field in a parabolic fashion across the aperture $\mathbb{A}$. In the complex phasor representation of the wave, this physical effect is modeled by the addition of a complex phase term on the support of the lens. We represent such a lens by the function

$$\phi(\boldsymbol{x}) = \exp\left(\mathcal{X}_{\mathbb{A}}(\boldsymbol{x})\frac{-ik}{2l}(x_1^2 + x_2^2)\right), \tag{21}$$

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

The Fraunhofer approximation (17) to the Fresnel kernel (14) also arises in models of optical systems with a lens. To see this, consider a wavefront of the form h_{Fre} at the x_1x_2-plane. The field immediately after the lens is given by multiplying h_{Fre} by the lens (21). If $l = \xi_3$, this multiplication yields the identity $\phi h_{Fre} = h_{Fra}$.

We now detail Kirchhoff's diffraction theory for the following imaging model, based on Huygens's principle (12), with diffraction kernel h_{Fre} and a lens of the form (21):

$$U(\boldsymbol{\xi}) \approx \int_{\mathbb{A}} U(\boldsymbol{x})\phi(\boldsymbol{x})h_{Fre}(\boldsymbol{\xi};\boldsymbol{x})d\boldsymbol{x}. \tag{22}$$

The derivation of (12) requires the conditions (8) and (9), where U satisfies (1). Here we encounter the difficulty that we have not specified any boundary conditions on the region $\mathbb{A}$, without which we cannot obtain an explicit approximation for U at $\boldsymbol{\xi}$. Kirchhoff's diffraction theory is based on the conditions (8) and (9) together with the additional boundary condition

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

where G_0 is given by (3). Since $\boldsymbol{x}'$ enters as a parameter on the right-hand side, we write the field U on $\mathbb{A}$ satisfying the above equation as

$$U(\boldsymbol{x};\boldsymbol{x}') = G_0(\boldsymbol{x};\boldsymbol{x}') \quad \text{for } \boldsymbol{x}' \notin \Omega \text{ and } \boldsymbol{x} \in \mathbb{A}. \tag{23}$$

Similarly, we write $U(\boldsymbol{\xi}) = U(\boldsymbol{\xi}; \boldsymbol{x}')$ to indicate that the field U on $\mathbb{I}$ is also parameterized by the location of the point source $\boldsymbol{x}'$. Conditions (8) and (23) are called *Kirchhoff's boundary conditions*.

The function satisfying (1), (8)–(9), and (23) is very special indeed. For most applications, however, it is sufficient to approximate the field U by the field that would result from a point source at $\boldsymbol{x}'$ in the absence of the screen $\mathbb{S}$, that is, $U(\cdot; \boldsymbol{x}') \approx G_0(\cdot; \boldsymbol{x}')$ everywhere to the left of the screen *except* on $\mathbb{A}'$, where $U(\cdot; \boldsymbol{x}') = 0$. The justification of such an approximation is beyond the scope of this work. There is a vast classical literature surrounding this problem. Interested readers are referred to [15, Chap. 11] and references therein.

Assume next that $\boldsymbol{x}'$ satisfies $|(x_1', x_2', 0)| \ll x_3'$, where $x_3' = \operatorname{dist}(\boldsymbol{x}', \mathbb{T})$. Then, as in the derivation of the Fresnel kernel (14), the field at any point $\boldsymbol{x} \in \mathbb{A}$ can be approximated by

$$U(\boldsymbol{x}; \boldsymbol{x}') \approx \frac{\exp(ikx_3')}{i\lambda x_3'} \exp\left(\frac{ik}{2x_3'}\left((x_1 - x_1')^2 + (x_2 - x_2')^2\right)\right). \tag{24}$$

Substituting (24) into (22) with the lens (21) yields

$$\begin{aligned} U(\boldsymbol{\xi}; \boldsymbol{x}') \equiv{}& \frac{\exp(ik(x_3' + \xi_3))}{-\lambda^2 x_3' \xi_3} \tilde{C}(\boldsymbol{\xi}) \tilde{C}(\boldsymbol{x}') \\ &\times \iint_{-\infty}^{\infty} \mathcal{X}_{\mathbb{A}}(\boldsymbol{x}) \exp\left(\frac{ik}{2}(1/x_3' + 1/\xi_3 - 1/l)(x_1^2 + x_2^2)\right) \\ &\qquad \times \exp\left(\frac{-2\pi i}{\lambda x_3' \xi_3}(x_1' \xi_3 + \xi_1 x_3',\ x_2' \xi_3 + \xi_2 x_3') \cdot (x_1, x_2)\right) dx_1 dx_2. \end{aligned} \tag{25}$$

Here $\tilde{C}(\boldsymbol{\xi}) = \exp\left(\frac{ik}{2\xi_3}(\xi_1^2 + \xi_2^2)\right)$, and likewise for $\tilde{C}(\boldsymbol{x}')$. When the *lens law* [52, (5)–(30)] is satisfied, that is, when

$$1/x_3' + 1/\xi_3 - 1/l = 0, \tag{26}$$

then the rays along which the light wave travels depend only linearly on the coordinates in the aperture $\mathbb{A}$. The field at $\mathbb{I}$ is said to be *in focus* when the lens law is satisfied, since this plane (where the receptors lie) coincides with the level surface of the wave.[6] We consider only those points $(\xi_1, \xi_2, \xi_3) \in \mathbb{I}$ and $(x_1', x_2', x_3') \in \mathbb{I}'$ for which

$$\xi_3 \gg \frac{k(\xi_1^2 + \xi_2^2)}{2} \quad \text{and} \quad x_3' \gg \frac{k({x_1'}^2 + {x_2'}^2)}{2}, \tag{27}$$

where $\mathbb{I}$ and $\mathbb{I}'$ are the planes depicted in Figure 2. Then, as with the Fraunhofer approximation, the $\tilde{C}(\cdot)$ factors are nearly unity. Thus, when (27) and the lens law (26) hold,

$$\begin{aligned} U(\boldsymbol{\xi}; \boldsymbol{x}') \approx{}& \frac{-\exp(ik(x_3' + \xi_3))}{\lambda^2 x_3' \xi_3} \\ &\times \iint_{-\infty}^{\infty} \mathcal{X}_{\mathbb{A}}(\boldsymbol{x}) \exp\left(\frac{-2\pi i}{\lambda x_3' \xi_3}(\xi_3 x_1' + x_3' \xi_1,\ \xi_3 x_2' + x_3' \xi_2) \cdot (x_1, x_2)\right) dx_1 dx_2. \end{aligned} \tag{28}$$

[6] Note that the lens law depends entirely on the parabolic approximation to the incident wavefront given by (24).

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

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

to obtain the following Fourier transform representation,

$$\begin{aligned} U(\boldsymbol{\xi};\boldsymbol{x}') &\approx c\iint_{-\infty}^{\infty} \mathcal{X}_{\mathbb{A}}(\boldsymbol{x}) \exp\left(\frac{-2\pi i}{\lambda\xi_3}(\hat{\xi}_1,\hat{\xi}_2)\cdot(x_1,x_2)\right) dx_1 dx_2 \\ &= c\mathcal{X}_{\mathbb{A}}^{\wedge_T}\left(\frac{\hat{\boldsymbol{\xi}}}{\lambda\xi_3}\right), \end{aligned} \tag{29}$$

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

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

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

If every point in the support of the source in the object plane satisfies $|(x_1', x_2', 0)| \ll x_3'$, as we have been assuming all along, then we can approximate the dependence of $U(\boldsymbol{\xi};\boldsymbol{x}')$ on $\boldsymbol{x}'$ by $U(\boldsymbol{\xi};\boldsymbol{x}') \approx U(\boldsymbol{\xi}) = U(\hat{\boldsymbol{\xi}} - \hat{\boldsymbol{x}})$. This approximation implies that the system's response to a point source $U(\boldsymbol{\xi};\boldsymbol{x}')$ remains invariant under translation of the source in the $x_1'x_2'$-plane.[7] The superposition is thus represented by the two-dimensional convolution, denoted $*$:

$$\begin{aligned} \psi(\hat{\boldsymbol{\xi}}) &= \int_{\mathbb{R}^2} U\left(\hat{\boldsymbol{\xi}} - \hat{\boldsymbol{x}}\right)\hat{\varphi}(\hat{\boldsymbol{x}})d\hat{x}_1 d\hat{x}_2 \\ &= U * \hat{\varphi}(\hat{\boldsymbol{\xi}}), \end{aligned} \tag{30}$$

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

3.1.3. Incoherent Fields. The last piece of physics to be added to the mathematical model is the fact that what is actually measured in many optical devices is the *intensity* of an *incoherent* field. In this setting, the incoherence of the field is related to statistical properties of waves. In the interest of brevity, our discussion is superficial. Interested readers are referred to [51]. In (1) we have only accounted for the spatial component of a time-harmonic wave. The entire wave is of the form $U_\omega(\boldsymbol{x},t) = U(\boldsymbol{x})\exp(i\omega t)$ for fixed frequency ω (see (1)). Define the *mutual coherence function*, Γ, to be the cross correlation of light at $\boldsymbol{x}$ and $\boldsymbol{y}$,

$$\Gamma(\boldsymbol{x},\boldsymbol{y},\tau) \equiv \left\langle\!\left\langle U_\omega(\boldsymbol{x},\cdot+\tau),\ \overline{U}_\omega(\boldsymbol{y},\cdot)\right\rangle\!\right\rangle, \tag{31}$$

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

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

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

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

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

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

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

The intensity of the image ψ in (30) at a point $\hat{\boldsymbol{\xi}}$ is thus given by

$$|\psi|^2 = \left|U * \hat{\varphi}(\hat{\boldsymbol{\xi}})\right|^2. \tag{32}$$

Rearranging the integrals yields

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

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

$$\left|\psi(\hat{\boldsymbol{\xi}})\right|^2 = \int_{\mathbb{R}^2}\int_{\mathbb{R}^2} J(\hat{\boldsymbol{\xi}} - \hat{\boldsymbol{x}}, \hat{\boldsymbol{\xi}} - \hat{\boldsymbol{y}})\hat{\varphi}(\hat{\boldsymbol{x}})\overline{\hat{\varphi}}(\hat{\boldsymbol{y}})d\hat{x}_1 d\hat{x}_2\, d\hat{y}_1 d\hat{y}_2. \tag{33}$$

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

$$J(\hat{\boldsymbol{\xi}}, \hat{\boldsymbol{\eta}}) \approx \hat{c}|U(\hat{\boldsymbol{\xi}})|^2\delta(\hat{\boldsymbol{x}} - \hat{\boldsymbol{\eta}}), \tag{34}$$

where $\hat{c}$ is some real constant. For a detailed discussion of this intricate theory see [51, section 5.5]. Substituting (34) into (33) yields

$$\left|\psi(\hat{\boldsymbol{\xi}})\right|^2 \approx \hat{c}\,|U|^2 * \left|\hat{\varphi}(\hat{\boldsymbol{\xi}})\right|^2. \tag{35}$$

If the coherence of the light is resolvable, then one must work with the less convenient representation of (32).

3.1.4. Rescaling the Model. We assume that $\lambda\xi_3 = 1$ (this is equivalent to resizing the aperture). The contribution of the x_3' component to the field U given in (29) is just a scalar multiple. This is normalized so that the scaling in (35) is unity,

$$\hat{c}|U|^2 * |\varphi|^2(\boldsymbol{\xi}) = |\tilde{c}\mathcal{X}_{\mathbb{A}}^\wedge|^2 * |\varphi|^2(\boldsymbol{\xi}), \tag{36}$$

where $\tilde{c} = -\exp(ik(x_3' + \xi_3))$. We represent the field at the exit pupil of the optical system, that is, on the right "side" of the pupil plane of the imaging system in Figure 1, by the function $u : \mathbb{R}^2 \to \mathbb{C}$. In (36) this field is known:

$$u = \tilde{c}\mathcal{X}_{\mathbb{A}} \quad \text{and} \quad U = u^\wedge. \tag{37}$$

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

The normalized mathematical model for the intensity mapping in the focal plane of a diffracted, incoherent, monochromatic, far-field electromagnetic field (35) becomes

$$|\psi|^2(\boldsymbol{\xi}) \approx |u^\wedge|^2 * |\varphi|^2(\boldsymbol{\xi}). \tag{38}$$

The kernel of the convolution $|u^\wedge|^2$ is known as the *point-spread function* of the idealized optical system of Figure 1. This kernel characterizes the optical system.

3.1.5. Aberrated Optical Systems. It has been assumed that the optical system is in the far field (with respect to some source) of a homogeneous medium; thus the wave at the entrance pupil, that is, on the *left* side of the pupil plane, is characterized by a constant amplitude plane wave, $\arg(u_-) \equiv 0$ and $|u_-| = const$ across the aperture $\mathbb{A}$, where u_- indicates the field at the entrance pupil. This is often called the *reference* wave. In most applications, however, the assumption of homogeneity is not correct for the field at the exit pupil. Inhomogeneities in the media cause deviations in the true wave from the reference wave. There are two types of deviations from the reference wave. We refer to deviations in the phase as *phase aberrations* and deviations from the amplitude as the *throughput* of the optical system. Deviations may occur at any point along the path of propagation and can be caused by an intervening medium such as atmosphere, crystal structure, or mirror surface. In geometric optics, the wave is assumed to travel along rays normal to the wavefront. The phase represents differences in the optical path length along different rays. The locations of the deviations along the rays are not important. Accordingly, all deviations are taken to occur at the pupil plane depicted in Figure 1.

A simple example of a phase aberration is defocus, which can be modeled by use of a lens as in (21). The field due to a defocused generalized pupil function is given by (25), where the lens law (26) is not satisfied, that is, $1/z_0 + 1/\zeta - 1/l = \epsilon,\ 0 < |\epsilon| \ll 1$. It often happens, however, that the aberration is unknown. Defocus is added to an optical system to improve signal-to-noise ratios in the tails of images. Defocus is also used to stabilize numerical schemes for recovering $\arg(u)$ (phase retrieval) and φ (deconvolution) from the image ψ.

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

The function u accounting for all of the above aberrations is referred to as the *generalized pupil function*. The generalized pupil function uniquely characterizes the optical system. For a perfect, deviation-free normalized optical system (where, in particular, $\lambda\xi_3 = 1$) the generalized pupil function is given by $u = \tilde{c}\mathcal{X}_{\mathbb{A}}$, as in (37). For a field with deviations from the reference wave u_-, that is, with phase aberration $\theta(\boldsymbol{x})$ and throughput $A(\boldsymbol{x})$, the generalized pupil function can be represented in complex phasor form by

$$u[A(\boldsymbol{x}), \theta(\boldsymbol{x})] = A(\boldsymbol{x}) \exp[i\theta(\boldsymbol{x})]. \tag{39}$$

The corresponding imaging model for an aberrated optical system is the same as (38).

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

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

The imaging model thus takes the form

$$\psi^2(\boldsymbol{\xi}) \approx |u^\wedge|^2 * \varphi^2(\boldsymbol{\xi}). \tag{40}$$

3.2. Inverse Problems. In the previous sections we have taken great care to develop the *forward* model for image formation. We now turn our attention to the *inverse* problem. If u is known and φ unknown, (40) is a Fredholm integral equation of the first kind. Recovering φ^2 from u and ψ^2 is called, for good reason, *deconvolution*. The *phase retrieval* problem arises when the amplitude of the generalized pupil function u is known, but the phase aberration, θ in (39), is unknown. When both the object φ and the phase aberrations in u are unknown,[8] one is faced with the problem of *simultaneous* deconvolution and phase retrieval. This work is limited to the case of phase retrieval.

3.2.1. Phase Retrieval. While phase aberrations can be recovered from extended sources with the full imaging model (40) (see [23, 50, 68, 76, 94, 95, 96, 120, 123, 122, 105]), they are most often (and more reliably) recovered from images of point sources $\varphi^2 = \delta$. In this case,

$$\psi^2(\boldsymbol{\xi}) \approx |u^\wedge|^2(\boldsymbol{\xi}).$$

The amplitude $|u|$ is assumed to be known and satisfies the equation

$$A = |u|, \tag{41}$$

where $A : \mathbb{R}^2 \to \mathbb{R}_+$ is known. This is often modeled as an indicator function for the aperture, $A = \mathcal{X}_{\mathbb{A}}$. For the purposes of this work it is only necessary to note that $A \in \mathbb{U}_+$, where $\mathbb{U}_+$ is a cone of nonnegative functions to be explicitly defined below. According to the uniqueness results proved in [58], for discrete band-limited images in two dimensions, if a solution to the phase retrieval problem exists, knowledge about both $|u|$ and $|u^\wedge|$ uniquely characterizes u and thus the optical system, up to a complex constant, linear shifts, and rotations by 180 degrees.

In most cases there is no closed-form analytic solution to the phase retrieval problem. Notable exceptions were first recognized in [67, 34, 103]. In numerical approaches, the problem is further constrained by the addition of known phase aberrations to the system. The corresponding images are called *diversity images*. The problem is then to find the unknown phase common to all images given the amplitude constraints. For $m = 1, \ldots, M$, let $\tilde{\theta}_m : \mathbb{R}^2 \to \mathbb{R}$ denote a known phase aberration added to the system across the aperture. The corresponding diversity images are denoted by $\psi_m : \mathbb{R}^2 \to \mathbb{R}$. These images are approximated by

$$\psi_m^2 \approx |\mathcal{P}_m[u]|^2, \tag{42}$$

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

$$\mathcal{P}_m[u] \equiv \left[u \exp[i\tilde{\theta}_m]\right]^\wedge. \tag{43}$$

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

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

$$\begin{pmatrix} A^2 \\ \psi_1^2 \\ \vdots \\ \psi_M^2 \end{pmatrix} = \begin{pmatrix} |u|^2 \\ |\mathcal{P}_1[u]|^2 \\ \vdots \\ |\mathcal{P}_M[u]|^2 \end{pmatrix}. \tag{44}$$

3.2.2. An Optimization Perspective. In the presence of noise it is unlikely that an exact solution to the system of equations given by (44) exists.[9] One therefore seeks the best estimate, u_*, for a given performance measure, ρ. While many different algorithms can be applied to recover the best estimate u_* numerically, it is our view that they all address some type of optimization problem. The method by which the best estimate is found involves some sort of optimality principle that depends on the formulation of the underlying optimization problem. Before stating this optimization problem, some remarks about the spaces in which the operators lie are necessary.

To establish a well-posed optimization problem the domain must be closed. The Fourier transform defined by (19) is only valid on $L^1 \cap L^2$, which is not closed. This technicality is avoided by defining the corresponding transform on L^2. The *Fourier–Plancherel* transform is the unique L^2 limit of the Fourier transform of elements in $L^1 \cap L^2$ [66]. All of the properties of the standard Fourier transform hold for this extended definition. In addition to being closed, the space L^2 has the advantage of being a Hilbert space. In all of the following, the transforms $\mathcal{P}_m : L^2[\mathbb{R}^2, \mathbb{R}^2] \to L^2[\mathbb{R}^2, \mathbb{C}]$ are defined by

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

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

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

$$\mathcal{R}(\boldsymbol{v}) \equiv v_1 + iv_2,$$

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

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

is given by

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

The mapping $\mathcal{R}$ is a unitary bounded linear operator with $\mathcal{R}^{-1} = \mathcal{R}^*$. Our discussion switches frequently between finite-dimensional and infinite-dimensional settings. Therefore, it is convenient to think of $\mathcal{R}$ as a mapping from $L^2[\mathbb{R}^2, \mathbb{R}^2]$ to $L^2[\mathbb{R}^2, \mathbb{C}]$.

[9] The uniqueness results studied in [58] therefore do not apply.

Whenever there is chance for confusion, square brackets are used to indicate a mapping, e.g.,

$$\mathcal{R}[\boldsymbol{v}] \equiv \mathcal{R}(\boldsymbol{v}(\cdot)) \tag{45}$$

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

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

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

The imaging equation (42) is equivalently written as

$$\psi^2(\boldsymbol{\xi}) \approx \left|\mathcal{F}_m[\boldsymbol{u}]\right|^2(\boldsymbol{\xi}), \tag{46}$$

where

$$\mathcal{F}_m[\boldsymbol{u}] \equiv \mathcal{R}^*\left[\mathcal{P}_m[\mathcal{R}[\boldsymbol{u}]]\right]. \tag{47}$$

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

For convenience define

$$\mathcal{F}_0 \equiv \mathcal{I}, \tag{48}$$

where $\mathcal{I}$ is the identity operator. Define the the aperture constraint (41) to be ψ_0:

$$\psi_0 \equiv A.$$

The optimization problem over $L^2[\mathbb{R}^2, \mathbb{R}^2]$ becomes

$$\begin{aligned} &\text{minimize} \sum_{m=0}^{M} \rho\left[\psi_m,\ |\mathcal{F}_m[\boldsymbol{u}]|\ \right] \\ &\text{over} \quad \boldsymbol{u} \in L^2[\mathbb{R}^2, \mathbb{R}^2]. \end{aligned} \tag{49}$$

We have much more flexibility with regard to restrictions on the data ψ_m. These functions are restricted to subsets of the space $\mathbb{U}$, a set of finite-valued functions for which the Fourier transform is well defined and whose tails tend to zero sufficiently fast. The data, ψ_m and A, belong to $\mathbb{U}_+$. For easy reference, the following hypothesis is assumed throughout.

