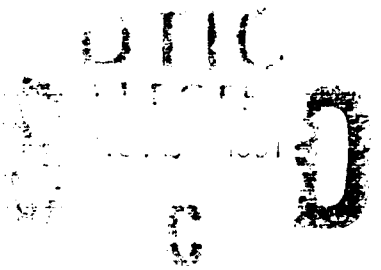
196800-18-F



(2)

HIGH RESOLUTION IMAGING USING PHASE RETRIEVAL

Final Report, Volume II

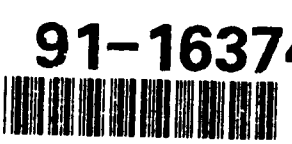
OCTOBER 1991

James R. Fienup
John D. Gorman
John H. Seldin
Jack N. Cederquist

Optical and Infrared Science Laboratory
Advanced Concepts Division
Environmental Research Institute of Michigan
P.O. Box 134001
Ann Arbor, MI 48113-4001

Prepared for:
Office of Naval Research
800 North Quincy Street
Arlington, Virginia 22217-5000
Attn: Dr. Fred W. Quelle

Contract No. N00014-86-C-0587



ERIM P.O. Box 134001
Ann Arbor, MI 48113-4001

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE

REPORT DOCUMENTATION PAGE

Form Approved OMB No. 0704-0188

1a REPORT SECURITY CLASSIFICATION Unclassified	1b RESTRICTIVE MARKINGS
2a SECURITY CLASSIFICATION AUTHORITY	3 DISTRIBUTION/AVAILABILITY OF REPORT Approved for public release; distribution unlimited
2b DECLASSIFICATION/DOWNGRADING SCHEDULE	
4 PERFORMING ORGANIZATION REPORT NUMBERS(S) 196800-18-F	5 MONITORING ORGANIZATION REPORT NUMBER(S)
6a NAME OF PERFORMING ORGANIZATION Environmental Research Institute of Michigan 6b OFFICE SYMBOL *(if applicable)*	7a NAME OF MONITORING ORGANIZATION Office of Naval Research
6c ADDRESS (City, State, and ZIP Code) P.O. Box 134001 Ann Arbor, MI 48113-4001	7b ADDRESS (City, State, and ZIP Code) 800 North Quincy Street Arlington, VA 22217-5000
8a NAME OF FUNDING /SPONSORING ORGANIZATION Office of Naval Research 8b OFFICE SYMBOL *(if applicable)*	9 PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER N00014-86-C-0587
8c ADDRESS (City, State, and ZIP Code) 800 N. Quincy Arlington, VA 22217-5000	10 SOURCE OF FUNDING NUMBERS PROGRAM ELEMENT NO / PROJECT NO / TASK NO / WORK UNIT ACCESSION NO

11 TITLE (Include Security Classification)

High Resolution Imaging Using Phase Retrieval

12 PERSONAL AUTHOR(S)

J.R. Fienup, J.D. Gorman, J.H. Seldin and J.N. Cederquist

13a TYPE OF REPORT	13b TIME COVERED	14 DATE OF REPORT (Year, Month, Day)	15 PAGE COUNT
Final Report, Vol 2	FROM 8/86 TO 12/89	1991 October	179

16 SUPPLEMENTARY NOTATION

17 COSATI CODES			18 SUBJECT TERMS *(Continue on reverse if necessary and identify by block number)*
FIELD	GROUP	SUB-GROUP	Unconventional Imaging, Discrimination, Phase Retrieval, Image Reconstruction, Amplitude Interferometer
20	05		
20	06		

19 ABSTRACT *(Continue on reverse if necessary and identify by block number)*

This report describes a technique for obtaining fine-resolution images, suitable for SDI midcourse discrimination, using an inexpensive, lightweight telescope that would ordinarily yield a poor image. If, instead of collecting a blurred image with the telescope, one sends the light through an amplitude interferometer, then the modulus, but not the phase, of the Fourier transform of the object can be measured, despite the aberrations. We have developed and analyzed phase retrieval algorithms that recover the unknown Fourier phase, which allows a fine-resolution image to be reconstructed despite the aberrations of the telescope. It will also correct aberrations due to atmospheric turbulence for a ground-based telescope, and can be used with several other imaging modalities.

20 DISTRIBUTION/AVAILABILITY OF ABSTRACT	21 ABSTRACT SECURITY CLASSIFICATION	
☐ UNCLASSIFIED/UNLIMITED ☐ SAME AS RPT ☐ DTIC USERS	Unclassified	
22a NAME OF RESPONSIBLE INDIVIDUAL Matt White	22b TELEPHONE (Include Area Code) (617) 451-3185	22c OFFICE SYMBOL Code 1112

DD Form 1473, JUN 86
ERIM 116

Previous editions are obsolete.