HYPOTHESIS 3.1. *Let $\boldsymbol{u} = (u_{re}, u_{im})$, where u_{re} and $u_{im} \in L^2[\mathbb{R}^2, \mathbb{R}]$. Assume that ψ_m satisfies $\psi_m \in \mathbb{U}_+$ for $m = 0, 1, \ldots, M$, where $\mathbb{U}_+$ is the cone of nonnegative functions given by*

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

The remainder of this work is devoted to the study of numerical methods for the solution of the above optimization problem. We restrict our attention to the optimization problem and optimality principles that underlie methods related to the iterative transform algorithms of Gerchberg, Saxton, Misell, and Fienup [48, 85, 43].

4. Geometric Approaches. Projection algorithms, such as iterative transform methods, are well-known numerical techniques for solving the phase retrieval problem [16, 32, 43, 48, 72, 85, 131, 129]. Much is known about projections onto convex sets [124, 17, 112, 56, 53, 132, 11]. However, the problem of phase retrieval involves projections onto nonconvex sets. It is shown below that as a consequence of nonconvexity the projections can be multivalued. This is the principal obstacle to proving the convergence of projection-type algorithms. For special classes of nonconvex sets, a convergence theory can be provided [31, 25]. The nonconvex sets considered here do not belong to these classes. The geometric analysis of [31] applies to the phase retrieval problem, although it requires assumptions that are difficult to satisfy. A convergence theory for generalized projection algorithms is developed in [6]; however, there are no known nonconvex sets to which their hypotheses apply. In particular, the hypotheses required in Proposition 2 of [6] are not satisfied in the case of phase retrieval.

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

The amplitude data for a one-dimensional example is depicted in Figure 4. The functions satisfying the data belong to sets that are collections of functions that lie on the surface of the tube-like structures depicted in Figure 5.

Given $\psi_m \not\equiv 0$ and $\tilde{\theta}_m$ measurable, the mathematical description of these tube-like sets is

$$\mathbb{Q}_m \equiv \left\{ \boldsymbol{u} \in L^2[\mathbb{R}^2, \mathbb{R}^2] \mid |\mathcal{F}_m[\boldsymbol{u}]| = \psi_m \text{ a.e.} \right\}. \tag{50}$$

PROPERTY 4.1 (see Appendix A for proof). *The sets $\mathbb{Q}_m$ defined by* (50) *are neither weakly closed nor convex in $L^2[\mathbb{R}^2, \mathbb{R}^2]$ whenever ψ_m is not identically zero.*

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

We now develop the projection theory for sets of the form (50). Let $\mathbb{X}$ be a metric space with metric $\rho : \mathbb{X} \to \mathbb{R}_+$ and let $\mathbb{Q} \subset \mathbb{X}$. Define the distance of a point $x \in \mathbb{X}$ to the set $\mathbb{Q}$ by

$$\operatorname{dist}(x; \mathbb{Q}) \equiv \inf_{u \in \mathbb{Q}} \rho(x, u). \tag{51}$$

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

$$\Pi_{\mathbb{Q}}[v] \equiv \arg\min_{u \in \mathbb{Q}} \|v - u\| = \{\bar{u} \in \mathbb{Q} : \|v - \bar{u}\| = \inf_{u \in \mathbb{Q}} \|v - u\|\}. \tag{52}$$

There is a general theory that addresses the question of the existence of projections onto sets in a metric space [121, 38]. Fortunately, we are able to provide a simple *constructive* proof of existence of sets of the type given by (50), which simultaneously

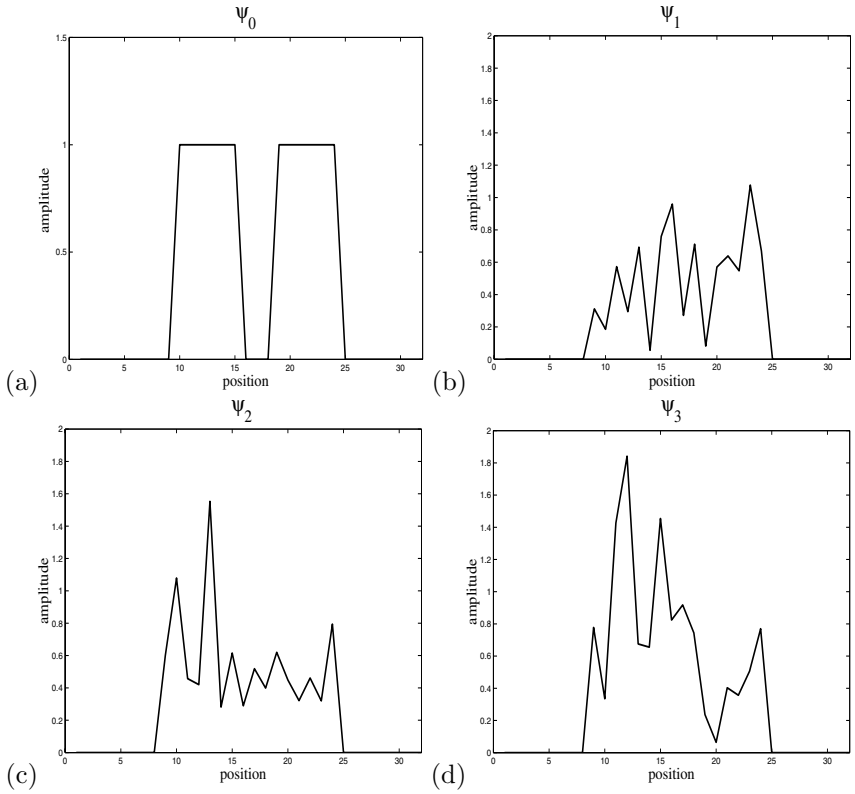

Fig. 4 *One-dimensional pupil with corresponding image data. Frame* (a) *is a one-dimensional cross section of the amplitude across the aperture shown in Figure* 6a. *Frames* (b)–(d) *are cross sections of the corresponding point-spread functions for the aperture in frame* (a) *with some unknown phase aberration as well as a known defocus.*

provides a complete description of these projections. The formulation agrees for the most part with what has heretofore been called the *projection* in the literature. While it is elementary, we are not aware of any other proof of the existence of this specific projection, much less its precise characterization.

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

$$\mathbb{Q}[b] \equiv \left\{ \boldsymbol{u} \in L^2[\mathbb{R}^2, \mathbb{R}^2] \;\middle|\; |\boldsymbol{u}| = b \text{ a.e.} \right\}, \tag{53}$$

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

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

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

$$\pi[\boldsymbol{u}; b, \theta](\boldsymbol{x}) \equiv \begin{cases} b(\boldsymbol{x}) \frac{\boldsymbol{u}(\boldsymbol{x})}{|\boldsymbol{u}(\boldsymbol{x})|} & \text{for } \boldsymbol{u}(\boldsymbol{x}) \neq 0, \\ b(\boldsymbol{x}) \mathcal{R}^* \left[\exp[i\theta(\boldsymbol{x})] \right] & \text{for } \boldsymbol{u}(\boldsymbol{x}) = 0 \end{cases} \tag{54}$$

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

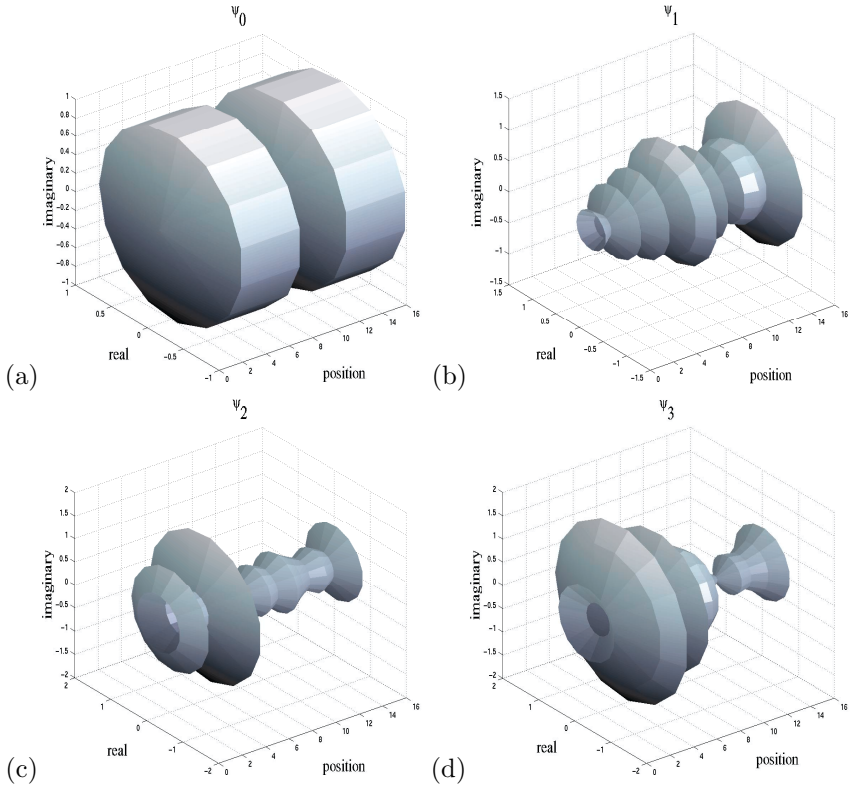

Fig. 5 *Tube constraints. The vertical axis and the axis coming out of the page correspond to the real and imaginary components of the tubes. The horizontal axes correspond to the horizontal axes of Figure* 4. *Frame* (a) *represents the constraint set corresponding to Figure* 4a. *Frames* (b)–(d) *represent the constraint sets corresponding to Figures* 4b–4d.

THEOREM 4.2. *For every* $b \in L^2[\mathbb{R}^2, \mathbb{R}_+]$ *and* $\boldsymbol{u}, \boldsymbol{v} \in L^2[\mathbb{R}^2, \mathbb{R}^2]$, *we have*

$$\boldsymbol{v} \in \Pi[\boldsymbol{u}; b] \iff |\boldsymbol{v}(\boldsymbol{x}) - \boldsymbol{u}(\boldsymbol{x})| = \operatorname{dist}(\boldsymbol{u}(\boldsymbol{x}); b(\boldsymbol{x})\mathbb{S}) \ \textit{a.e.}, \tag{55}$$

$$\Pi_{\mathbb{Q}[b]}[\boldsymbol{u}] = \Pi[\boldsymbol{u}; b], \quad \textit{and} \tag{56}$$

$$\operatorname{dist}(\boldsymbol{u}; \mathbb{Q}[b]) = \| \, |\boldsymbol{u}| - b\|. \tag{57}$$

Proof. We first show (55). Let $\boldsymbol{u} \in L^2[\mathbb{R}^2, \mathbb{R}^2]$ and $b \in L^2[\mathbb{R}^2, \mathbb{R}_+]$ be given. Observe that if $\pi[\boldsymbol{u}; b, \theta] \in \Pi[\boldsymbol{u}; b]$, then $\pi[\boldsymbol{u}; b, \theta] \in \mathbb{Q}[b]$ and

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

That is, the function $\pi[\boldsymbol{u}; b, \theta]$ attains the pointwise distance of $\boldsymbol{u}(\boldsymbol{x})$ to the set $b(\boldsymbol{x})\mathbb{S}$ on $\mathbb{R}^2$. Conversely, suppose that $\boldsymbol{v} \in L^2[\mathbb{R}^2, \mathbb{R}^2]$ attains the pointwise distance of $\boldsymbol{u}(\boldsymbol{x})$ to the set $b(\boldsymbol{x})\mathbb{S}$ on $\mathbb{R}^2$. Then by [104, Corollary 1.9.e] there exists a complex measurable function $\alpha : \mathbb{R}^2 \to \mathbb{C}$ such that $|\alpha(\boldsymbol{x})| = 1$ for all $\boldsymbol{x} \in \mathbb{R}^2$ and $\mathcal{R}[\boldsymbol{v}] = \alpha|\boldsymbol{v}|$. Define the measurable function $\theta : \mathbb{R}^2 \to \mathbb{R}$ by $\theta = \cos^{-1}(\operatorname{Re}(\alpha))$, where we take the

principal branch of $\cos^{-1}$. Then $\alpha = \exp[i\theta]$. Consequently,

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

which implies that $\boldsymbol{v} \in \Pi[\boldsymbol{u}; b]$. Therefore (55) holds.

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

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

Therefore,

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

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

Conversely, if $\boldsymbol{v} \in \Pi_{\mathbb{Q}[b]}[\boldsymbol{u}]$, then, in particular, $\boldsymbol{v} \in \mathbb{Q}[b]$. If $\boldsymbol{v} \notin \Pi[\boldsymbol{u}; b]$, then, as above, there is a set of positive measure on which $\boldsymbol{v}$ does not attain the pointwise distance to the set $b(\boldsymbol{x})\mathbb{S}$, which implies the contradiction $\|\boldsymbol{u} - \pi[\boldsymbol{u}; b, \theta]\| < \|\boldsymbol{u} - \boldsymbol{v}\|$ for any function $\pi[\boldsymbol{u}; b, \theta] \in \Pi[\boldsymbol{u}; b]$. Thus we have established (56).

We now show (57). Choose $\pi[\boldsymbol{u}; b, \theta]$ from $\Pi_{\mathbb{Q}[b]}[\boldsymbol{u}]$. Then

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

As an elementary consequence of Theorem 4.2 we are also able to characterize the projection onto the sets $\mathbb{Q}_m$ defined in (50). The projection mappings $\Pi_{\mathbb{Q}[b]}$ are multivalued mappings from $L^2[\mathbb{R}^2,\mathbb{R}^2]$ to $\mathbb{R}^2$. For any linear operator $A : L^2 \to \mathbb{X}$, where $\mathbb{X}$ is any topological vector space, we define the image of the multivalued mapping $\Pi_{\mathbb{Q}[b]}$ under A to be the set

$$A\Pi_{\mathbb{Q}[b]}[\boldsymbol{u}] = \left\{ A[\boldsymbol{v}] \;\middle|\; \boldsymbol{v} \in \Pi_{\mathbb{Q}[b]}[\boldsymbol{u}] \;\right\}.$$

COROLLARY 4.3. *Let the set $\mathbb{Q}_m$ be defined as in* (50) *and let the operators $\Pi_{\mathbb{Q}_m}$ and $\mathcal{F}_m$ be as defined in* (52) *and* (47)*, respectively. Then*

$$\Pi_{\mathbb{Q}_m}[\boldsymbol{u}] = \mathcal{F}_m^* \left[\Pi_{\mathbb{Q}[\psi_m]}[\mathcal{F}_m[\boldsymbol{u}]]\right] \tag{58}$$

and

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

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

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

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

Therefore,

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

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

Finally, since $\mathcal{F}_m$ is unitary, we obtain from Theorem 4.2 that

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

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

4.2. Projection Algorithms. A general framework for projection algorithms can be found in [11], which considered sequences of weighted relaxed projections of the form

$$\boldsymbol{u}^{(\nu+1)} \in \left(\sum_{m=0}^{M} \gamma_m^{(\nu)} \left[(1-\alpha_m^{(\nu)})\mathcal{I} + \alpha_m^{(\nu)}\Pi_{\mathbb{Q}_m}\right]\right)[\boldsymbol{u}^{(\nu)}]. \tag{59}$$

Here $\mathcal{I}$ is the identity mapping, $\alpha_m^{(\nu)}$ is a relaxation parameter usually in the interval $[0,2]$, and the weights $\gamma_m^{(\nu)}$ are nonnegative scalars summing to 1. General results

for these types of algorithms apply only to convex sets. In the convex setting the inclusion in algorithm (59) is an equality since projections onto convex sets are single valued. In the nonconvex setting this is not the case.

The Gerchberg–Saxton algorithm [48] and its variants can be viewed as an instance of algorithm (59). To see this, define the set of *active indices* at iteration ν by

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

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

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

This is an instance of what is called a *cyclic* projection algorithm [11]. Projections onto the sets $\mathbb{Q}_m$ are calculated one at a time in a sequential manner. Thus $M+1$ iterations of this cyclic algorithm are the same as one iteration of the following sequential projection algorithm, known in the optics community as the *iterative transform algorithm*:

$$\boldsymbol{u}^{(\nu+1)} \in \left(\prod_{m=0}^{M} \left[(1 - \alpha_m^{(\nu)})\mathcal{I} + \alpha_m^{(\nu)} \Pi_{\mathbb{Q}_m} \right] \right) [\boldsymbol{u}^{(\nu)}]. \tag{60}$$

The Gerchberg–Saxton algorithm [48] is obtained by setting $M = 1$ and $\alpha_0^{(\nu)} = \alpha_1^{(\nu)} = 1$. Variants of this algorithm [85, 78] involve increasing the number of diversity images, that is, $M > 1$, and adjusting the relaxation parameters $\alpha_m^{(\nu)}$. Convergence results often cited for the Gerchberg–Saxton algorithm refer to the observation that the set distance error, defined as the sum of the distances of an iterate $\boldsymbol{u}^{(\nu)}$ to two constraint sets, $\mathbb{Q}_0$ and $\mathbb{Q}_1$, will not increase as the iteration proceeds [72]. For $M > 1$, this may not be the case. That is, the set distance error can *increase*. In all cases, the algorithm may fail to converge due to the nonconvexity of the sets $\mathbb{Q}_m$ (see Levi and Stark [72] for an example of this behavior).

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

$$\lambda^{(\nu)} \beta_m^{(\nu)} \equiv \gamma_m^{(\nu)} \alpha_m^{(\nu)} \tag{61}$$

to rewrite algorithm (59) as

$$\boldsymbol{u}^{(\nu+1)} \in \left(\mathcal{I} - \lambda^{(\nu)} \mathcal{G}^{(\nu)} \right) [\boldsymbol{u}^{(\nu)}], \tag{62}$$

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

$$\mathcal{G}^{(\nu)} \equiv \sum_{m=0}^{M} \mathcal{G}_m^{(\nu)}, \tag{63}$$

where

$$\mathcal{G}_m^{(\nu)} \equiv \beta_m^{(\nu)} \left(\mathcal{I} - \Pi_{\mathbb{Q}_m} \right). \tag{64}$$

In algorithm (62) the nonnegative weights $\beta_m^{(\nu)}$ do not necessarily sum to 1, and the parameters $\lambda^{(\nu)}$ are to be interpreted as a *step length*. This formulation of the projection algorithm is shown in section 5 to be equivalent to a steepest descent algorithm for a weighted squared distance function under very special circumstances. In our opinion the multivalued nature of the projections has not been adequately addressed in the numerical theory for the phase retrieval problem. Insufficient attention to this detail can result in unstable numerical calculations. This is discussed in section 7. Several authors have proposed extensions to projection algorithms to overcome stagnation [43, 116]. These methods are a valuable topic for further study; however, in order to illustrate the comparison between geometric methods and analytic methods studied in the following sections, we restrict our attention to simple projection algorithms of the form of algorithm (59) and algorithm (60)

5. Analytic Methods. Convergence results for projection methods applied to the phase retrieval problem are not possible in general due to the nonconvexity of the constraint sets. In this section we show that the nonconvexity of the constraint sets is related to the nonsmoothness of the square of the set distance error $\operatorname{dist}(\boldsymbol{u};\mathbb{Q}_m)$ defined in (51). This is fundamentally different from the convex setting in a Hilbert space, where the squared distance function is smooth. The nonsmoothness of the squared distance function in the nonconvex setting is a consequence of the multivaluedness of the projection operator. In this section some insight into this relationship is given.

Our numerical methods are based on smooth approximations to the squared set distance error E. This allows us to provide a convergence theory that is easily derived from standard results in the optimization literature. By relating the smooth approximations to the projection operators, we are able to provide an interpretation of iterative transform methods in the context of the analytic methods studied in this section.

There are many different approaches based on other error metrics (maximum entropy, for example). Since the primary focus of this work is on projection methods and related techniques, we limit our discussion to two analytic approaches. The first is a direct application of smoothing methods, which we refer to as perturbed least squares; the second is an extended least squares approach that allows us to adaptively correct for the relative variability in the diversity measurements, ψ_m.

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

$$E[\boldsymbol{u}] = \sum_{m=0}^{M} \frac{\beta_m}{2} \operatorname{dist}^2(\boldsymbol{u};\mathbb{Q}_m), \tag{65}$$

where $\beta_m \geq 0$ for $m = 0,\ldots,M$, and by Corollary 4.3,

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

With this least squares objective the optimization problem (49) becomes

$$\begin{aligned} &\text{minimize } E[\boldsymbol{u}] \\ &\quad \text{over} \quad \boldsymbol{u} \in L^2[\mathbb{R}^2,\mathbb{R}^2]. \end{aligned} \tag{66}$$

In general the optimal value for this problem is nonzero, and so classical techniques for solving the problem numerically are based on satisfying a first-order necessary condition for optimality. For smooth functions, this condition simply states that

the gradient takes the value zero at any local solution to the optimization problem. However, the functions $\operatorname{dist}^2(\boldsymbol{u};\mathbb{Q}_m)$ are not differentiable. The easiest way to see this is to consider the one-dimensional function $a(x) = |\,|x| - b|^2$, where $b > 0$. This function is not differentiable at $x = 0$ (indeed, it is not even subdifferentiably regular at $x = 0$ [101, Def. 7.25]). It is precisely at these points that the pointwise projection operator $\Pi_{\mathbb{Q}_m}$ is multivalued. Similarly, the functions $\operatorname{dist}^2(\boldsymbol{u};\mathbb{Q}_m)$ are not differentiable at functions $\boldsymbol{u}$ for which there exists a set $\Omega \subset \operatorname{supp}(\psi_m)$ of positive measure on which $\boldsymbol{u} \equiv 0$. Nondifferentiability in this context is related to the nondifferentiability of the Euclidean norm at the origin: $\nabla|\boldsymbol{x}| = \boldsymbol{x}/|\boldsymbol{x}|$. A common technique to avoid division by zero is to add a small positive quantity to the denominator of any suspect rational expression. This device was used in [46] to avoid division by zero in the representation of the derivative of the modulus function. However, this is not a principled approach to the need for approximating the modulus function and its derivatives globally. In the next section we study an alternative approximation to the modulus function itself that possesses excellent global approximation properties.

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

$$0 \in \partial E[\boldsymbol{u}_*],$$

where ∂ denotes a *subgradient* operator such as those studied in [27, 26, 86, 87, 63, 64]. It was shown in [75] that one can apply the calculus of subdifferentials to obtain the identity

$$\partial\left(\operatorname{dist}^2(\boldsymbol{u};\mathbb{Q}_m)\right) = 2\operatorname{cl}^*\left(\mathcal{I} - \Pi_{\mathbb{Q}_m}\right)[\boldsymbol{u}], \tag{67}$$

where cl^* denotes the weak-star closure. Using this fact and standard calculus rules for the subdifferential, we obtain that

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

where $\mathcal{G}_m$ is defined by (64). An understanding of the theory of subdifferentials is not required for the numerical theory developed in subsequent sections. Readers interested in these relationships are referred to [75]. In order to avoid the difficulties associated with nondifferentiability, we consider smooth objectives that are perturbations of the least squares objective functional E.

5.2. Perturbed Least Squares. One obvious solution to the problem of nonsmooth objectives is simply to square the data and the modulus. The modulus squared is a smooth function. For this reason, analytic techniques tend to favor objectives based on the modulus squared. See [35] for a very careful treatment of analytic techniques for the modulus squared. In our experiments, however, objectives based on the modulus squared, while robust, suffer from very slow rates of convergence compared to the nonsmooth or nearly nonsmooth objectives studied in section 5. An intuitive explanation for this is that the modulus squared smooths out curvature information in the objective [65, 71]. Another explanation is that the singular values of the operator $|\mathcal{F}_m[\boldsymbol{u}]|^2$ are much more spread out compared to those of the operator $|\mathcal{F}_m[\boldsymbol{u}]|$; that is, the squared modulus system is more ill conditioned than the modulus system. This results in slower convergence of methods based on linearizations of the operator $|\mathcal{F}_m[\boldsymbol{u}]|^2$. See [4, 57] for a discussion. While it is difficult to work with, we

have found that the modulus function outperforms the modulus squared function as an objective in optimization techniques. The principal goal of this work is to develop tools for taking advantage of these "good" aspects of the modulus, while avoiding instabilities.

The smooth least squares objective function we consider in this section is based on a smooth perturbation of the modulus function $|\boldsymbol{u}|$ of the form

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

This smoothing of the modulus function enjoys three key properties,

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

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

$$E_\epsilon[\boldsymbol{u}] = \sum_{m=0}^{M} \frac{\beta_m}{2} \left\| \frac{|\mathcal{F}_m[\boldsymbol{u}]|^2}{\left(|\mathcal{F}_m[\boldsymbol{u}]|^2 + \epsilon^2\right)^{1/2}} - \psi_m \right\|^2, \tag{68}$$

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

Next we study the analytic properties of E_ϵ. Define the pointwise residual $r : \mathbb{R}^2 \times \mathbb{R}_+ \to \mathbb{R}$ by

$$r(\boldsymbol{u}; b, \epsilon) = \frac{|\boldsymbol{u}|^2}{(|\boldsymbol{u}|^2+\epsilon^2)^{1/2}} - b. \tag{69}$$

When the arguments of r are *functions*, this is denoted as usual with square brackets. Denote the extended reals by $\overline{\mathbb{R}} = \mathbb{R} \cup \{\infty\}$, and define the integral functional $J : L^2[\mathbb{R}^2,\mathbb{R}^2] \to \overline{\mathbb{R}}$ by

$$J[\boldsymbol{u}; b, \epsilon] = \int_{\mathbb{R}^2} r^2(\boldsymbol{u}(\boldsymbol{x}); b(\boldsymbol{x}), \epsilon) d\boldsymbol{x}. \tag{70}$$

For $\boldsymbol{u}(\cdot) \in L^2[\mathbb{R}^2,\mathbb{R}^2]$, $b(\cdot) \in L^1[\mathbb{R}^2,\mathbb{R}_+] \cap L^2[\mathbb{R}^2,\mathbb{R}_+] \cap L^\infty[\mathbb{R}^2,\mathbb{R}_+]$, and all ϵ, $J[\boldsymbol{u}(\cdot); b(\cdot), \epsilon]$ is finite valued and Fréchet differentiable with globally Lipschitz continuous derivative (see Appendix B). Using this notation, we can rewrite E_ϵ in composition form as

$$E_\epsilon[\boldsymbol{u}] \equiv \sum_{m=0}^{M} \frac{\beta_m}{2} \left(J[\cdot; \psi_m, \epsilon] \circ \mathcal{F}_m\right)[\boldsymbol{u}]. \tag{71}$$

In Appendix B it is shown that E_ϵ is Fréchet differentiable as a function on L^2, with Fréchet derivative given by

$$E'_\epsilon[\boldsymbol{u}][\boldsymbol{w}] = \sum_{m=0}^{M} \frac{\beta_m}{2} \left(J[\cdot; \psi_m, \epsilon] \circ \mathcal{F}_m\right)'[\boldsymbol{u}][\boldsymbol{w}], \tag{72}$$

where

(73)
$$\left(J[\cdot;\psi_m,\epsilon]\circ\mathcal{F}_m\right)'[\boldsymbol{u}][\boldsymbol{w}] = 2\left\langle \mathcal{F}_m^*\left[r[\mathcal{F}_m[\boldsymbol{u}];\psi_m,\epsilon]\frac{|\mathcal{F}_m[\boldsymbol{u}]|^2+2\epsilon^2}{\left(|\mathcal{F}_m[\boldsymbol{u}]|^2+\epsilon^2\right)^{3/2}}\mathcal{F}_m[\boldsymbol{u}]\right],\ \boldsymbol{w}\right\rangle.$$

Since $E_\epsilon : L^2 \to \mathbb{R}$ is Fréchet differentiable, $E'_\epsilon[\boldsymbol{u}]$ is an element of the dual of L^2 for all $\boldsymbol{u}$. Since L^2 is a Hilbert space, we can identify $E'_\epsilon[\boldsymbol{u}]$ with an element of L^2. We denote this element by $\nabla E_\epsilon[\boldsymbol{u}]$. Using (72) and (73) we obtain the formula

$$\nabla E_\epsilon[\boldsymbol{u}] = \sum_{m=0}^{M}\beta_m\mathcal{F}_m^*\left[r[\mathcal{F}_m[\boldsymbol{u}];\psi_m,\epsilon]\frac{|\mathcal{F}_m[\boldsymbol{u}]|^2+2\epsilon^2}{\left(|\mathcal{F}_m[\boldsymbol{u}]|^2+\epsilon^2\right)^{3/2}}\mathcal{F}_m[\boldsymbol{u}]\right]. \tag{74}$$

See Appendix B for the complete calculation.

We now establish the principal relationship between ∇E_ϵ and the operator $\mathcal{G}$ given by (63).

THEOREM 5.1 (see Appendix B). *Let the functions $\boldsymbol{u}$ and ψ_m satisfy Hypothesis 3.1. At each $\boldsymbol{u}$ with $E[\boldsymbol{u}] < \delta$, there exists an $\epsilon > 0$ such that*

$$\|\nabla E_\epsilon[\boldsymbol{u}] \ - \ \boldsymbol{v}\| \ < \ C\delta^{1/2} \tag{75}$$

for all $\boldsymbol{v} \in \mathcal{G}[\boldsymbol{u}]$, where

$$\mathcal{G} = \ \sum_{m=0}^{M}\beta_m\left(\mathcal{I}-\Pi_{\mathbb{Q}_m}\right)$$

and

$$C = \sqrt{2}\sum_{m=0}^{M}\beta_m^{1/2}\left(1+\sqrt{2}\beta_m^{1/2}\right).$$

REMARK 5.2. *Though $\mathcal{G}$ is a multivalued mapping, the norm $\|\nabla E_\epsilon[\boldsymbol{u}]-\boldsymbol{v}\|$ is the same for every $\boldsymbol{v}\in\mathcal{G}[\boldsymbol{u}]$. See Appendix B for details.*

Suppose $E[\boldsymbol{u}] < \delta$. Then from (75) we have

$$\|\nabla E_\epsilon[\boldsymbol{u}]\|^2 - 2\langle\nabla E_\epsilon[\boldsymbol{u}],\,\boldsymbol{v}\rangle + \|\boldsymbol{v}\|^2 \le C^2\delta$$

for every $\boldsymbol{v}\in\mathcal{G}[\boldsymbol{u}]$. Therefore, if $\|\nabla E_\epsilon[\boldsymbol{u}]\|^2 + \|\boldsymbol{v}\|^2 \ \ge C^2\delta$, then the direction $-\boldsymbol{v}$ is necessarily a direction of descent for $E_\epsilon[\boldsymbol{u}]$ for every $\boldsymbol{v}\in\mathcal{G}[\boldsymbol{u}]$. In particular, if a line search algorithm

$$\boldsymbol{u}^{(\nu+1)} = \boldsymbol{u}^{(\nu)} - \lambda^{(\nu)}\nabla E_\epsilon[\boldsymbol{u}^{(\nu)}]$$

produces a sequence with $E_\epsilon(\boldsymbol{u}^{(\nu)})\to 0$, then the corresponding projection algorithm

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

behaves similarly. That is, the qualitative convergence behavior of the projection algorithm can be studied by examining the convergence properties of the corresponding line search algorithm for the perturbed objective. However, in the presence of noise,

where the global solution to (66) is greater than zero, the behavior of the algorithms near the solution could differ significantly since the bound (75) does not guarantee that dist $(\nabla E_\epsilon[\boldsymbol{u}], \mathcal{G}[\boldsymbol{u}]) \to 0$.

The principal obstacle to a bound of the form (75) depending only on ϵ and not on the value of $E[\boldsymbol{u}]$ is the possibility that the estimate $\boldsymbol{u}$ has a domain of positive measure over which $\boldsymbol{u}$ is near zero but the data is nonzero. In the numerical literature for wavefront reconstruction, this difficulty is often circumvented by either implicitly or explicitly assuming that none of the estimates $\boldsymbol{u}$ have this property. If one is willing to make this assumption, then a bound of the form (75) that depends only on ϵ is possible. The discrepancy in the general case is consistent with the fact that ∇E_ϵ is a smooth approximation of the multivalued projection operator.

Define $\mathbb{V} \subset L^2[\mathbb{R}^2, \mathbb{R}^2]$ by

$$\mathbb{V} \equiv \bigcap_{m=0}^{M} \mathbb{V}_m,$$

where

$$\mathbb{V}_m \equiv \{\boldsymbol{v} \mid |\mathcal{F}_m[\boldsymbol{v}]| \neq 0 \text{ a.e. on } \operatorname{supp}(\psi_m)\}.$$

In the next corollary we establish that for every $\boldsymbol{v} \in \mathbb{V}$ the projection operator is single valued and the gradient ∇E_ϵ converges pointwise to the operator $\mathcal{G}$.

COROLLARY 5.3 (see Appendix B). *Let the hypotheses of Theorem* 5.1 *hold and let* $\boldsymbol{u} \in \mathbb{V} \neq \emptyset$; *then* $\mathcal{G}[\boldsymbol{u}]$ *is single valued. Suppose further that for each* $m = 0, 1, 2, \ldots$

$$\tilde{\psi}_m = \frac{\psi_m}{|\mathcal{F}_m[\boldsymbol{u}]|} \mathcal{X}_{\psi_m} \in L^\infty[\mathbb{R}^2, \mathbb{R}_+].$$

Then given any $\delta > 0$ *there exists an* $\epsilon > 0$ *such that*

$$\|\nabla E_\epsilon[\boldsymbol{u}] - \mathcal{G}[\boldsymbol{u}]\| \leq \delta.$$

REMARK 5.4. *The assumptions of Corollary* 5.3 *are extremely strong. While each of the sets* $\mathbb{V}_m$ *is dense in* $L^2[\mathbb{R}^2, \mathbb{R}^2]$, *this is not true for the intersection. Indeed, it is common that* $\mathbb{V} = \emptyset$, *as in the case of noisy data.*

Supposing $\mathbb{V} \neq \emptyset$, for $\boldsymbol{u} \in \mathbb{V}$ we define the "gradient" of the unperturbed set distance error by

$$\nabla E[\boldsymbol{u}] \equiv \lim_{\epsilon \to 0} \nabla E_\epsilon[\boldsymbol{u}].$$

Together with Lebesgue's dominated convergence theorem [66, p. 133], the above corollary implies that for $\boldsymbol{u} \in \mathbb{V} \neq \emptyset$

$$\nabla E[\boldsymbol{u}] = \mathcal{G}[\boldsymbol{u}] \quad \text{a.e.}$$

Note that $\nabla E[\boldsymbol{u}]$ is not the gradient in the Fréchet sense. In [6] the authors impose assumptions that allow them to prove that this object *is* the gradient in the Fréchet sense. Again, in most practical situations $\mathbb{V} = \emptyset$; thus the applicability of any such assumption is very narrow. Applying this theory to algorithms is also problematic. Supposing that $\mathbb{V} \neq \emptyset$, then one must find an initial point $\boldsymbol{u}_0 \in \mathbb{V}$. Once an initial admissible point is found, one must guarantee that all subsequent iterates remain in $\mathbb{V}$ as well. Algorithms that do not take this into account suffer from numerical instabilities. This issue is revisited in section 7.

5.3. Extended Least Squares. The projection algorithm (59) allows the user to choose the relaxation parameters $\alpha_m^{(\nu)}$ and weightings $\gamma_m^{(\nu)}$ at each iteration ν. This begs the question as to what the *optimal* choice of these parameters might be. Under the change of variables (61), one is similarly confronted with the issue of optimally selecting the step lengths $\lambda^{(\nu)}$ and weights $\beta_m^{(\nu)}$. Step lengths are discussed in section 6.1. In this section we consider an approach to optimal weight selection. This requires an extension of (66) to accommodate variable weights.

Following the work of Bell, Burke, and Shumitzky [13], define the objective

$$L_\epsilon[\boldsymbol{u},\boldsymbol{\beta}] = \sum_{m=0}^{M} -\ln(2\pi\beta_m) + \beta_m \left(J[\mathcal{F}_m[\boldsymbol{u}];\psi_m,\epsilon] + G_m[\boldsymbol{u}]\right), \tag{76}$$

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

$$\begin{aligned} &\text{minimize } L_\epsilon[\boldsymbol{u},\boldsymbol{\beta}] \\ &\quad \text{over} \quad \boldsymbol{u} \in L^2[\mathbb{R}^2,\mathbb{R}^2],\ 0 < \boldsymbol{\beta}. \end{aligned}$$

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

LEMMA 5.1. *Let* $L_\epsilon : L^2[\mathbb{R}^2,\mathbb{R}^2] \times \mathbb{R}_+^{M+1} \to \mathbb{R}$ *be defined by* (76) *and let* $\boldsymbol{u} \in L^2[\mathbb{R}^2,\mathbb{R}^2]$. *Let*

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

where

$$\beta_{m*}[\boldsymbol{u}] = (J[\mathcal{F}_m[\boldsymbol{u}];\psi_m,\epsilon] + c_m)^{-1} \qquad \text{for } m = 0,\ldots,M. \tag{77}$$

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

Proof. This is nearly identical to Lemma 1 of Bell, Burke, and Shumitzky [13]. Their proof also holds in this setting. □

Substituting $\boldsymbol{\beta}_*[\boldsymbol{u}]$ for $\boldsymbol{\beta}$ into (76) yields

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

Dropping the constants yields the reduced objective

$$R_\epsilon[\boldsymbol{u}] = \sum_{m=0}^{M} \ln(J[\mathcal{F}_m[\boldsymbol{u}];\psi_m,\epsilon] + c_m). \tag{78}$$

The corresponding optimization problem is

$$\begin{aligned} &\text{minimize } R_\epsilon[\boldsymbol{u}] \\ &\text{over} \quad \boldsymbol{u} \in L^2[\mathbb{R}^2, \mathbb{R}^2]. \end{aligned} \tag{79}$$

For Fréchet differentiable J, the objective above is Fréchet differentiable with derivative given by

$$R'_\epsilon[\boldsymbol{u}][\boldsymbol{w}] = \sum_{m=0}^{M} \left(J[\mathcal{F}_m[\boldsymbol{u}]; \psi_m, \epsilon] + c_m\right)^{-1} \left(J[\cdot; \psi_m, \epsilon] \circ \mathcal{F}_m\right)'[\boldsymbol{u}][\boldsymbol{w}],$$

where, by (73) and (77),

$$\nabla R_\epsilon[\boldsymbol{u}] = 2 \sum_{m=0}^{M} \beta_{m*}[\boldsymbol{u}] \mathcal{F}_m^* \left[r[\mathcal{F}_m[\boldsymbol{u}]; \psi_m, \epsilon] \frac{|\mathcal{F}_m[\boldsymbol{u}]|^2 + 2\epsilon^2}{\left(|\mathcal{F}_m[\boldsymbol{u}]|^2 + \epsilon^2\right)^{3/2}} \mathcal{F}_m[\boldsymbol{u}] \right]. \tag{80}$$

This is simply the Fréchet derivative of the perturbed least squares objective scaled by the inverse of the perturbed squared set distance plus some constant. The complete calculation can be found in Appendix B.

6. Numerical Methods. In this section we present two basic numerical approaches for the minimization of the perturbed least squares objective, E_ϵ, and the perturbed extended least squares objective, R_ϵ. The first algorithm is a simple first-order line search method, while the second is a trust region algorithm that incorporates curvature information using limited memory techniques.

6.1. Line Search. Let $F : L^2[\mathbb{R}^2, \mathbb{R}^2] \to \mathbb{R}$ be Fréchet differentiable. Given an initial estimate of the solution $\boldsymbol{u}^{(0)}$, a descent algorithm for the minimization of F generates iterates $\boldsymbol{u}^{(\nu)}$ by the rule

$$\boldsymbol{u}^{(\nu+1)} = \boldsymbol{u}^{(\nu)} + \lambda^{(\nu)} \boldsymbol{w}^{(\nu)},$$

where

$$\boldsymbol{w}^{(\nu)} \in \mathbb{D}[\boldsymbol{u}^{(\nu)}] = \left\{ \boldsymbol{w} \in L^2[\mathbb{R}^2, \mathbb{R}^2] \;\middle|\; F'[\boldsymbol{u}^{(\nu)}][\boldsymbol{w}] < 0 \right\} \tag{81}$$

and $\lambda^{(\nu)}$ is a well-chosen step length parameter.

There are several methods for computing a suitable step length [89]. The criteria we use is the *sufficient decrease* condition:

$$F[\boldsymbol{u}^{(\nu)} + \lambda^{(\nu)} \boldsymbol{w}^{(\nu)}] \leq F^{(\nu)} + \eta \lambda^{(\nu)} \left\langle \nabla F^{(\nu)},\ \boldsymbol{w}^{(\nu)} \right\rangle, \tag{82}$$

where $0 < \eta < 1$ is a fixed parameter and

$$F^{(\nu)} \equiv F[\boldsymbol{u}^{(\nu)}] \qquad \text{and} \qquad \nabla F^{(\nu)} \equiv \nabla F[\boldsymbol{u}^{(\nu)}].$$

THEOREM 6.1 (see Appendix B). *Let $F : L^2[\mathbb{R}^2, \mathbb{R}^2] \to \mathbb{R}$ be Fréchet differentiable and bounded below. Consider the following algorithm.*

Step 0: *(Initialization) Choose $\gamma \in (0, 1)$, $\eta \in (0, 1)$, $c \geq 1$, and $\boldsymbol{u}^{(0)} \in L^2[\mathbb{R}^2, \mathbb{R}^2]$, and set $\nu = 0$.*

Step 1: *(Search Direction) If $\mathbb{D}[\boldsymbol{u}^{(\nu)}] = \emptyset$, STOP; otherwise, choose $\boldsymbol{w}^{(\nu)} \in \mathbb{D}[\boldsymbol{u}^{(\nu)}] \cap c\mathbb{B}$, where $\mathbb{D}$ is defined by (81) and $\mathbb{B}$ is the closed unit ball in $L^2[\mathbb{R}^2, \mathbb{R}^2]$.*

Step 2: *(Step Length) Set*

$$\begin{aligned}\lambda^{(\nu)} \equiv\ & \textit{maximize } \gamma^s \\ & \textit{subject to } s \in \mathbb{N} \equiv \{0,1,2,\dots\} \\ & \quad \textit{with} \quad F[\boldsymbol{u}^{(\nu)} + \gamma^s \boldsymbol{w}^{(\nu)}] - F^{(\nu)} \leq \eta\gamma^s \left\langle \nabla F^{(\nu)},\ \boldsymbol{w}^{(\nu)} \right\rangle.\end{aligned}$$

Step 3: *(Update) Set* $\boldsymbol{u}^{(\nu+1)} \equiv \boldsymbol{u}^{(\nu)} + \lambda^{(\nu)}\boldsymbol{w}^{(\nu)}$ *and* $\nu = \nu + 1$. *Return to Step* 1.

If ∇F *is globally Lipschitz continuous, then the sequence* $\{\boldsymbol{u}^{(\nu)}\}$ *satisfies*

$$\left\langle \nabla F^{(\nu)},\ \boldsymbol{w}^{(\nu)} \right\rangle \to 0.$$

In particular, if $\boldsymbol{w}^{(\nu)}$ *is chosen so that*

$$\boldsymbol{w}^{(\nu)} = -\frac{\tilde{c}}{\|\nabla F^{(\nu)}\|}\nabla F^{(\nu)}$$

for $0 < \tilde{c} \leq c$, *then*

$$\|\nabla F^{(\nu)}\| \to 0.$$

Since for noisy examples it is not known a priori what the optimal value of the objective is, the algorithm in Theorem 6.1 provides the norm of the gradient as a suitable exit criterion.

6.2. Acceleration Techniques: Limited Memory BFGS with Trust Regions. Ideally, for a twice differentiable function one would want to use Newton's method near the solution. In "nonparametric" techniques, however, explicit calculation of the Hessian is often impossible. For example, if the objective $F[\boldsymbol{u}]$ is discretized into a pixel basis for a 512×512 image, the number of unknowns is $2*2^{18}$. The corresponding Hessian, assuming it exists, is a dense $2^{19} \times 2^{19}$ matrix. Limited memory methods provide an efficient way to use approximate Hessian information without explicitly forming the matrix. These methods are derived from matrix secant methods that approximate curvature information of the objective function from preceding steps and gradients. Limited memory methods are made robust with the introduction of trust regions. For a thorough treatment of matrix secant and trust region methods, see [33].

Denote the discretized unknown functions $\boldsymbol{u}$ by the same variable with the two dimensions stacked into one column vector, that is, $\boldsymbol{u} \in \mathbb{R}^n$ for some integer n. Matrix secant iterates are generated by

$$\boldsymbol{u}^{(\nu+1)} = \boldsymbol{u}^{(\nu)} - \left(M^{(\nu)}\right)^{-1} \nabla F^{(\nu)}, \tag{83}$$

where $M^{(\nu)} \in \mathbb{R}^{n\times n}$ is an approximation to $\nabla^2 F^{(\nu)}$ satisfying the matrix secant equation:

$$M^{(\nu)}(\boldsymbol{u}^{(\nu-1)} - \boldsymbol{u}^{(\nu)}) = \nabla F^{(\nu-1)} - \nabla F^{(\nu)}. \tag{84}$$

Infinitely many solutions are possible since (84) is a system of n equations in n^2 unknowns. Common choices for the secant approximation $M^{(\nu)}$ are Broyden's update, the symmetric-rank-one (SR1) update, and the Broyden–Fletcher–Goldfarb–Shanno

(BFGS) update. Limited memory techniques for BFGS matrices are reviewed here; however, similar methods for alternative updates are possible.

The BFGS update to the true Hessian is given by

$$M^{(\nu)} = M^{(\nu-1)} + \frac{\boldsymbol{y}^{(\nu)}\boldsymbol{y}^{(\nu)^T}}{\boldsymbol{y}^{(\nu)^T}\boldsymbol{s}^{(\nu)}} - \frac{M^{(\nu-1)}\boldsymbol{s}^{(\nu)}\boldsymbol{s}^{(\nu)^T}M^{(\nu-1)}}{\boldsymbol{s}^{(\nu)^T}M^{(\nu-1)}\boldsymbol{s}^{(\nu)}},\ \nu = 1, 2, \ldots,$$

where

$$\boldsymbol{y}^{(\nu)} \equiv \nabla F^{(\nu+1)} - \nabla F^{(\nu)}, \qquad \boldsymbol{s}^{(\nu)} \equiv \boldsymbol{u}^{(\nu+1)} - \boldsymbol{u}^{(\nu)}. \tag{85}$$

The BFGS approximation is symmetric and positive definite as long as $\boldsymbol{s}^{(\nu)^T}\boldsymbol{y}^{(\nu)} > 0$ and $M^{(\nu-1)}$ is symmetric and positive definite.

Define

$$S^{(\nu)} \equiv [\boldsymbol{s}^{(\nu-m)}, \ldots, \boldsymbol{s}^{(\nu-1)}] \in \mathbb{R}^{n\times m} \quad \text{and} \quad Y^{(\nu)} \equiv [\boldsymbol{y}^{(\nu-m)}, \ldots, \boldsymbol{y}^{(\nu-1)}] \in \mathbb{R}^{n\times m}.$$

Limited memory techniques involve generating *at each iteration* the BFGS matrix from the m most recent of the pairs $\{\boldsymbol{y}_i, \boldsymbol{s}_i\}_{i=1}^{\nu-1}$ and generating matrix $M^{(0,\nu)}$. Typically $m \in [5, 10]$. The choice of $M^{(0,\nu)}$ most often used is $M^{(0,\nu)} = \mu^{(\nu)}I$, where I is the identity matrix and $\mu^{(\nu)}$ is some scaling (see [108]). With this generating matrix, limited memory BFGS (L-BFGS) is equivalent to doing m steps of conjugate gradient at each iteration. It can be shown that the complexity of calculating the L-BFGS update (83) is on the order of $mn + m^3$. See [22] for details.

Acceptance of the step to the next iterate depends on the accuracy of the quadratic approximation

$$\widetilde{F}^{(\nu+1)} = F^{(\nu)} + \nabla F^{(\nu)^T} \cdot \boldsymbol{s}^{(\nu)} + \frac{1}{2}\boldsymbol{s}^{(\nu)^T}M^{(\nu)}\boldsymbol{s}^{(\nu)} \tag{86}$$

against the true function value $F^{(\nu+1)}$. A measurement of this accuracy is given by the ratio of the actual change in the function value between iterates $\boldsymbol{u}^{(\nu)}$ and $\boldsymbol{u}^{(\nu+1)}$ and the predicted change,

$$\rho(\boldsymbol{s}^{(\nu)}) = \frac{\text{actual change}^{(\nu)}}{\text{predicted change}^{(\nu)}} = -\frac{F^{(\nu)} + F^{(\nu+1)}}{\nabla F^{(\nu)^T} \cdot \boldsymbol{s}^{(\nu)} + \frac{1}{2}\boldsymbol{s}^{(\nu)^T}M^{(\nu)}\boldsymbol{s}^{(\nu)}}. \tag{87}$$

If the ratio is below some tolerance $\tilde{\eta}$, then the step is restricted. A line search strategy such as the one given in Theorem 6.1 can be employed to find an acceptable step size; however, this often requires several function evaluations. In applications such as nonparametric phase retrieval, function and gradient evaluations are the most expensive part of each iteration, thus we consider alternative strategies for finding acceptable steps. We have found in practice that a single application of a *trust region* strategy is usually all that is required to find a step that satisfies (82). A trust region is a ball around the current iterate $\boldsymbol{u}^{(\nu)}$ within which the quadratic approximation is reliable.

The trust region subproblem with trust region radius $\Delta^{(\nu)}$ is given by

$$TR(\Delta^{(\nu)}) \text{ minimize } \nabla F^{(\nu)^T}\boldsymbol{s} + \frac{1}{2}\boldsymbol{s}^T M^{(\nu)}\boldsymbol{s},$$
$$\|\boldsymbol{s}\| \le \Delta^{(\nu)}.$$

Using compact representations of the L-BFGS approximation derived by Byrd, Nocedal, and Schnabel in [22], Burke and Wiegmann [19] derived a compact representation of the inverse of the matrix $\tau I + M^{(\nu)}$ required to solve the trust region subproblem without actually forming the matrix. This inverse can be computed with the same order of computational complexity as the computation of $\left[M^{(\nu)}\right]^{-1}$. With the proper scaling, $\mu^{(\nu)}$, the trust region is required only a small fraction of the time. This is consistent with observations noted in [19]. The scaling is key to the success of the algorithm. There are many proposals for the scaling $\mu^{(\nu)}$ [93]. We employ the scaling suggested by Shanno and Phua [108]:

$$\mu^{(\nu)} = \frac{{\boldsymbol{y}^{(\nu-1)}}^T \boldsymbol{y}^{(\nu-1)}}{{\boldsymbol{s}^{(\nu-1)}}^T \boldsymbol{y}^{(\nu-1)}}. \tag{88}$$

As in [19], we default to the unconstrained L-BFGS method at the beginning of each iteration, that is, $\Delta_0^{(\nu)} = \infty$ in $TR(\Delta^{(\nu)})$. The trust region is invoked only if the ratio $\rho(\boldsymbol{s}^{(\nu)})$ falls below a given tolerance, indicating that the quadratic model (86) is not reliable.

ALGORITHM 6.2 (Limited Memory BFGS with Trust Regions).

Step 0: *(Initialization) Choose* $\tilde{\eta} > 0$, $\zeta > 0$, $\overline{m} \in \{1, 2, \ldots, n\}$, *and* $\boldsymbol{u}^{(0)} \in \mathbb{R}^n$, *and set* $\nu = m = 0$. *Compute* $\nabla F^{(0)}$, $F^{(0)}$, *and* $\|\nabla F^{(0)}\|$.

Step 1: *(L-BFGS step) If* $m = 0$, *compute* $\boldsymbol{u}^{(\nu+1)}$ *by some line search algorithm (e.g., the algorithm in Theorem* 6.1*); otherwise compute* $\boldsymbol{s}^{(\nu)} = -\left(M^{(\nu)}\right)^{-1} \nabla F^{(\nu)}$, *where* $M^{(\nu)}$ *is the L-BFGS update* [22], $\boldsymbol{u}^{(\nu+1)} = \boldsymbol{u}^{(\nu)} + \boldsymbol{s}^{(\nu)}$, $F^{(\nu+1)}$, *and the predicted change* (86).

Step 2: *(Trust Region) If* $\rho(\boldsymbol{s}^{(\nu)}) < \tilde{\eta}$, *where* ρ *is given by* (87), *reduce the trust region* $\Delta^{(\nu)}$, *solve the trust region subproblem for a new step* $\boldsymbol{s}^{(\nu)}$ [19], *and return to the beginning of Step* 2. *If* $\rho(\boldsymbol{s}^{(\nu)}) \geq \tilde{\eta}$, *compute* $\boldsymbol{u}^{(\nu+1)} = \boldsymbol{u}^{(\nu)} + \boldsymbol{s}^{(\nu)}$ *and* $F^{(\nu+1)}$.

Step 3: *(Update) Compute* $\nabla F^{(\nu+1)}$, $\|\nabla F^{(\nu+1)}\|$, $\boldsymbol{y}^{(\nu)}$ *from* (85), *and* ${\boldsymbol{s}^{(\nu)}}^T \boldsymbol{y}^{(\nu)}$. *Discard the vector pair* $\{\boldsymbol{s}^{(\nu-m)}, \boldsymbol{y}^{(\nu-m)}\}$ *from storage. If* ${\boldsymbol{s}^{(\nu)}}^T \boldsymbol{y}^{(\nu)} \leq \zeta$, *set* $m = \max\{m-1, 0\}$, $\Delta^{(\nu+1)} = \infty$, $\mu^{(\nu+1)} = \mu^{(\nu)}$, *and* $M^{(\nu+1)} = M^{(\nu)}$ *(i.e., shrink the memory and don't update); otherwise set* $\mu^{(\nu+1)} = \frac{{\boldsymbol{y}^{(\nu)}}^T \boldsymbol{y}^{(\nu)}}{{\boldsymbol{s}^{(\nu)}}^T \boldsymbol{y}^{(\nu)}}$, $\Delta^{(\nu+1)} = \infty$, $m = \min\{m+1, \overline{m}\}$, *add the vector pair* $\{\boldsymbol{s}^{(\nu)}, \boldsymbol{y}^{(\nu)}\}$ *to storage, and update* $M^{(\nu+1)}$ [22]. *Set* $\nu = \nu + 1$ *and return to Step* 1.

REMARK 6.3. *With a slight modification, Algorithm* 6.2 *can be used as a backtracking line search algorithm, where* $\overline{m} = 1$ *and* $M^{(\nu)} = \mu^{(\nu)} I$ *for all* ν.

7. Numerical Results. This section details the results of numerical experiments comparing the average performance of line search and L-BFGS methods with projection methods of similar type for noiseless and noisy data for the phase retrieval problem.

The aperture of the pupil consists of seven meter-class panels shown in Figures 6a and 7. This design is one of several configurations being studied at NASA's Goddard Space Flight Center for use on the NGST, Hubble's replacement. To recover the phase, three diversity images shown in Figures 6b–6d are used, two out-of-focus and one in-focus. From this example the advantage of choosing a pixel basis over some parameterization (for example, Zernike polynomials [133, 80, 14, 15]) is apparent. Most obvious is the irregular shape of the pupil and the phase jumps across the separate panels, which make it difficult to find an orthogonal parameterization [114]. Another

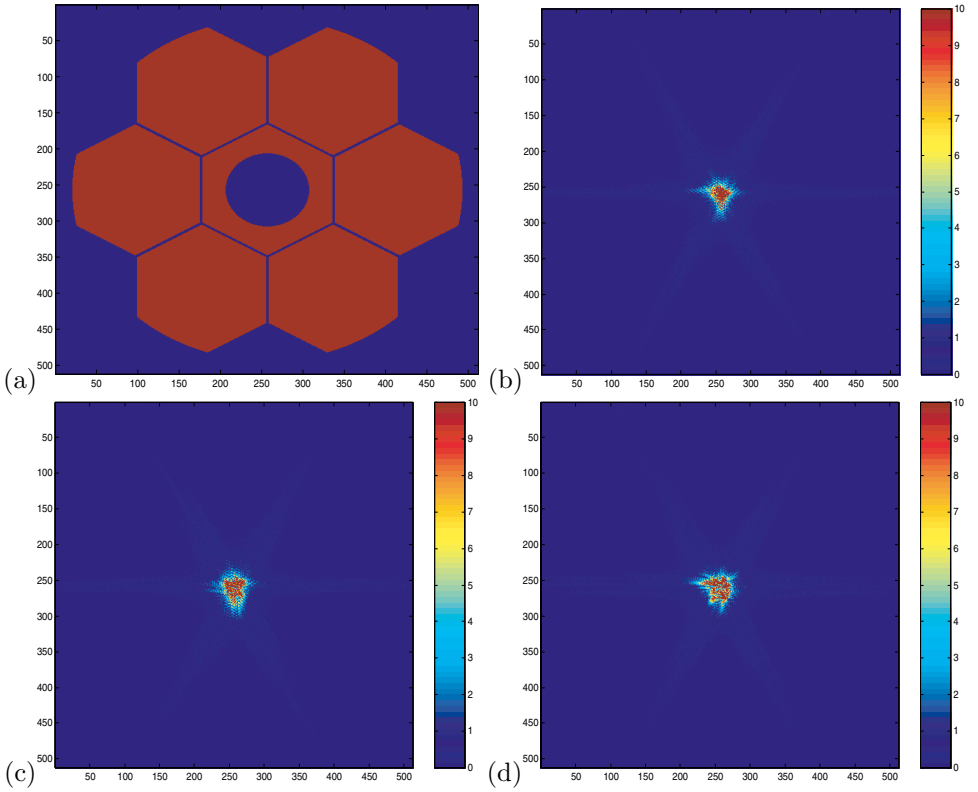

Fig. 6 *Aperture* (a) *and noiseless image data* (b)–(d) *for a segmented pupil on a* 512 *by* 512 *grid. The three diversity images are the optical system's response to a point source at focus and plus/minus defocus, respectively.*

advantage of the pixel basis is that it allows for the most accurate representation of the domain without introducing any regularization implicit in less precise parameterizations. The noisy data shown in Figure 8 generated the results given in Figure 9. Using a pixel basis the methods recover artifacts such as the Gibbs phenomenon associated with the filtering of the data. Issues surrounding filtering and regularization of the data are independent of the numerical method and depend on the types of observations being made [77].

Two projection algorithms are compared to line search and L-BFGS algorithms for the least squares and extended least squares objectives (68) and (79). The first projection algorithm is evenly averaged ($\gamma_m^{(\nu)} = 1/4$ for all ν and $m = 0, \ldots, 3$) and unrelaxed ($\alpha_m = 1$ for all ν and $m = 0, \ldots, 3$) (algorithm (59)). This algorithm is denoted AP for averaged projections. The second projection algorithm is an unrelaxed implementation of algorithm (60), denoted SP for sequential projections. In this implementation the pupil domain projection is computed at every second iterate. This is consistent with higher end implementations that compute the pupil projection more often because it is less computationally expensive than the image domain projections. The projection algorithms are compared to line search algorithms for the evenly weighted least squares measure E_ϵ (LS) and the extended least squares reduced objective R_ϵ (ELS). An additional comparison is made to an L-BFGS trust region

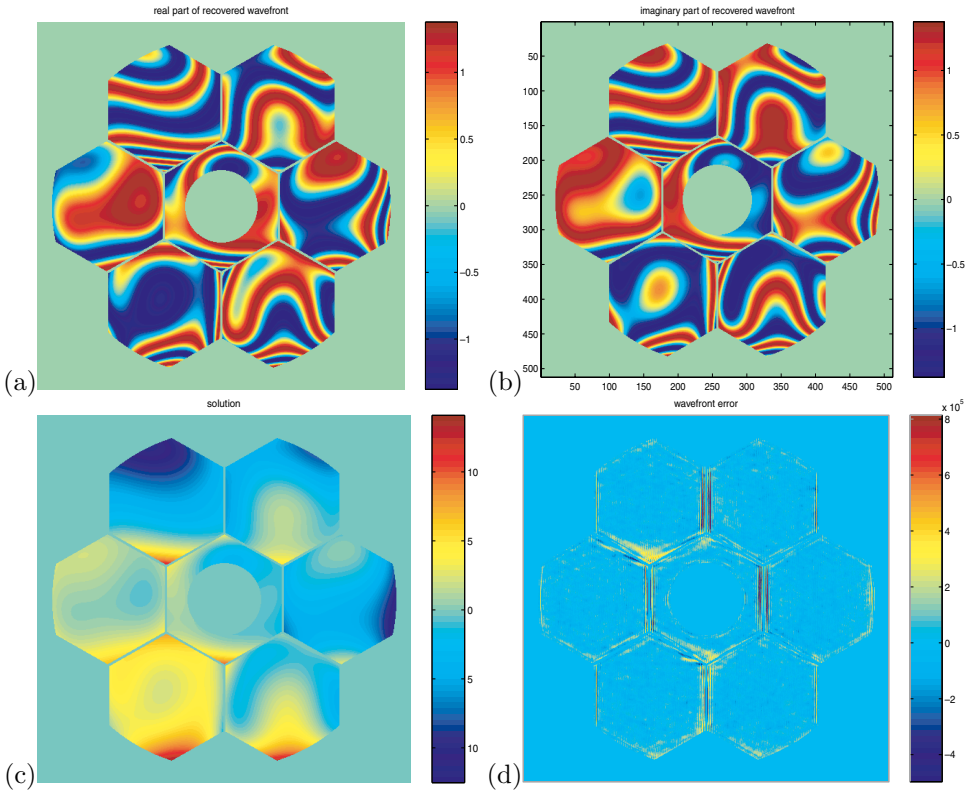

Fig. 7 *Real and imaginary parts,* (a) *and* (b), *respectively, of an aberrated wavefront for the segmented pupil recovered from three noiseless diversity point source images on a* 512 *by* 512 *grid. The wavefront phase is unwrapped* (c) *and compared to the true phase. The wavefront error* (d) *is in units of wavelength.*

algorithm applied to the reduced objective R_ϵ (*L-BFGS*). See Algorithm 6.2 and Remark 6.3. The value of the constants in R_ϵ is taken to be $c_m = 1$ for $m = 0, \ldots, 3$. For the limited memory implementation, a memory length of 4 was chosen.

The formulation of the projections in (58) is numerically unstable. There are several sources of this instability, the most elementary being the possibility of division by zero. In order to achieve a reasonable comparison of computational complexity to line search methods applied to E_ϵ or R_ϵ, the projections are calculated naively as prescribed by (54). We observe that about 6% of the projection runs are exited due to divide-by-zero errors. A second source of instability arises when $\Pi_{\mathbb{Q}_m}[\boldsymbol{u}]$ is multivalued. This is easily remedied by taking a selection $\pi[\boldsymbol{u}; \psi_m, \theta]$ given by (54). While it is unlikely that an iterate will be exactly zero, how one interprets machine zero in this context is an important consideration for numerical stability. In a neighborhood of zero corresponding to machine precision, the phase and amplitude of the estimated wavefront at a grid point $\boldsymbol{u}(\boldsymbol{x}_j)$ are not reliable. If at the same point the data $\psi_m(\boldsymbol{x}_j)$ is relatively large, then, even though the projection $\Pi_{\mathbb{Q}_m}[\boldsymbol{u}]$ is single valued, the error will be amplified. This error amplification could result in stagnation of projection algorithms. About 6% of our trials with projection algorithms resulted in little or no progress from the initial guess. Since the norm of the gradient of a slightly perturbed E in these regions was found to be well away from zero, we attribute this outcome

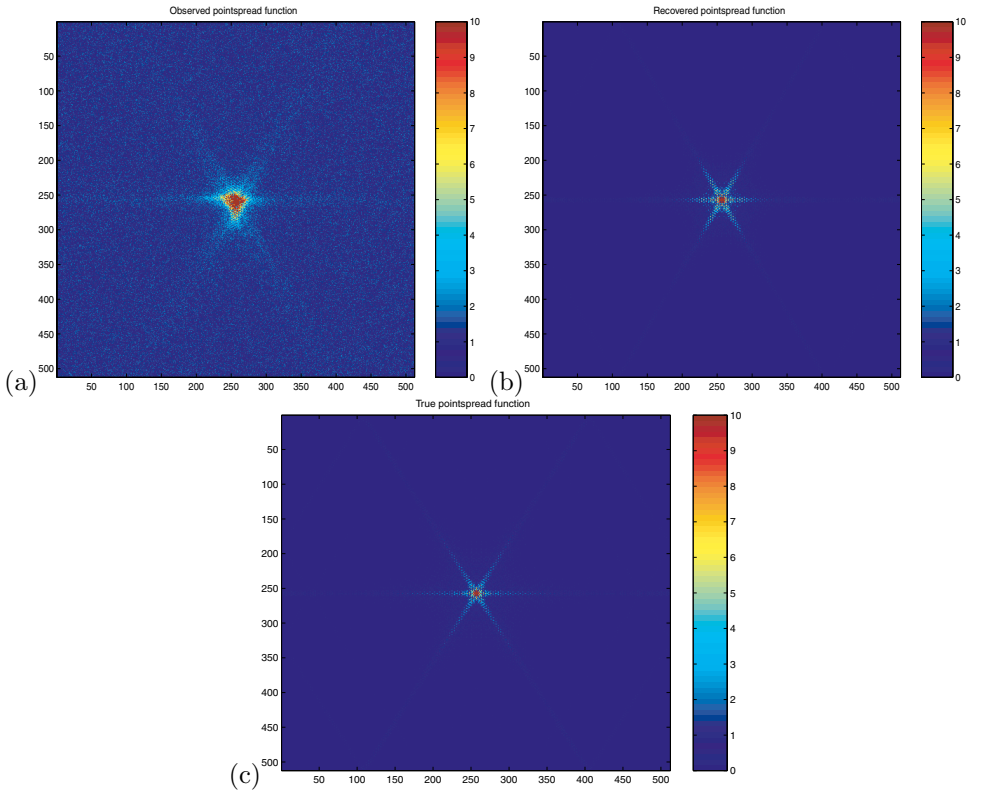

Fig. 8 *Noisy point-spread function* (a) *for a segmented pupil on a* 512 *by* 512 *grid. The recovered point-spread function* (b) *was first filtered with a Fourier window filter before processing by the wavefront reconstruction algorithm. Frame* (c) *shows the true, unaberrated point-spread function.*

to the instability due to phase error amplification. Nonconvergence due to divide-by-zero errors and possible phase error amplification was discounted from the averages computed in Table 1. That is, approximately 12% of the runs for which the projection algorithm fails are not included in Table 1. On the other hand, all of the runs for the analytic algorithms converged and are included in the table.

The behavior of the squared set distance error for a sample run for each of the algorithms is illustrated in Figure 10. Each of the algorithms behaves qualitatively the same, as would be expected. Each spends the majority of time in a flat region where little progress is made until a neighborhood of a solution is found, and error reduction in all cases is rapid. In the flat region the gradient and curvature of the objective are very small. This region corresponds to what is described in projection methods as a "tunnel." The notoriously slow convergence of projection methods is easily understood in terms of the notoriously slow convergence of first-order methods. The limited memory implementation does much better in the flat region, though it too is slowed considerably.

The behavior of the algorithms varies considerably depending on the initialization, hence the average performance of the algorithms over 30 random initial guesses is tabulated in Table 1. The initial guesses all have unit magnitude in the pupil domain with random phase uniformly distributed on $[0, 2\pi]$. In Table 1 average cpu

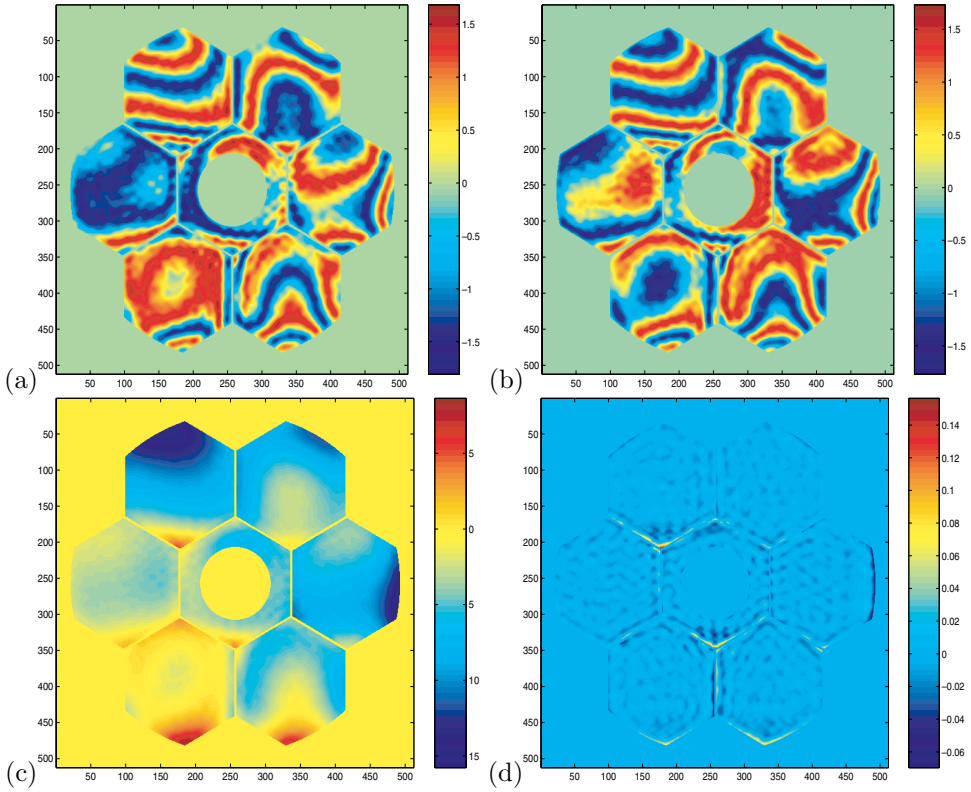

Fig. 9 *The real and imaginary parts,* (a) *and* (b)*, respectively, of the aberrated wavefront for the segmented pupil recovered from three filtered noisy diversity point source images on a* 512 *by* 512 *grid. The wavefront phase* (c) *is unwrapped and compared to the true phase. The wavefront error* (d) *is in units of wavelength. The ridges in the wavefront error are due to the Gibbs phenomenon associated with the noise filter.*

Table I *Relative cpu time of projection and analytic algorithms averaged over* 30 *random trials with the least squares (LS) algorithm as baseline. Outliers are not included in the totals for algorithms with a* $*$*.*

	No noise $E \leq 20e^{-9}$			Noise $E \leq 0.05$			$E \leq 0.0138$		
	Mean	Low	High	Mean	Low	High	Mean	Low	High
LS	248	99	970	161	68	483	222	159	518
AP*	2.29	99	1680	2.7	126	1765	2.3	162	1808
SP*	.96	72	591	1.19	35	746	—	—	—
ELS	.66	74	365	.77	35	258	.84	76	304
L-BFGS	.29	41	196	.44	37	159	.47	72	182

times, along with maxima and minima of the experiments, are compared using the least squares (LS) algorithm as a baseline; the results for the other algorithms are normalized by the least squares performance given at the far left of the table. The standard deviations reflect the robustness of the algorithm and consistency of per-

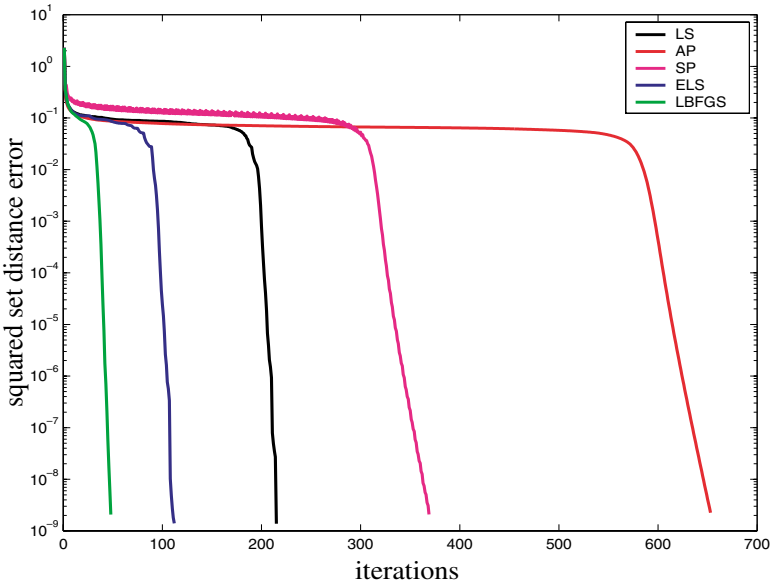

Fig. 10 *Comparison of algorithms applied to the example shown in Figure 7. The algorithms have different objectives, so we compare the behavior of the squared set distance error for each.*

formance. Note that the averaged projections (AP) algorithm is analogous to taking $\epsilon = 0$ in the objective E_ϵ defined by (68). The difference between this and a gradient descent algorithm is that the step length is not optimized. Overflow problems and instability aside, the difference between the performance of AP and least squares can be regarded as a measure of the value of the step length in algorithm performance. With the exception of the SP algorithm, on average each algorithm requires the same number of function evaluations per iteration. The limiting calculation for this application is the Fourier transform, which is accomplished with the fast Fourier transform (FFT) algorithm. Each squared set distance error evaluation requires one FFT per diversity image. Each gradient or projection calculation requires two FFTs per diversity image. The SP algorithm requires three fewer FFTs per iteration than the line search or AP algorithms, since only one projection is calculated at each iteration. Hence the per iteration cost of the SP algorithm is 0.6 times that of the other algorithms. For L-BFGS and least squares implementations, when the trust region is invoked or when backtracking is required to generate the proper step size, additional function evaluations are needed. When the trust region is restricted, usually only one restriction is necessary when the scaling (88) is used. For backtracking, usually three backtracking steps are required. The added computational cost for implementing limited memory methods is not noticeable in cpu time. The average time per iteration for L-BFGS methods is 1.047 seconds for a 512×512 image using a parallel cluster of 16 processors, compared to 1.017 seconds for line search methods. There is, however, a considerable difference in the memory requirements depending on how many previous steps are stored.

The performance of the algorithms on apodized (i.e., filtered) noisy data shown in Figure 8 is very similar in character to the noiseless experiments. Since the methods use a pixel basis, all of the algorithms attempt to match the data exactly, including the noise. Filtering for data analysis is treated as a separate issue from filtering for

numerical efficiency or stability. While it has been noted that other noise models are more appropriate [95], the noise in these experiments is additive and normally distributed, consistent with the least squares performance measure. The squared set distance error $E = 0.050$ is the outer edge of the neighborhood of the solution, i.e., the "knee" in the error reduction shown in Figure 10. Once inside this neighborhood, error reduction is rapid in all cases. With the exception of the SP algorithm, error reduction flattens out at $E = 0.0138$. In every trial the SP algorithm fails to reduce the error below $E = 0.02$. In practice, however, this difference between the SP "solution" and that of the other algorithms does not result in noticeable differences in the eyeball norm for the phase estimate.

8. Conclusion. At their very best, projection methods will behave as well as comparable line search methods. Until now the relaxation parameters $\alpha_m^{(\nu)}$ and weights $\gamma_m^{(\nu)}$ for iterative transform algorithms such as (60) have been chosen in an ad hoc manner independent of any performance criterion. Thus, traditional iterative transform algorithms could behave much worse than line search methods for which suitable parameters reinterpreted as $\lambda^{(\nu)}$ and $\beta_m^{(\nu)}$ in (61) have been extensively studied.

In section 3 of this work we provided a derivation of the optical phase retrieval problem from first principles. In section 4 we reviewed geometric solution techniques, among which are iterative transform techniques. Many fundamental questions regarding convergence of projection algorithms remain. When the intersection of even convex constraints is empty, convergence is an open question. This is often the case in image processing with noisy data. When algorithms stagnate it is impossible to tell if the method has found a local solution or is stuck in what is often referred to as a tunnel. We noted that extensions to projection algorithms have been proposed to overcome stagnation [43]. These methods seem to be very robust and efficient in practice [116]. Their success warrants precise mathematical analysis, which has yet to be done. In section 5 we reviewed analytic perturbation approaches to the problem and quantified their relationship to geometric methods. Two performance measures were considered, and their associated optimization problems were formulated in (66) and (79). The first measure is a perturbed weighted least squares measure. The second is a new approach, which we call extended least squares. This objective allows us to adaptively correct for the relative variability in the diversity measurements, ψ_m. In section 6 we reviewed two numerical methods. The first was a standard line search algorithm for which convergence to first-order necessary conditions for optimality was proven for the perturbed least squares and extended least squares objectives. The line search method is accelerated by a limited memory approach, which allows us to efficiently approximate curvature information in large problems. The use of limited memory techniques for phase retrieval and deconvolution has appeared in recent work [76, 122]. The method is made robust with a novel use of explicit trust regions. The trust region strategy also allows for precise scaling of the step size, thus avoiding costly function evaluations that are common to more trial-and-error–type methods such as implicit trust regions and backtracking. The resulting algorithm was given as Algorithm 6.2. In section 7 we compared the performance of the different approaches on noiseless and noisy data. The results indicate that while certain implementations of iterative transform algorithms can be competitive (see the SP algorithm), their performance varies more from one example to the next than the algorithms based on analytic techniques. Other implementations of the iterative transform algorithm such as AP are clearly not competitive approaches. Limited memory and trust region techniques reduce the variability of performance without adding significant computational cost.

Further cpu speed-up is possible with the introduction of multiresolution techniques as discussed in [76, 75, 90]. These are similar to windowing techniques used for noise filtering. In tests with MATLAB we have achieved 17-fold speed-up in time to convergence with the use of these techniques. With optimal parallelization and multiresolution techniques we expect that the per iteration cpu time for a cluster of 16 PCs with three 512×512 diversity images could be brought down, conservatively, to a tenth of a second.

The extended least squares approach presented in section 5.3 has great potential for future research. In our implementations we chose the simplest possible regularizing functional in (76), that is, $G_m[\boldsymbol{u}] = \text{const}$. Even this simple choice had a dramatic effect on the performance of the algorithms. This opens the door to a search for an optimal $G_m[\boldsymbol{u}]$. There are two different ways to interpret $G_m[\boldsymbol{u}]$; the first and perhaps most natural is statistical, the second is purely algorithmic. Under the statistical interpretation, $G_m[\boldsymbol{u}]$ is viewed as the variance or spatial correlation of the data sets. The method is very general and applies to a wide variety of observations and statistical models. Under the algorithmic interpretation, $G_m[\boldsymbol{u}]$ is a regularizing term in a penalty function and can be used to tackle the problem of algorithm stagnation in the middle iterations (see Figure 10). The adaptive weighting strategy allows one to include several different metrics in the same objective, one that is more effective for the middle regions and one that is more effective near a local solution.

Other directions for research include partial function evaluation algorithms similar to the SP algorithms discussed in section 4. The trust region methodology reviewed in section 6.2 is a first step to stably implementing this strategy. Regularization techniques are also central to numerical methods for solving the more general problem of *simultaneous* wavefront reconstruction *and* deconvolution, known as the *phase diversity* problem. Here, both the wavefront aberration as well as the field source, or object, are unknown. The theory developed here is intended as a starting point for numerical solutions to both the phase retrieval problem and the more general phase diversity problem.

Appendix A. Properties of Constraint Sets.

Proof of Property 4.1. First we show that the set $\mathbb{Q}_0$ is not convex. If $\boldsymbol{u}$ belongs to $\mathbb{Q}_0$, then so does $\boldsymbol{u}' = -\boldsymbol{u}$. Thus for any nontrivial convex combination of $\boldsymbol{u}$ and $\boldsymbol{u}'$,

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

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

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

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

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

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

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

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

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

$$\mathcal{R}[\hat{\boldsymbol{u}}_n] = [\mathcal{R}(\boldsymbol{u}(\boldsymbol{x}))\exp[-2\pi i \boldsymbol{n}\cdot\boldsymbol{x}]]^\wedge = \mathcal{R}[\hat{\boldsymbol{u}}](\boldsymbol{\xi} + \boldsymbol{n}).$$

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

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

By the Riemann–Lebesgue lemma [66, p. 297],

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

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

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

Appendix B. Analytic Properties of the Perturbed Objective. In the proofs that follow, it suffices to consider integral functionals of the form J given in (70).

THEOREM B.1. *Let $f : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ satisfy the following:*

(1) *$f(\cdot, u(\cdot))$ is integrable on $\mathbb{R}^n$ for all $u(\cdot) \in L^2[\mathbb{R}^n, \mathbb{R}^m]$;*

(2) *for all $x \in \mathbb{R}^n$, $f(x, u)$ is Gâteaux differentiable with respect to u as a function on $\mathbb{R}^n \times \mathbb{R}^m$ with Gâteaux derivative denoted by*

$$D_u f(x, u);$$

(3) *there exists a K such that for all x, $D_u f(x, \cdot)$ is globally Lipschitz on $\mathbb{R}^m$ with Lipschitz constant K.*

Define the integral functional $J : L^2[\mathbb{R}^n, \mathbb{R}^m] \to \mathbb{R}$ by

$$J[u] = \int_{\mathbb{R}^n} f(x, u(x))\, dx.$$

Then $J[u]$ is Fréchet differentiable as a function on $L^2[\mathbb{R}^n, \mathbb{R}^m]$ with Fréchet derivative

$$J'[u][w] = \int_{\mathbb{R}^n} D_u f(x, u(x))(w(x)) dx. \tag{89}$$

Moreover, the Fréchet derivative J' is Lipschitz continuous on $L^2[\mathbb{R}^n, \mathbb{R}^m]$ with constant K.

Proof.

$$\left| J[u+w] - J[u] - \int_{\mathbb{R}^n} D_u f(x, u(x))(w(x)) dx \right|$$

$$\leq \int_{\mathbb{R}^n} |f(x, u(x) + w(x)) - f(x, u(x)) - D_u f(x, u(x))(w(x))|\, dx.$$

For fixed x,

$$|f(x, u(x) + w(x)) - f(x, u(x)) - D_u f(x, u(x))(w(x))|$$

$$\leq \int_0^1 |D_u f(x, u(x) + \tau w(x))(w(x)) - D_u f(x, u(x))(w(x))|\ d\tau$$

$$\leq \int_0^1 |D_u f(x, u(x) + \tau w(x)) - D_u f(x, u(x))|\ |w(x)|\ d\tau.$$

Since $D_u f$ is globally Lipschitz continuous with constant K, for all u and x

$$|D_u f(x, u(x) + \tau w(x)) - D_u f(x, u(x))| \leq K\tau \, |w(x)| \, ,$$

thus

$$\begin{aligned}\int_0^1 |D_u f(x, u(x) + \tau w(x))(w(x)) - D_u f(x, u(x))(w(x))| \, d\tau &\leq \int_0^1 K\tau |w(x)|^2 d\tau \\ &= \frac{K}{2}|w(x)|^2,\end{aligned}$$

and hence

$$\begin{aligned}\left| J[u + w] - J[u] - \int_{\mathbb{R}^n} D_u f(x, u(x))(w(x)) dx \right| &\leq \int_{\mathbb{R}^n} \frac{K}{2}|w(x)|^2 dx \\ &= \frac{K}{2}\|w\|^2.\end{aligned}$$

Consequently, J is Fréchet differentiable with $J'[u][w]$ given by (89).

Since L^2 is a Hilbert space, the kernel of the integral operator $J'[u]$ is equal to $D_u f(\cdot, u(\cdot))$. Thus if $D_u f(x, u(x))$ is globally Lipschitz with respect to u with constant K for all x, then $J'[u]$ is globally Lipschitz with constant K. □

REMARK B.2. *Conditions* (2) *and* (3) *in Theorem* B.1 *imply that, for all* $x \in \mathbb{R}^n$, *the integrand* $f(x, u)$ *is Fréchet differentiable with respect to* u *as a function on* $\mathbb{R}^n \times \mathbb{R}^m$. *It is not true in general that Gâteaux differentiability implies Fréchet differentiability. See* [27, Ex. 1.11.20] *for a counterexample. Moreover, it is not true in general that a Fréchet differentiable function has a globally Lipschitz continuous Fréchet derivative. We state Theorem* B.1 *in terms of Gâteaux differentiable functions instead of Fréchet differentiable functions because it is often easier to show Gâteaux differentiability than it is to show Fréchet differentiability.*

REMARK B.3. *Since* $L^2[\mathbb{R}^n, \mathbb{R}^m]$ *is a Hilbert space, the derivative of* $J[u]$ *also belongs to* $L^2[\mathbb{R}^n, \mathbb{R}^m]$. *We denote this mapping by* $\nabla J[u] = D_u f(\cdot, u(\cdot))$.

Denote the space of linear mappings from $\mathbb{R}^2$ to $\mathbb{R}$ by $\mathcal{L}(\mathbb{R}^2, \mathbb{R})$. From elementary vector calculus, the function $r^2(\boldsymbol{u}; b, \epsilon)$ defined by (69) is Gâteaux differentiable with respect to $\boldsymbol{u}$ for $\epsilon > 0$. In fact, $r^2(\boldsymbol{u}; b, \epsilon)$ is *analytic*. The derivative is given by

$$D_{\boldsymbol{u}} r^2(\boldsymbol{u}; b, \epsilon) \equiv 2r(\boldsymbol{u}; b, \epsilon) D_{\boldsymbol{u}} r(\boldsymbol{u}; b, \epsilon), \tag{90}$$

where

$$D_{\boldsymbol{u}} r(\boldsymbol{u}; b, \epsilon) = \frac{|\boldsymbol{u}|^2 + 2\epsilon^2}{(|\boldsymbol{u}|^2 + \epsilon^2)^{3/2}} \boldsymbol{u}^T. \tag{91}$$

The next lemma shows that $D_{\boldsymbol{u}} r^2(\boldsymbol{u}; b, \epsilon)$ is globally Lipschitz continuous.

LEMMA B.4. *The derivative* $D_{\boldsymbol{u}} r^2$ *for* r *defined by* (69) *is globally Lipschitz continuous on* $\mathbb{R}^2$ *with global Lipschitz constant*

$$K = 16 + \frac{12}{\epsilon}|b|. \tag{92}$$

Proof. The setting here is finite dimensional. The finite-dimensional norm is assumed to be the 2-norm and is denoted by $|\cdot|$. From (90)–(91), for $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^2$,

$$\begin{aligned} D(r^2(\boldsymbol{u}; b, \epsilon)) &= 2\frac{\left(|\boldsymbol{u}|^2 + 2\epsilon^2\right) r(\boldsymbol{u}; b, \epsilon)}{\left(|\boldsymbol{u}|^2 + \epsilon^2\right)^{3/2}}\boldsymbol{u} \\ &= 2\left[\frac{|\boldsymbol{u}|^2\boldsymbol{u}}{|\boldsymbol{u}|^2+\epsilon^2} + \epsilon^2\frac{|\boldsymbol{u}|^2\boldsymbol{u}}{(|\boldsymbol{u}|^2+\epsilon^2)^2} - \frac{b\boldsymbol{u}}{(|\boldsymbol{u}|^2+\epsilon^2)^{1/2}} - \epsilon^2\frac{b\boldsymbol{u}}{(|\boldsymbol{u}|^2+\epsilon^2)^{3/2}}\right]. \end{aligned} \tag{93}$$

We proceed by calculating the Lipschitz constant for each of the terms in (93). Each of these terms takes the form

$$\frac{|\boldsymbol{u}|^p\boldsymbol{u}}{(|\boldsymbol{u}|^2+\epsilon^2)^q}.$$

The Lipschitz constant is obtained by bounding terms of the form

$$\left|\frac{|\boldsymbol{u}|^p\boldsymbol{u}}{(|\boldsymbol{u}|^2+\epsilon^2)^q} - \frac{|\boldsymbol{v}|^p\boldsymbol{v}}{(|\boldsymbol{v}|^2+\epsilon^2)^q}\right|.$$

Add and subtract $\frac{|\boldsymbol{v}|^p\boldsymbol{u}}{(|\boldsymbol{v}|^2+\epsilon^2)^q}$ to obtain

$$\begin{aligned} \left|\frac{|\boldsymbol{u}|^p\boldsymbol{u}}{(|\boldsymbol{u}|^2+\epsilon^2)^q} - \frac{|\boldsymbol{v}|^p\boldsymbol{v}}{(|\boldsymbol{v}|^2+\epsilon^2)^q}\right| &= \left|\left(\frac{|\boldsymbol{u}|^p}{(|\boldsymbol{u}|^2+\epsilon^2)^q} - \frac{|\boldsymbol{v}|^p}{(|\boldsymbol{v}|^2+\epsilon^2)^q}\right)\boldsymbol{u} + \frac{|\boldsymbol{v}|^p}{(|\boldsymbol{v}|^2+\epsilon^2)^q}(\boldsymbol{u}-\boldsymbol{v})\right| \\ &\le \left|\frac{|\boldsymbol{u}|^p(|\boldsymbol{v}|^2+\epsilon^2)^q - |\boldsymbol{v}|^p(|\boldsymbol{u}|^2+\epsilon^2)^q}{(|\boldsymbol{u}|^2+\epsilon^2)^q(|\boldsymbol{v}|^2+\epsilon^2)^q}\boldsymbol{u}\right| + \frac{|\boldsymbol{v}|^p}{(|\boldsymbol{v}|^2+\epsilon^2)^q}|\boldsymbol{v}-\boldsymbol{u}|. \end{aligned}$$

Unfortunately this general form must be analyzed case by case. We examine three different cases.

Case 1. $p = 2$, $q = 1$:

$$\begin{aligned} \left|\frac{|\boldsymbol{u}|^2\boldsymbol{u}}{|\boldsymbol{u}|^2+\epsilon^2} - \frac{|\boldsymbol{v}|^2\boldsymbol{v}}{|\boldsymbol{v}|^2+\epsilon^2}\right| &\le \left|\frac{|\boldsymbol{u}|^2(|\boldsymbol{v}|^2+\epsilon^2) - |\boldsymbol{v}|^2(|\boldsymbol{u}|^2+\epsilon^2)}{(|\boldsymbol{u}|^2+\epsilon^2)(|\boldsymbol{v}|^2+\epsilon^2)}\boldsymbol{u}\right| + \frac{|\boldsymbol{v}|^2}{|\boldsymbol{v}|^2+\epsilon^2}|\boldsymbol{v}-\boldsymbol{u}| \\ &\le \epsilon^2\left|\frac{|\boldsymbol{u}|^2-|\boldsymbol{v}|^2}{(|\boldsymbol{u}|^2+\epsilon^2)^{1/2}(|\boldsymbol{v}|^2+\epsilon^2)}\right| + |\boldsymbol{u}-\boldsymbol{v}|, \end{aligned}$$

where we have used the inequality

$$\left|\frac{\boldsymbol{u}}{(|\boldsymbol{u}|^2+\epsilon^2)^{1/2}}\right| \le 1. \tag{94}$$

Without loss of generality assume that $|\boldsymbol{v}| \le |\boldsymbol{u}|$. Then

$$|\boldsymbol{u}|^2 - |\boldsymbol{v}|^2 \le 2|\boldsymbol{u}|\ |\boldsymbol{v}-\boldsymbol{u}| \quad \text{for} \quad |\boldsymbol{v}| \le |\boldsymbol{u}|. \tag{95}$$

Using this, inequality (94), and

$$\frac{1}{|\boldsymbol{v}|^2+\epsilon^2} \le \frac{1}{\epsilon^2} \tag{96}$$

yields the bound

$$\left|\frac{|\boldsymbol{u}|^2\boldsymbol{u}}{|\boldsymbol{u}|^2+\epsilon^2}-\frac{|\boldsymbol{v}|^2\boldsymbol{v}}{|\boldsymbol{v}|^2+\epsilon^2}\right| \le \left(\frac{2|\boldsymbol{u}|\epsilon^2}{(|\boldsymbol{u}|^2+\epsilon^2)^{1/2}(|\boldsymbol{v}|^2+\epsilon^2)}+1\right)|\boldsymbol{v}-\boldsymbol{u}|$$

$$\le 3|\boldsymbol{v}-\boldsymbol{u}|. \tag{97}$$

Case 2. $p=2$, $q=2$: As in case 1, assume without loss of generality that $|\boldsymbol{v}|\le|\boldsymbol{u}|$. Then the inequalities (94)–(96) yield

$$\left|\frac{|\boldsymbol{u}|^2\boldsymbol{u}}{(|\boldsymbol{u}|^2+\epsilon^2)^2}-\frac{|\boldsymbol{v}|^2\boldsymbol{v}}{(|\boldsymbol{v}|^2+\epsilon^2)^2}\right| \le \left|\frac{|\boldsymbol{u}|^2(|\boldsymbol{v}|^2+\epsilon^2)^2-|\boldsymbol{v}|^2(|\boldsymbol{u}|^2+\epsilon^2)^2}{(|\boldsymbol{u}|^2+\epsilon^2)^2(|\boldsymbol{v}|^2+\epsilon^2)^2}\boldsymbol{u}\right|+\frac{|\boldsymbol{v}|^2}{(|\boldsymbol{v}|^2+\epsilon^2)^2}|\boldsymbol{v}-\boldsymbol{u}|$$

$$=\left|\frac{(|\boldsymbol{u}|^2|\boldsymbol{v}|^2-\epsilon^4)(|\boldsymbol{v}|^2-|\boldsymbol{u}|^2)}{(|\boldsymbol{u}|^2+\epsilon^2)^2(|\boldsymbol{v}|^2+\epsilon^2)^2}\boldsymbol{u}\right|+\frac{|\boldsymbol{v}|^2}{(|\boldsymbol{v}|^2+\epsilon^2)^2}|\boldsymbol{v}-\boldsymbol{u}|$$

$$\le\left(\frac{2|\boldsymbol{u}|^2\left(|\boldsymbol{u}|^2|\boldsymbol{v}|^2+\epsilon^2\right)}{(|\boldsymbol{u}|^2+\epsilon^2)^2(|\boldsymbol{v}|^2+\epsilon^2)^2}+\frac{1}{\epsilon^2}\right)|\boldsymbol{v}-\boldsymbol{u}|$$

$$\le\left(\frac{2}{\epsilon^2}\frac{\left(|\boldsymbol{u}|^2|\boldsymbol{v}|^2+\epsilon^2\right)}{(|\boldsymbol{u}|^2+\epsilon^2)(|\boldsymbol{v}|^2+\epsilon^2)}+\frac{1}{\epsilon^2}\right)|\boldsymbol{v}-\boldsymbol{u}|$$

$$\le\frac{5}{\epsilon^2}|\boldsymbol{v}-\boldsymbol{u}|. \tag{98}$$

Case 3. $p=0$, $q=n/2$:

$$\left|\frac{\boldsymbol{u}}{(|\boldsymbol{u}|^2+\epsilon^2)^{n/2}}-\frac{\boldsymbol{v}}{(|\boldsymbol{v}|^2+\epsilon^2)^{n/2}}\right|$$

$$\le\left|\frac{(|\boldsymbol{v}|^2+\epsilon^2)^{n/2}-(|\boldsymbol{u}|^2+\epsilon^2)^{n/2}}{(|\boldsymbol{u}|^2+\epsilon^2)^{n/2}(|\boldsymbol{v}|^2+\epsilon^2)^{n/2}}\boldsymbol{u}\right|+\frac{1}{(|\boldsymbol{v}|^2+\epsilon^2)^{n/2}}|\boldsymbol{u}-\boldsymbol{v}|$$

$$\le\left|\frac{\left(|\boldsymbol{v}|^2+\epsilon^2\right)^{n/2}-\left(|\boldsymbol{u}|^2+\epsilon^2\right)^{n/2}}{(|\boldsymbol{u}|^2+\epsilon^2)^{(n-1)/2}\left(|\boldsymbol{v}|^2+\epsilon^2\right)^{n/2}}\right|+\frac{1}{\epsilon^n}|\boldsymbol{u}-\boldsymbol{v}|.$$

The last expression uses inequalities (94) and (96). By the mean value theorem there exists a $w\in[|\boldsymbol{v}|,|\boldsymbol{u}|]$ such that

$$\left(|\boldsymbol{v}|^2+\epsilon^2\right)^{n/2}-\left(|\boldsymbol{u}|^2+\epsilon^2\right)^{n/2}=nw\left(w^2+\epsilon^2\right)^{n/2-1}(|\boldsymbol{v}|-|\boldsymbol{u}|)$$

and so

$$\left|\left(|\boldsymbol{v}|^2+\epsilon^2\right)^{n/2}-\left(|\boldsymbol{u}|^2+\epsilon^2\right)^{n/2}\right|\le nw\left(w^2+\epsilon^2\right)^{n/2-1}|\boldsymbol{v}-\boldsymbol{u}|$$

$$\le n\left(w^2+\epsilon\right)^{(n-1)/2}|\boldsymbol{v}-\boldsymbol{u}|.$$

This yields

$$\left|\frac{\left(|\boldsymbol{v}|^2+\epsilon^2\right)^{n/2}-\left(|\boldsymbol{u}|^2+\epsilon^2\right)^{n/2}}{(|\boldsymbol{u}|^2+\epsilon^2)^{(n-1)/2}\left(|\boldsymbol{v}|^2+\epsilon^2\right)^{n/2}}\right|\le\left|\frac{n|\boldsymbol{v}-\boldsymbol{u}|\left(w^2+\epsilon\right)^{(n-1)/2}}{(|\boldsymbol{u}|^2+\epsilon^2)^{(n-1)/2}(|\boldsymbol{v}|^2+\epsilon^2)^{n/2}}\right|$$

$$\le\frac{n}{\epsilon^n}|\boldsymbol{v}-\boldsymbol{u}|.$$

Finally,

$$\left| \frac{\boldsymbol{u}}{(|\boldsymbol{u}|^2+\epsilon^2)^{n/2}} - \frac{\boldsymbol{v}}{(|\boldsymbol{v}|^2+\epsilon^2)^{n/2}} \right| \leq \frac{n+1}{\epsilon^n}|\boldsymbol{v}-\boldsymbol{u}|. \tag{99}$$

The bounds (97)–(98) and the bound (99) for $n = 1, 3$ yield the following global bound, which completes the proof:

$$\left|Dr^2(\boldsymbol{u};b,\epsilon) - Dr^2(\boldsymbol{v};b,\epsilon)\right| \leq \left(16 + \frac{12}{\epsilon}|b|\right)|\boldsymbol{u}-\boldsymbol{v}|. \qquad \square$$

The constant K given in (92) is the pointwise Lipschitz constant for the Gâteaux derivative of the functional $r^2(\boldsymbol{u}(\boldsymbol{x});b(\boldsymbol{x}),\epsilon)$. If $b \in L^\infty$, then for all $\boldsymbol{x}$

$$\left|D_{\boldsymbol{u}}r^2(\boldsymbol{u}(\boldsymbol{x});b(\boldsymbol{x}),\epsilon) - D_{\boldsymbol{u}}r^2(\boldsymbol{v}(\boldsymbol{x});b(\boldsymbol{x}),\epsilon)\right| \leq \left(16 + \frac{12}{\epsilon}\|b\|_\infty\right)|\boldsymbol{u}(\boldsymbol{x}) - \boldsymbol{v}(\boldsymbol{x})|. \tag{100}$$

We can therefore apply Theorem B.1 to the integral operator J defined by (70) for $\epsilon > 0$ to obtain the Fréchet derivative

$$J'[\boldsymbol{u};b,\epsilon][\boldsymbol{w}] = \int_{\mathbb{R}^2} \left(D_{\boldsymbol{u}}r^2(\boldsymbol{u}(\boldsymbol{x});b(\boldsymbol{x}),\epsilon),\ \boldsymbol{w}(\boldsymbol{x})\right) d\boldsymbol{x},$$

where $(\cdot,\ \cdot)$ denotes the standard finite-dimensional inner product. Equations (69), (90), and (91) yield the gradient of J at $\boldsymbol{u}$

$$\nabla J[\boldsymbol{u};b,\epsilon] = 2\left(\frac{|\boldsymbol{u}|^2}{(|\boldsymbol{u}|^2+\epsilon^2)^{1/2}} - b\right)\frac{|\boldsymbol{u}|^2+2\epsilon^2}{(|\boldsymbol{u}|^2+\ \epsilon^2)^{3/2}}\boldsymbol{u}. \tag{101}$$

By Lemma B.4, (100), and Theorem B.1, $\nabla J[\boldsymbol{u};b,\epsilon]$ is globally Lipschitz continuous with global Lipschitz constant

$$K_{\nabla J} = \left(16 + \frac{12}{\epsilon}\|b\|_\infty\right). \tag{102}$$

The preceding results extend immediately to the perturbed squared set distance error $E_\epsilon[\boldsymbol{u}]$ defined by (71). Since $\mathcal{F}_m[\boldsymbol{u}]$ defined by (47) and (48) is a linear operator on L^2, it is Fréchet differentiable there with Fréchet derivative given by

$$\mathcal{F}_m[\boldsymbol{u}]'[\boldsymbol{w}] = \mathcal{F}_m[\boldsymbol{w}].$$

For $\boldsymbol{u}$ and ψ_m satisfying Hypothesis 3.1, Theorem B.1 together with the chain rule for Fréchet differentiable functions and (101) yields (73):

$$\begin{aligned}
\left(J[\cdot;\psi_m,\epsilon]\circ\mathcal{F}_m[\boldsymbol{u}]\right)'[\boldsymbol{w}] &= J'[\mathcal{F}_m[\boldsymbol{u}];\psi_m,\epsilon][\mathcal{F}_m'[\boldsymbol{u}][\boldsymbol{w}]] \\
&= \langle \nabla J[\mathcal{F}_m[\boldsymbol{u}];\psi_m,\epsilon], \mathcal{F}_m[\boldsymbol{w}]\rangle \\
&= 2\left\langle \mathcal{F}_m^*\left[r[\mathcal{F}_m[\boldsymbol{u}];\psi_m,\epsilon]\frac{|\mathcal{F}_m[\boldsymbol{u}]|^2+2\epsilon^2}{(|\mathcal{F}_m[\boldsymbol{u}]|^2+\ \epsilon^2)^{3/2}}\mathcal{F}_m[\boldsymbol{u}]\right],\ \boldsymbol{w}\right\rangle
\end{aligned}$$

for $m = 0, \ldots, M$. Thus

$$\nabla\left(J[\cdot;\psi_m,\epsilon]\circ\mathcal{F}_m[\boldsymbol{u}]\right) = 2\mathcal{F}^*\left[r[\mathcal{F}_m[\boldsymbol{u}];\psi_m,\epsilon]\frac{|\mathcal{F}_m[\boldsymbol{u}]|^2+2\epsilon^2}{(|\mathcal{F}_m[\boldsymbol{u}]|^2+\ \epsilon^2)^{3/2}}\mathcal{F}_m[\boldsymbol{u}]\right]. \tag{103}$$

Extending this to $E_\epsilon[\boldsymbol{u}]$ we have

$$E'_\epsilon[\boldsymbol{u}][\boldsymbol{w}] = \langle \nabla E_\epsilon[\boldsymbol{u}],\ \boldsymbol{w} \rangle ,$$

where, by (72) and (103),

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

which yields (74). By Parseval's relation, (102), and the triangle inequality, the global Lipschitz constant $K_{\nabla E_\epsilon}$ for $\nabla E_\epsilon[\boldsymbol{u}]$ is

$$K_{\nabla E_\epsilon} = \sum_{m=0}^{M} \beta_m \left(8 + \frac{6\|\psi_m\|_\infty}{\epsilon} \right).$$

Similarly, the extended least squares objective $R_\epsilon[\boldsymbol{u}]$ defined by (78) is Fréchet differentiable with derivative given by

$$R'_\epsilon[\boldsymbol{u}][\boldsymbol{w}] = \sum_{m=0}^{M} \left(J[\mathcal{F}_m[\boldsymbol{u}]; \psi_m, \epsilon] + c_m \right)^{-1} \left(J[\cdot; \psi_m, \epsilon] \circ \mathcal{F}_m \right)'[\boldsymbol{u}][\boldsymbol{w}].$$

And so, by (73), for $\nabla \left(J[\cdot; \psi_m, \epsilon] \circ \mathcal{F}_m \right)[\boldsymbol{u}]$ given by (103),

$$\nabla R_\epsilon[\boldsymbol{u}] = \sum_{m=0}^{M} \left(\left(J[\cdot; \psi_m, \epsilon] \circ \mathcal{F}_m \right)[\boldsymbol{u}] + c_m \right)^{-1} \nabla \left(J[\cdot; \psi_m, \epsilon] \circ \mathcal{F}_m \right)[\boldsymbol{u}],$$

which yields (80). Together with the fact that $\ln(x + c_m)$ has a derivative bounded by $1/c_m$ on $\mathbb{R}_+$, (102) yields the global Lipschitz constant $K_{\nabla R_\epsilon}$ for ∇R_ϵ,

$$K_{\nabla R_\epsilon} = \sum_{m=0}^{M} \frac{1}{c_m} \left(16 + \frac{12\|\psi_m\|_\infty}{\epsilon} \right).$$

Theorem 5.1 gives a bound on the distance between the projection and the gradient of the perturbed objective in the phase retrieval problem. But first, we prove that the perturbed objective is a continuous function of ϵ.

LEMMA B.5. *Let $\boldsymbol{u} \in L^2[\mathbb{R}^2, \mathbb{R}^2]$ and $b \in \mathbb{U}_+$ defined in Hypothesis* 3.1. *The integral functional $J[\boldsymbol{u}; b, \epsilon]$ defined by* (70) *is a continuous function of ϵ.*

Proof. Let $\boldsymbol{u} \in L^2[\mathbb{R}^2, \mathbb{R}^2]$, $b \in \mathbb{U}_+$. From (70)–(69),

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

Since $\boldsymbol{u}$ and b satisfy Hypothesis 3.1, by Hölder's inequality for all ϵ,

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

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

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

Proof of Theorem 5.1. The theorem follows from careful splitting of the norm and repeated application of Lebesgue's dominated convergence theorem. Define

$$\mathbb{G}_m = \operatorname{supp}(\mathcal{F}_m[\boldsymbol{u}]), \qquad m = 0, 1, \ldots.$$

Denote the complements of these sets by $\widetilde{\mathbb{G}}_m$. Denote the norm over the domain $\Omega \subset \mathbb{R}^2$ by

$$\|\cdot\|_\Omega \equiv \|\cdot \mathcal{X}_\Omega\|,$$

where $\mathcal{X}_\Omega$ is the indicator function for Ω defined by (20). Let $\boldsymbol{v}_m \in \Pi_{\mathbb{Q}_m}[\boldsymbol{u}]$, $m = 0, 1, 2, \ldots$, and let $\boldsymbol{v} = \sum_{m=0}^{M} \beta_m(\boldsymbol{u} - \boldsymbol{v}_m)$. Then

$$\begin{aligned}
\|\nabla E_\epsilon[\boldsymbol{u}] - \boldsymbol{v}\| &\le \sum_{m=0}^{M} \left\| \frac{\beta_m}{2} \nabla J[\mathcal{F}_m[\boldsymbol{u}]; \psi_m, \epsilon] - \beta_m(\boldsymbol{u} - \boldsymbol{v}_m) \right\| \\
&= \sum_{m=0}^{M} \beta_m \left\| \mathcal{F}_m^* \left[r\left[\mathcal{F}_m[\boldsymbol{u}]; \psi_m, \epsilon\right] \nabla r\left[\mathcal{F}_m[\boldsymbol{u}]; \psi_m, \epsilon\right] \mathcal{F}_m[\boldsymbol{u}] \right] - (\boldsymbol{u} - \boldsymbol{v}_m) \right\| \\
&= \sum_{m=0}^{M} \beta_m \left\| r\left[\mathcal{F}_m[\boldsymbol{u}]; \psi_m, \epsilon\right] \nabla r\left[\mathcal{F}_m[\boldsymbol{u}]; \psi_m, \epsilon\right] \mathcal{F}_m[\boldsymbol{u}] - \mathcal{F}_m[\boldsymbol{u} - \boldsymbol{v}_m] \right\| \\
&= \sum_{m=0}^{M} \beta_m \left\| r\left[\mathcal{F}_m[\boldsymbol{u}]; \psi_m, \epsilon\right] \nabla r\left[\mathcal{F}_m[\boldsymbol{u}]; \psi_m, \epsilon\right] \mathcal{F}_m[\boldsymbol{u}] - \mathcal{F}_m[\boldsymbol{u} - \boldsymbol{v}_m] \right\|_{\mathbb{G}_m} \\
&\qquad + \beta_m \|\mathcal{F}_m[\boldsymbol{v}_m]\|_{\widetilde{\mathbb{G}}_m}.
\end{aligned}$$

Now, by the definition of $\mathbb{Q}_m$ (50), $|\mathcal{F}_m[\boldsymbol{v}_m]| = \psi_m$. Also, on $\mathbb{G}_m[\boldsymbol{u}]$ we have $\mathcal{F}_m[\boldsymbol{v}_m] = \frac{\mathcal{F}_m[\boldsymbol{u}]}{|\mathcal{F}_m[\boldsymbol{u}]|}\psi_m$, which yields the inequality

$$\begin{aligned}
\|\nabla E_\epsilon[\boldsymbol{u}] - \boldsymbol{v}\| \le \sum_{m=0}^{M} \beta_m \Big\| & r\left[\mathcal{F}_m[\boldsymbol{u}]; \psi_m, \epsilon\right] \nabla r\left[\mathcal{F}_m[\boldsymbol{u}]; \psi_m, \epsilon\right] |\mathcal{F}_m[\boldsymbol{u}]| \\
& - \left(|\mathcal{F}_m[\boldsymbol{u}]| - \psi_m\right) \Big\|_{\mathbb{G}_m} + \beta_m \|\psi_m\|_{\widetilde{\mathbb{G}}_m}. \qquad (104)
\end{aligned}$$

Note that this bound is achieved *for any* $\boldsymbol{v}_m \in \Pi_{\mathbb{Q}_m}[\boldsymbol{u}]$, $m = 0, 1, 2, \ldots$.

Now, by assumption $E < \delta$, which yields the following bound on the rightmost term of (104):

$$\begin{aligned}
&\sum_{m=0}^{M} \frac{\beta_m}{2} \|\psi_m\|_{\widetilde{\mathbb{G}}_m}^2 < \delta \\
\implies &\frac{\beta_m}{2} \|\psi_m\|_{\widetilde{\mathbb{G}}_m}^2 < \delta \\
\implies &\|\psi_m\|_{\widetilde{\mathbb{G}}_m} < \sqrt{\frac{2}{\beta_m}\delta} \\
\implies &\sum_{m=0}^{M} \beta_m \|\psi_m\|_{\widetilde{\mathbb{G}}_m} < (2\delta)^{1/2} \sum_{m=0}^{M} \beta_m^{1/2}. \qquad (105)
\end{aligned}$$

For the remaining terms of (104) consider any $a \in L^2[\mathbb{R}^2, \mathbb{R}_+]$ and $b \in \mathbb{U}_+$ satisfying $\|a-b\|^2 < \delta$. Let

$$\mathbb{G} = \operatorname{supp}(a) \qquad \text{and} \qquad \mathbb{G}_\epsilon = \left\{ \boldsymbol{x} \;\middle|\; a(\boldsymbol{x}) > \sqrt{\epsilon} \right\}.$$

The remaining norms in (104) take the form

$$\left\| \left(\frac{a^2}{(a^2 + \epsilon^2)^{1/2}} - b \right) \frac{a^3 + 2a\epsilon^2}{(a^2 + \epsilon^2)^{3/2}} + (b-a) \right\|_{\mathbb{G}}$$

$$\leq \left\| \frac{a\epsilon^4}{(a^2+\epsilon^2)^2} \right\| + \left\| \left(1 - \frac{a^3 + 2a\epsilon^2}{(a^2 + \epsilon^2)^{3/2}} \right) b \right\|_{\mathbb{G}}. \tag{106}$$

Consider the first norm on the right-hand side of (106):

$$\left\| \frac{a\epsilon^4}{(a^2+\epsilon^2)^2} \right\| \leq \left\| \frac{a\epsilon^4}{(a^2+\epsilon^2)^2} \right\|_{\mathbb{B}(\frac{1}{\sqrt{\epsilon}})} + \left\| \frac{a\epsilon^4}{(a^2+\epsilon^2)^2} \right\|_{\widetilde{\mathbb{B}}(\frac{1}{\sqrt{\epsilon}})},$$

where $\mathbb{B}(\frac{1}{\sqrt{\epsilon}})$ is the ball of radius $\frac{1}{\sqrt{\epsilon}}$. The argument of the norm over the interior of $\mathbb{B}(\frac{1}{\sqrt{\epsilon}})$ is bounded by $\frac{a\epsilon^4}{(a^2+\epsilon^2)^2} \leq \epsilon$, thus

$$\left\| \frac{a\epsilon^4}{(a^2+\epsilon^2)^2} \right\|_{\mathbb{B}(\frac{1}{\sqrt{\epsilon}})} \leq \sqrt{\pi\epsilon}.$$

The norm over the complement $\widetilde{\mathbb{B}}(\frac{1}{\sqrt{\epsilon}})$ cannot be bounded by ϵ without an additional assumption that a has compact support. However, since $a \in L^2$ the norm can be made arbitrarily small, i.e., given ϵ' there is an $\epsilon_0 > 0$ such that

$$\left\| \frac{a\epsilon^4}{(a^2+\epsilon^2)^2} \right\|_{\widetilde{\mathbb{B}}(\frac{1}{\sqrt{\epsilon}})} \leq \epsilon' \quad \forall \epsilon \geq \epsilon_0.$$

Thus

$$\left\| \frac{a\epsilon^4}{(a^2+\epsilon^2)^2} \right\| \leq \sqrt{\pi\epsilon} + \epsilon' \quad \forall \epsilon \geq \epsilon_0. \tag{107}$$

Next consider the rightmost norm of (106). Rearranging terms yields

$$\frac{a^3 + 2a\epsilon^2}{(a^2 + \epsilon^2)^{3/2}} = \frac{a}{(a^2+\epsilon^2)^{1/2}} \left(1 + \frac{\epsilon^2}{a^2+\epsilon^2} \right).$$

From this it is clear that for all a and ϵ,

$$0 \leq \frac{a^2}{a^2+\epsilon^2} \leq \frac{a^3 + 2a\epsilon^2}{(a^2 + \epsilon^2)^{3/2}} \leq \left(1 + \frac{\epsilon^2}{a^2+\epsilon^2} \right)^2 \leq 2.$$

Define

$$g(\alpha, \epsilon) = \left| 1 - \frac{\alpha^3 + 2\alpha\epsilon^2}{(\alpha^2 + \epsilon^2)^{3/2}} \right|.$$

For all (α, ϵ), we have $0 \leq g(\alpha, \epsilon) \leq 1$. Indeed, for all (α, ϵ),

$$\begin{aligned} g(\alpha,\epsilon) &\leq \max\left[1 - \frac{\alpha^2}{\alpha^2+\epsilon^2},\ \left(1+\frac{\epsilon^2}{a^2+\epsilon^2}\right)^2 - 1\right] \\ &= \max\left[\frac{\epsilon^2}{\alpha^2+\epsilon^2},\ \frac{\epsilon^2}{\alpha^2+\epsilon^2}\left(\frac{2\alpha^2+3\epsilon^2}{a^2+\epsilon^2}\right)\right] \\ &\leq 5\frac{\epsilon^2}{\alpha^2+\epsilon^2}. \end{aligned}$$

On the interval $\alpha \in [\sqrt{\epsilon}, \infty)$ we have $\frac{\epsilon^2}{\alpha^2+\epsilon^2} \leq \frac{\epsilon^2}{\epsilon+\epsilon^2} \leq \epsilon$ and

$$g(\alpha,\epsilon) \leq 5\epsilon \quad \forall\ \alpha \in [\sqrt{\epsilon}, \infty).$$

Thus, given $\epsilon' > 0$, there is an $\epsilon > 0$ such that

$$\left\| \left(1 - \frac{a^3+2a\epsilon^2}{(a^2+\ \epsilon^2)^{3/2}}\right) b \right\|_{\mathbb{G}_\epsilon} \leq 5\epsilon\|b\| \leq \epsilon'. \tag{108}$$

On $\widetilde{\mathbb{G}}_\epsilon \cap \mathbb{G}$, from the above, we have that

$$\left\| \left(1 - \frac{a^3+2a\epsilon^2}{(a^2+\ \epsilon^2)^{3/2}}\right) b \right\|_{\widetilde{\mathbb{G}}_\epsilon \cap \mathbb{G}} \leq \|b\|_{\widetilde{\mathbb{G}}_\epsilon \cap \mathbb{G}}\ .$$

Since $\|a-b\|^2 < \delta$,

$$\|b\|_{\widetilde{\mathbb{G}}_\epsilon \cap \mathbb{G}} < \|a\|_{\widetilde{\mathbb{G}}_\epsilon \cap \mathbb{G}} + \delta^{1/2} = \|a\mathcal{X}_{\widetilde{\mathbb{G}}_\epsilon \cap \mathbb{G}}\| + \delta^{1/2}.$$

Since the function $a\mathcal{X}_{\widetilde{\mathbb{G}}_\epsilon \cap \mathbb{G}}$ converges pointwise to zero as $\epsilon \to 0$, we obtain from the Lebesgue dominated convergence theorem that

$$\lim_{\epsilon\to 0} \|a\|_{\widetilde{\mathbb{G}}_\epsilon \cap \mathbb{G}} = \lim_{\epsilon \to 0} \|a\mathcal{X}_{\widetilde{\mathbb{G}}_\epsilon \cap \mathbb{G}}\| = 0.$$

Hence there exists an $\epsilon' > 0$ such that for all $\epsilon \in [0, \epsilon']$,

$$\|b\|_{\widetilde{\mathbb{G}}_\epsilon \cap \mathbb{G}} < \delta^{1/2}. \tag{109}$$

Without applying additional constraints on a the bound of (109) cannot be made tighter.

Letting $\epsilon' = \delta^{1/2}$ in (107)–(108) and substituting the bounds (105)–(109) into (104) completes the proof. □

Proof of Corollary 5.3. The single-valuedness $\mathcal{G}[\boldsymbol{u}]$ follows directly from the definition of the projections. To prove the next statement of the corollary, note that the only terms on the right-hand side of (104) that could not be made arbitrarily small were the terms with bounds (105) and (109). With the assumptions of the corollary, these bounds are much tighter. Indeed, since the support of ψ_m is contained in the support of $\mathcal{F}_m[\boldsymbol{u}]$, the bound in (105) is zero since

$$\|\psi_m\|_{\widetilde{\mathbb{G}}_m} = 0.$$

For the bound (109), define

$$\mathbb{G}_{m,\epsilon} = \left\{ \boldsymbol{\xi} \;\middle|\; |\mathcal{F}_m[\boldsymbol{u}]|(\boldsymbol{\xi}) > \sqrt{\epsilon} \right\}.$$

As usual, denote the complement of this set by $\widetilde{\mathbb{G}}_{m,\epsilon}$. Since $\tilde{\psi}_m \in L^\infty[\mathbb{R}^2, \mathbb{R}_+]$ with $\tilde{\psi}_m = \frac{\psi_m}{|\mathcal{F}_m[\boldsymbol{u}]|}\mathcal{X}_{\psi_m}$, then

$$\begin{aligned}\|\psi_m\|_{\widetilde{\mathbb{G}}_{m,\epsilon}\cap\mathbb{G}_m} &= \left\| \tilde{\psi}_m |\mathcal{F}_m[\boldsymbol{u}]| \right\|_{\widetilde{\mathbb{G}}_{m,\epsilon}\cap\mathbb{G}_m} \\ &\leq \left\| \tilde{\psi}_m \mathcal{X}_{\widetilde{\mathbb{G}}_{m,\epsilon}} \right\|_\infty \|\mathcal{F}_m[\boldsymbol{u}]\|_{\widetilde{\mathbb{G}}_{m,\epsilon}\cap\mathbb{G}_m}.\end{aligned}$$

As in the proof of the bound (109), we have that

$$\lim_{\epsilon\to 0} \|\mathcal{F}_m[\boldsymbol{u}]\|_{\widetilde{\mathbb{G}}_{m,\epsilon}\cap\mathbb{G}_m} = \lim_{\epsilon\to 0} \|\mathcal{F}_m[\boldsymbol{u}]\mathcal{X}_{\widetilde{\mathbb{G}}_{m,\epsilon}\cap\mathbb{G}_m}\| = 0.$$

Hence there exists an ϵ such that for any $\delta > 0$,

$$\|\psi_m\|_{\widetilde{\mathbb{G}}_{m,\epsilon}\cap\mathbb{G}_m} < \delta. \qquad \square$$

All the pieces are in place now to prove Theorem 6.1.

Proof of Theorem 6.1. The proof is by contradiction. Suppose there is a subsequence $\mathbb{K} \subset \mathbb{N}$ such that $\sup_{\mathbb{K}} \left\langle \nabla F^{(\nu)}, \boldsymbol{w}^{(\nu)} \right\rangle < \beta < 0$. Since F is bounded below, $F^{(\nu)} \searrow F_* \in \mathbb{R}$, and so $(F^{(\nu+1)} - F^{(\nu)}) \to 0$. By the choice of $\lambda^{(\nu)}$ in Step 2 we have that

$$\lambda^{(\nu)} \left\langle \nabla F^{(\nu)},\ \boldsymbol{w}^{(\nu)} \right\rangle \to 0.$$

Therefore $\lambda^{(\nu)} \underset{\mathbb{K}}{\to} 0$ and so, without loss of generality, $\lambda^{(\nu)} < 1$ for all $\nu \in \mathbb{K}$. Hence,

$$\eta\lambda^{(\nu)}\gamma^{-1}\langle \nabla F^{(\nu)},\ \boldsymbol{w}^{(\nu)} \rangle < F[\boldsymbol{u}^{(\nu)} + \lambda^{(\nu)}\gamma^{-1}\boldsymbol{w}^{(\nu)}] - F^{(\nu)} \tag{110}$$

for all $\nu \in \mathbb{K}$. Let K be the global Lipschitz constant for ∇F. Then

$$F[\boldsymbol{u}^{(\nu)}+\lambda^{(\nu)}\gamma^{-1}\boldsymbol{w}^{(\nu)}]-F^{(\nu)} \leq \lambda^{(\nu)}\gamma^{-1}\left[\langle \nabla F^{(\nu)},\ \boldsymbol{w}^{(\nu)}\rangle + K(\lambda^{(\nu)}\gamma^{-1}\|\boldsymbol{w}^{(\nu)}\|)\right]. \tag{111}$$

Together, (110)–(111) yield

$$0 < (1-\eta)\beta + K(\lambda^{(\nu)}\gamma^{-1}\|\boldsymbol{w}^{(\nu)}\|).$$

Taking limits over $\nu \in \mathbb{K}$,

$$\lambda^{(\nu)}\gamma^{-1}\|\boldsymbol{w}^{(\nu)}\| \to 0 \quad \Longrightarrow \quad K(\lambda^{(\nu)}\gamma^{-1}\|\boldsymbol{w}^{(\nu)}\|) \to 0,$$

which yields the contradiction $0 < (1-\eta)\beta < 0$.

We next show convergence of the norm of the gradient to zero for $\boldsymbol{w}^{(\nu)} = -\frac{\tilde{c}}{\|\nabla F^{(\nu)}\|}\nabla F^{(\nu)}$, where $0 < \tilde{c} \leq c$. This is a direction of descent lying within $c\mathbb{B}$. Thus, for this choice of $\boldsymbol{w}^{(\nu)}$,

$$\left\langle \nabla F^{(\nu)},\ \boldsymbol{w}^{(\nu)} \right\rangle = -\tilde{c}\|\nabla F^{(\nu)}\| \to 0. \qquad \square$$

Acknowledgments. The first author gratefully acknowledges NASA's VSEP program, through which he was first introduced to the phase retrieval problem, and subsequently NASA's GSRP program, which has provided valuable support and stimulating exposure to research in the adaptive optics field. This work is dedicated to the memory of Dr. Gerald A. Soffen, whose boundless enthusiasm and curiosity touched so many young scientists.

REFERENCES

[1] E. J. Akutowicz, *On the determination of the phase of a Fourier integral,* I, Trans. Amer. Math. Soc., 83 (1956), pp. 179–192.

[2] E. J. Akutowicz, *On the determination of the phase of a Fourier integral,* II, Proc. Amer. Math. Soc., 8 (1957), pp. 234–238.

[3] B. E. Allman, P. J. McMahon, K. A. Nugent, D. Paganin, D. Jacobson, M. Arif, and S. A. Werner, *Imaging—phase radiography with neutrons*, Nature, 408 (2000), pp. 158–159.

[4] O. Axelsson and G. Lindskog, *On the rate of convergence of the preconditioned conjugate gradient algorithm*, Numer. Math., 48 (1986), pp. 499–523.

[5] N. Baba and K. Mutoh, *Measurement of telescope aberrations through atmospheric turbulence by use of phase diversity*, Appl. Opt., 40 (2001), pp. 544–552.

[6] R. Barakat and G. Newsam, *Algorithms for reconstruction of partially known, band-limited Fourier-transform pairs from noisy data.* II. *The nonlinear problem of phase retrieval*, J. Integral Equations, 9 (1985), pp. 77–125.

[7] R. Barakat and G. Newsam, *Algorithms for reconstruction of partially known, band-limited Fourier-transform pairs from noisy data.*, J. Opt. Soc. Amer. A, 2 (1985), pp. 2027–2038.

[8] T. K. Barrett and D. G. Sandler, *Artificial neural network for the determination of Hubble Space Telescope aberration from stellar images*, Appl. Opt., 32 (1993), pp. 1720–1727.

[9] H. H. Bauschke, *The composition of finitely many projections onto closed convex sets in Hilbert space is asymptotically regular*, Proc. Amer. Math. Soc., to appear.

[10] H. H. Bauschke and J. M. Borwein, *On the convergence of von Neumann's alternating projection algorithm for two sets*, Set-Valued Anal., 1 (1993), pp. 185–212.

[11] H. H. Bauschke and J. M. Borwein, *On projection algorithms for solving convex feasibility problems*, SIAM Rev., 38 (1996), pp. 367–426.

[12] H. H. Bauschke, J. M. Borwein, and A. S. Lewis, *The method of cyclic projections for closed convex sets in Hilbert space*, in Recent Developments in Optimization Theory and Nonlinear Analysis (Jerusalem, 1995), AMS, Providence, RI, 1997, pp. 1–38.

[13] B. Bell, J. V. Burke, and A. Shumitzky, *A relative weighting method for estimating parameters and variances in multiple data sets*, Comput. Statist. Data Anal., 22 (1996), pp. 119–135.

[14] A. Bhatia and E. Wolf, *On the circle polynomials of Zernike and related orthogonal sets*, Proc. Camb. Phil. Soc., 50 (1954), pp. 40–48.

[15] M. Born and E. Wolf, *Principles of Optics*, 6th ed., Pergamon Press, New York, 1980.

[16] R. H. Boucher, *Convergence of algorithms for phase retrieval from two intensity measurements*, in Proc. SPIE 231, SPIE, Bellingham, WA, 1980, pp. 130–141.

[17] L. M. Brègman, *The method of successive projection for finding a common point of convex sets*, Soviet Math. Dokl., 6 (1965), pp. 688–692.

[18] Y. M. Bruck and L. G. Sodin, *On the ambiguity of the image reconstruction problem*, Opt. Comm., 30 (1979), pp. 304–308.

[19] J. V. Burke and A. Wiegmann, *Low-Dimensional Quasi-Newton Updating Strategies for Large-Scale Unconstrained Optimization*, manuscript.

[20] C. J. Burrows, *Hubble Space Telescope optics status*, in Proc. SPIE 1567, SPIE, Bellingham, WA, 1991, pp. 284–293.

[21] C. J. Burrows, J. A. Holtzman, S. M. Faber, P. Y. Beley, H. Hasan, C. R. Lynds, and D. Schroeder, *The imaging performance of the Hubble Space Telescope*, Astrophys. J., 369 (1991), pp. L21–L25.

[22] R. H. Byrd, J. Nocedal, and R. B. Schnabel, *Representations of quasi-Newton matrices and their use in limited memory methods*, Math. Programming, 63 (1994), pp. 129–156.

[23] R. Carreras, S. Restaino, G. Love, G. Tarr, and J. Fender, *Phase diversity experimental results: Deconvolution of ν Scorpii*, Opt. Comm., 130 (1996), pp. 13–19.

[24] J. N. Cederquist, J. R. Fienup, C. C. Wackerman, S. R. Robinson, and D. Kryskowski, *Wave-front phase estimation from Fourier intensity measurements*, J. Opt. Soc. Amer. A, 6 (1989), pp. 1020–1026.
[25] S. Chrétien and P. Bondon, *Cyclic projection methods on a class of nonconvex sets*, Numer. Funct. Anal. Optim., 17 (1996), pp. 37–56.
[26] F. H. Clarke, *Optimization and Nonsmooth Analysis*, Classics Appl. Math. 5, SIAM, Philadelphia, 1990.
[27] F. H. Clarke, R. J. Stern, Y. S. Ledyaev, and P. R. Wolenski, *Nonsmooth Analysis and Control Theory*, Springer-Verlag, New York, 1998.
[28] P. L. Combettes, *Inconsistent signal feasibility problems: Least-squares solutions in a product space*, IEEE Trans. Signal Process., 42 (1994), pp. 2955–2966.
[29] P. L. Combettes, *The convex feasibility problem in image recovery*, in Advances in Imaging and Electron Physics 95, P. W. Hawkes, ed., Academic Press, New York, 1996, pp. 155–270.
[30] P. L. Combettes and P. Bondon, *Hard-constrained inconsistent signal feasibility problems*, IEEE Trans. Signal Process., 47 (1999), pp. 2460–2468.
[31] P. L. Combettes and H. J. Trussell, *Method of successive projections for finding a common point of sets in metric spaces*, J. Optim. Theory Appl., 67 (1990), pp. 487–507.
[32] J. C. Dainty and J. R. Fienup, *Phase retrieval and image reconstruction for astronomy*, in Image Recovery: Theory and Application, H. Stark, ed., Academic Press, New York, 1987.
[33] J. E. Dennis and R. Schnabel, *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*, Prentice-Hall, Englewood Cliffs, NJ, 1983.
[34] D. Dialetis and E. Wolf, *The phase retrieval problem of coherence theory as a stability problem*, Nuovo Cimento (X), 47 (1967), pp. 113–116.
[35] D. C. Dobson, *Phase reconstruction via nonlinear least squares*, Inverse Problems, 8 (1992), pp. 541–558.
[36] B. Z. Dong, Y. Zhang, B. Y. Gu, and G. Yang, *Numerical investigation of phase retrieval in a fractional Fourier transform*, J. Opt. Soc. Amer. A, 14 (1997), pp. 2709–2714.
[37] A. J. J. Drenth, A. Huiser, and H. Ferwerda, *The problem of phase retrieval in light and electron microscopy of strong objects*, Optica Acta, 22 (1975), pp. 615–628.
[38] N. V. Efimov and S. B. Stechkin, *Approximative compactness and Chebyshev sets*, Soviet Math. Dokl., 2 (1961), pp. 1226–1228.
[39] M. W. Farn, *New iterative algorithm for the design of phase-only gratings*, in Proc. SPIE 1555, SPIE, Bellingham, WA, 1991, pp. 34–42.
[40] J. R. Fienup, *Reconstruction of an object from the modulus of its Fourier transform*, Opt. Lett., 3 (1978), pp. 27–29.
[41] J. R. Fienup, *Space object imaging through the turbulent atmosphere*, Opt. Engrg., 18 (1979), pp. 529–534.
[42] J. R. Fienup, *Iterative method applied to image reconstruction and to computer-generated holograms*, Opt. Engrg., 19 (1980), pp. 297–305.
[43] J. R. Fienup, *Phase retrieval algorithms: A comparison*, Appl. Opt., 21 (1982), pp. 2758–2769.
[44] J. R. Fienup, *Phase retrieval for Hubble Space Telescope using iterative propagation algorithms*, in Applications of Digital Image Processing XIV, A. Tescher, ed., Proc. SPIE 1567, SPIE, Bellingham, WA, 1991, pp. 327–332.
[45] J. R. Fienup, J. Marron, T. Schultz, and J. Seldin, *Hubble Space Telescope characterized by using phase retrieval algorithms*, Appl. Opt., 32 (1993), pp. 1747–1767.
[46] J. R. Fienup and C. C. Wackerman, *Phase retrieval stagnation problems and solutions*, J. Opt. Soc. Amer. A, 3 (1986), pp. 1897–1907.
[47] J. Frank, P. Penczek, R. K. Agrawal, R. A. Grassucci, and A. B. Heagle, *Three-dimensional cryoelectron microscopy of ribosomes*, in RNA-Ligand Interactions. Part A, Methods in Enzymology 317, Daniel W. Celander, ed., Academic Press, San Diego, 2000, pp. 276–291.
[48] R. W. Gerchberg and W. O. Saxton, *A practical algorithm for the determination of phase from image and diffraction plane pictures*, Optik, 35 (1972), pp. 237–246.
[49] R. A. Gonsalves, *Phase retrieval from modulus data*, J. Opt. Soc. Amer., 66 (1976), pp. 961–964.
[50] R. A. Gonsalves, *Phase retrieval and diversity in adaptive optics*, Opt. Engrg., 21 (1982), pp. 829–832.
[51] J. W. Goodman, *Statistical Optics*, Wiley, New York, 1985.
[52] J. W. Goodman, *Introduction to Fourier Optics*, 2nd ed., McGraw-Hill, New York, 1996.

[53] L. Gubin, B. Polyak, and E. Raik, *The method of projections for finding the common point of convex sets*, USSR Comput. Math. and Math. Phys., 7 (1967), pp. 1–24.
[54] T. E. Gureyev and K. A. Nugent, *Phase retrieval with the transport of intensity equation:* II. *Orthogonal series solution for nonuniform illumination*, J. Opt. Soc. Amer. A, 13 (1996), pp. 1670–1682.
[55] T. E. Gureyev, A. Roberts, and K. A. Nugent, *Phase retrieval with the transport of intensity equation: Matrix solution with the use of Zernike polynomials*, J. Opt. Soc. Amer. A, 12 (1995), pp. 1933–1941.
[56] I. Halpern, *The product of projection operators*, Acta Sci. Math., 23 (1962), pp. 96–99.
[57] P. C. Hansen, *Rank-Deficient and Discrete Ill-Posed Problems*, SIAM, Philadelphia, 1998.
[58] M. H. Hayes, *Signal Reconstruction from Phase or Magnitude*, Ph.D. thesis, Massachusetts Institute of Technology, Cambridge, MA, 1981.
[59] M. H. Hayes and A. V. Oppenheim, *Signal reconstruction from phase or magnitude*, IEEE Trans. Acoust. Speech Signal Process., ASSP-28 (1980), pp. 672–680.
[60] G. T. Herman, *Image Reconstruction from Projections—The Fundamentals of Computerized Tomography*, Academic Press, New York, 1980.
[61] A. Huiser and H. Ferwerda, *The problem of phase retrieval in light and electron microscopy of strong objects.* II. *On the uniqueness and stability of object reconstruction procedures using two defocused images*, Optica Acta, 23 (1976), pp. 445–456.
[62] N. E. Hurt, *Phase Retrieval and Zero Crossings*, Kluwer Academic Publishers, Norwell, MA, 1989.
[63] A. D. Ioffe, *Approximate subdifferentials and applications:* II, Mathematika, 33 (1986), pp. 111–128.
[64] A. D. Ioffe, *Approximate subdifferentials and applications:* III, Mathematika, 36 (1989), pp. 1–38.
[65] T. Isernia, G. Leone, R. Pierri, and F. Soldovieri, *Role of support information and zero locations in phase retrieval by a quadratic approach*, J. Opt. Soc. Amer. A, 16 (1999), pp. 1845 – 1856.
[66] F. Jones, *Lebesgue Integration on Euclidean Space*, Jones and Bartlett, Boston, 1993.
[67] Y. Kano and E. Wolf, *Temporal coherence of black body radiation*, Proc. Phys. Soc. (Lond.), 80 (1962), pp. 1273–1276.
[68] R. Kendrick, D. Acton, and A. Duncan, *Phase diversity wave-front sensor for imaging systems*, Appl. Opt., 33 (1994), pp. 6533–6546.
[69] J. E. Krist and C. J. Burrows, *Phase retrieval analysis of pre and post-repair Hubble Space Telescope images*, Appl. Opt., 34 (1995), pp. 4951–4964.
[70] D. J. Lee, M. C. Roggemann, B. M. Welsh, and E. R. Crosby, *Evaluation of least-squares phase-diversity technique for space telescope wave-front sensing*, Appl. Opt., 36 (1997), pp. 9186–9197.
[71] G. Leone, R. Pierri, and F. Soldovieri, *Reconstruction of complex signals from intensities of Fourier transform pairs*, J. Opt. Soc. Amer. A, 13 (1996), pp. 1546–1556.
[72] A. Levi and H. Stark, *Image restoration by the method of generalized projections with application to restoration from magnitude*, J. Opt. Soc. Amer. A, 1 (1984), pp. 932–943.
[73] H. M. Lloyd, S. M. Jefferies, J. R. P. Angel, and E. K. Hege, *Wave-front sensing with time-of-flight phase diversity*, Opt. Lett., 26 (2001), pp. 402–404.
[74] M. G. Lofdahl, G. B. Scharmer, and W. Wei, *Calibration of a deformable mirror and Strehl ratio measurements by use of phase diversity*, Appl. Opt., 39 (2000), pp. 94–103.
[75] D. R. Luke, *Analysis of Wavefront Reconstruction and Deconvolution in Adaptive Optics*, Ph.D. thesis, University of Washington, Seattle, 2001.
[76] D. R. Luke, J. V. Burke, and R. Lyon, *Fast algorithms for phase diversity and phase retrieval*, in Proceedings of the Workshop on Computational Optics and Imaging for Space Applications, NASA/Goddard Space Flight Center, 2000, Optical Society of America.
[77] R. Lyon, *DCATT Wavefront Sensing and Optical Control Study*, Tech. Report WFSC-0001, NASA/Goddard Space Flight Center, 1999.
[78] R. Lyon, J. Dorband, and J. Hollis, *Hubble Space Telescope faint object camera calculated point spread functions*, Appl. Opt., 36 (1997), pp. 1752–1765.
[79] R. Lyon, P. Miller, and A. Grusczak, *Hubble Space Telescope phase retrieval: A parameter estimation*, in Applications of Digital Image Processing XIV, Proc. SPIE 1567, A. Tescher, ed., SPIE, Bellingham, WA, 1991, pp. 317–326.
[80] V. Mahajan, *Zernike annular polynomials for imaging systems with annular pupils*, J. Opt. Soc. Amer., 71 (1981), pp. 75–85.
[81] J. C. Mather, *NGST*, in Proc. SPIE 4013, SPIE, Bellingham, WA, 2000, pp. 2–16.

[82] A. B. Meinel, M. P. Meinel, and D. H. Schulte, *Determination of the Hubble Space Telescope effective conic-constant error from direct image measurements*, Appl. Opt., 32 (1993), pp. 1715–1719.

[83] J. Miao, P. Charalambous, J. Kirz, and D. Sayre, *Extending the methodology of x-ray crystallography to allow imaging of micrometre-sized non-crystalline specimens*, Nature, 400 (1999), pp. 342–344.

[84] R. Millane, *Phase retrieval in crystallography and optics*, J. Opt. Soc. Amer. A., 7 (1990), pp. 394–411.

[85] D. L. Misell, *An examination of an iterative method for the solution of the phase problem in optics and electron optics* I. *Test calculations*, J. Phys. D, 6 (1973), pp. 2200–2216.

[86] B. S. Mordukhovich, *Approximation Methods in Problems of Optimization and Control*, Nauka, Moscow, 1988 (in Russian).

[87] B. S. Mordukhovich and Y. Shao, *Nonsmooth sequential analysis in Asplund spaces*, Trans. Amer. Math. Soc., 328 (1996), pp. 1235–1280.

[88] Z. Mouyan and R. Unbehauen, *Methods for reconstruction of* 2*-d sequences from Fourier transform magnitude*, IEEE Trans. Image Process., 6 (1997), pp. 222–234.

[89] J. Nocedal and S. Wright, *Numerical Optimization*, Springer-Verlag, New York, 2000.

[90] Y. Ohneda, N. Baba, N. Miura, and T. Sakurai, *Multiresolution approach to image reconstruction with phase-diversity technique*, Opt. Rev., 8 (2001), pp. 32–36.

[91] E. L. O'Neill and A. Walther, *The question of phase in image formation*, Optica Acta, 10 (1963), pp. 33–40.

[92] A. Oppenheim and R. Schafer, *Digital Signal Processing*, Prentice-Hall, Englewood Cliffs, NJ, 1975.

[93] S. S. Oren and E. Spedicato, *Optimal conditioning of self-scaling variable metric algorithms*, Math. Programming, 10 (1976), pp. 70–90.

[94] R. G. Paxman, T. J. Schultz, and J. R. Fienup, *Joint estimation of object and aberrations by using phase diversity*, J. Opt. Soc. Amer. A, 9 (1992), pp. 1072–1085.

[95] R. G. Paxman, J. H. Seldin, M. G. Lofdahl, G. B. Scharmer, and C. U. Keller, *Evaluation of phase-diversity techniques for solar-image restoration*, Astrophys. J., 466 (1996), pp. 1087–1099.

[96] R. J. Plemmons and V. P. Pauca, *Some computational problems arising in adaptive optics imaging systems*, J. Comput. Appl. Math., 123 (2000), pp. 467–487.

[97] T. Quatieri and A. Oppenheim, *Iterative techniques for minimum phase signal reconstruction from phase or magnitude*, IEEE Trans. Acoust. Speech Signal Process., ASSP-29 (1981), pp. 1187–1193.

[98] D. Redding, S. Basinger, D. Cohen, A. Lowman, F. Shi, P. Bely, C. Bowers, R. Burg, L. Burns, P. Davila, B. Dean, G. Mosier, T. Norton, P. Petrone, B. Perkins, and M. Wilson, *Wavefront Control for a Segmented Deployable Space Telescope*, Tech. Report, Jet Propulsion Labs, Pasadena, CA, NASA/Goddard Space Flight Center, Greenbelt, MD, and Space Telescope Science Institute, Baltimore, MD, 2000.

[99] D. Redding, S. Basinger, A. Lowman, A. Kissil, P. Bely, R. Burg, R. Lyon, G. Mosier, M. Wilson, M. Femiano, M. Wilson, G. Schunk, L. Craig, D. Jacobson, J. Rakoczy, and J. Hadaway, *Wavefront sensing and control for a Next Generation Space Telescope*, in Proc. SPIE 3356, SPIE, Bellingham, WA, 1998.

[100] D. Redding, B. M. Levine, J. Yu, and J. Wallace, *Hybrid ray-trace and diffraction propagation code for analysis of optical systems*, in Design, Modeling, and Control of Laser Beam Optics, Proc. SPIE 1625, Y. Kohanzadeh, G. N. Lawrence, G. McCoy, and H. Weichel, eds., SPIE, Bellingham, WA, 1992, pp. 95–107.

[101] R. T. Rockafellar and R. J. Wetts, *Variational Analysis*, Springer-Verlag, New York, 1998.

[102] C. Roddier and F. Roddier, *Combined approach to the Hubble Space Telescope wave-front distortion analysis*, Appl. Opt., 32 (1993), pp. 2992–3008.

[103] P. Roman and A. S. Marathay, *Analyticity and phase retrieval*, Nuovo Cimento (X), 30 (1963), pp. 1452–1464.

[104] W. Rudin, *Real and Complex Analysis*, 2nd ed., McGraw-Hill, New York, 1974.

[105] G. B. Scharmer, *Object-independent fast phase-diversity*, Astronomical-Society-of-the-Pacific-Conference-Series, 183 (1999), pp. 330–341.

[106] B. D. Seery and E. P. Smith, *NASA's Next Generation Space Telescope visiting a time when galaxies were young*, in Proc. SPIE 3356, SPIE, Bellingham, WA, 1998, pp. 2–13.

[107] J. H. Seldin and J. R. Fienup, *Numerical investigation of the uniqueness of phase retrieval*, J. Opt. Soc. Amer. A, 7 (1990), pp. 412–27.

[108] D. F. Shanno and K. Phua, *Matrix conditioning and nonlinear optimization*, Math. Programming, 14 (1978), pp. 149–160.
[109] A. Sommerfeld, *Optics*, Academic Press, New York, 1954.
[110] H. Stark, ed., *Image Recovery: Theory and Application*, Academic Press, New York, 1987.
[111] H. Stark and M. I. Sezan, *Image processing using projection methods*, in Real-Time Optical Information Processing, Academic Press, London, UK, 1994, pp. 185–232.
[112] W. J. Stiles, *Closest point maps and their product,* II, Nieuw Archief voor Wiskunde, 13 (1965), pp. 212–225.
[113] J. W. Strutt (Lord Rayleigh), *On the interference bands of approximately homogeneous light; in a letter to Prof. A. Michelson*, Phil. Mag., 34 (1892), pp. 407–411.
[114] W. Swantner and W. W. Chow, *Gram-Schmidt orthonormalization of Zernike polynomials for general aperture shapes*, Appl. Opt., 33 (1994), pp. 1832–1857.
[115] A. Szoke, *Holographic microscopy with a complicated reference*, J. Imaging Sci. Tech., 41 (1997), pp. 332–341.
[116] H. Takajo, T. Shizuma, T. Takahashi, and S. Takahata, *Reconstruction of an object from its noisy Fourier modulus: Ideal estimate of the object to be constructed and a method that attempts to find that object*, Appl. Opt., 38 (1999), pp. 5568–5576.
[117] M. R. Teague, *Deterministic phase retrieval: A Green's function solution*, J. Opt. Soc. Amer., 73 (1983), pp. 1434–1441.
[118] P. van Toorn and H. Ferwerda, *The problem of phase retrieval in light and electron microscopy of strong objects.* III. *Developments of methods for numerical solution*, Optica Acta, 23 (1976), pp. 456–468.
[119] P. van Toorn and H. Ferwerda, *The problem of phase retrieval in light and electron microscopy of strong objects.* IV. *Checking algorithms by means of simulated objects*, Optica Acta, 23 (1976), pp. 468–481.
[120] J. Véran, F. Rigaut, H. Maître, and D. Rouan, *Estimation of the adaptive optics long-exposure point-spread function using control-loop data*, J. Opt. Soc. Amer. A., 14 (1997), pp. 3057–3068.
[121] L. P. Vlasov, *Approximative properties of sets in normed linear spaces*, Russian Math. Surveys, (1973), pp. 1–66.
[122] C. Vogel, *A limited memory BFGS method for an inverse problem in atmospheric imaging*, in Methods and Applications of Inversion, Lecture Notes in Earth Sciences 92, P. C. Hansen, B. Jacobsen, and K. Mosegaard, eds., Springer-Verlag, New York, 2000, pp. 292–304.
[123] C. R. Vogel, T. Chan, and R. Plemmons, *Fast algorithms for phase diversity-based blind deconvolution*, in Adaptive Optical System Technologies, Proc. SPIE 3353, SPIE, Bellingham, WA, 1998.
[124] J. von Neumann, *On rings of operators, reduction theory*, Ann. Math., 50 (1949), pp. 401–485.
[125] A. Walther, *The question of phase retrieval in optics*, Optica Acta, 10 (1963), pp. 41–49.
[126] E. Wolf, *Is a complete determination of the energy spectrum of light possible from measurements of degree of coherence?*, Proc. Phys. Soc. (Lond.), 80 (1962), pp. 1269–1272.
[127] J. W. Wood, M. A. Fiddy, and R. E. Burge, *Phase retrieval using two intensity measurements in the complex plane*, Opt. Lett., 6 (1981), pp. 514–516.
[128] G. Z. Yang, B. Z. Dong, B. Y. Gu, J. Y. Zhuang, and O. K. Ersoy, *Gerchberg-Saxton and Yang-Gu algorithms for phase retrieval in a nonunitary transform system: A comparison*, Appl. Opt., 33 (1994), pp. 209–219.
[129] D. C. Youla, *Mathematical theory of image restoration by the method of convex projections*, in Image Recovery: Theory and Applications, H. Stark, ed., Academic Press, New York, 1987, pp. 29–77.
[130] D. C. Youla and V. Velasco, *Extensions of a result on the synthesis of signals in the presence of inconsistent constraints*, IEEE Trans. Circuits Syst., 33 (1986), pp. 465–468.
[131] D. C. Youla and H. Webb, *Image restoration by the method of convex projections: Part* I - *theory*, IEEE Trans. Med. Im., MI-1 (1982), pp. 81–94.
[132] E. H. Zarantonello, *Projections on convex sets in Hilbert space and spectral theory*, in Contributions to Nonlinear Functional Analysis, E. H. Zarantonello, ed., Academic Press, New York, 1971, pp. 237–424.
[133] F. Zernike, *Beugungstheorie des schneidenverfahrens und seiner verbesserten form, der phasenkontrastmethode*, Physica, 1 (1934), pp. 689–794.