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE

19. Abstract (Continued)

The theory of the amplitude interferometer was advanced and alternative estimators of the Fourier modulus (the visibility function) were derived and analyzed, in particular for operation at low light levels. The effects of both noise and the structure of the object on the performance of image reconstruction were determined both analytically and by computer simulation and reconstruction experiments. A specific SDI discrimination test scenario, the first Firefly launch as seen from the University of Maryland's multi-aperture amplitude interferometer (MAAI) attached to the Goddard 48-inch telescope, was analyzed and simulated. It was found that a poor image would result due to the loss of low and mid spatial frequency information because of a large central obscuration in the telescope. It was also found that high quality imagery, having resolution several times finer than would ordinarily be allowed by the turbulence of the atmosphere, could be obtained under the same circumstances if (a) a telescope with a smaller central obscuration were used or (b) if the interferometer axis were offset in such a way as to measure the low spatial frequencies.

For the case of interferometry through a partially obscured aperture, an algorithm was developed that combines phase retrieval with interpolation in order to restore the information at the missing spatial frequencies while retrieving the phase at the unobscured spatial frequencies. Another new phase retrieval algorithm, based on the Ayers/Dainty blind deconvolution algorithm, was also developed. A new methodology for exploring the uniqueness of phase retrieval in a practical sense was developed and tested. It involves finding the ambiguous image which is closest to a given image by a reduced-gradient search technique. The computational requirements for the phase retrieval algorithm were quantified. Laboratory experiments to test the technique were initiated.

PREFACE

The work reported here was performed in the Optical and IR Science Laboratory of the Advanced Concepts Division, Environmental Research Institute of Michigan (ERIM). The work was sponsored by the Office of Naval Research (ONR), Boston, Contract No. N00014-86-C-0587, funded from the Innovative Science and Technology Office at the Strategic Defense Initiative Office (SDIO/IST). The project monitor at ONR was Dr. Fred W. Quelle.

This final technical report covers work performed from 1 August 1986 to 31 December 1989. The principal investigator was James R. Fienup. Major contributions to this work also included Jack N. Cederquist, John D. Gorman, and John H. Seldin.

(Volume 1 of the Final Report is by J.N. Cederquist, J.R. Fienup and J.C. Marron, "High Resolution Imaging by Phase Retrieval and Discrimination Using Speckle," ERIM Report No. 201600-11-F, March 1989, which describes work sponsored by the Office of Naval Research and the Naval Research Laboratory.)

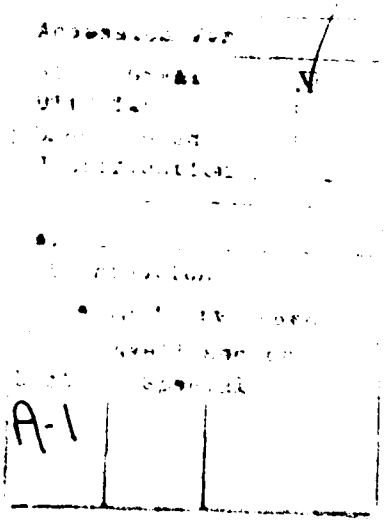


TABLE OF CONTENTS

TABLE OF CONTENTS
(Continued)

LIST OF FIGURES

LIST OF FIGURES
(Continued)

LIST OF TABLES

1.0 INTRODUCTION AND OVERVIEW

1.1 BACKGROUND

Discrimination of targets from decoys can be done using imagery having very fine resolution. The diffraction limit on resolution, $\rho = \lambda R/D$, obtained from an imaging sensor at a range R using wavelength λ and aperture diameter D, implies that, for SDI midcourse discrimination applications, the wavelength must be very short and/or the aperture diameter D must be very large. Such very large apertures would be impractically heavy and difficult to steer rapidly in space if they were made to be rigid in order to be without aberrations. On the other hand, mirrors that are inexpensive and lightweight would warp, causing phase errors and a severe blurring of the imagery.

An approach to circumventing these problems is to employ cheap, lightweight mirrors and obtain fine-resolution images from them using phase retrieval algorithms. By this approach, a computer algorithm corrects the errors after the data is collected. With the increasing speed and decreasing cost of computers, this trade-off of simpler optical hardware at the expense of additional computational requirements is increasingly attractive.

Phase retrieval can be employed to greatly improve the quality of imagery from a large number of sensors. In this study, we concentrated on a particular imaging sensor, the Multi-Aperture Amplitude Interferometer (MAAI), under development at the University of Maryland (UMd) by the group headed by Doug Currie. It is essentially a multi-channel, modernized Michelson stellar interferometer that gathers the Fourier transform of the target image, with all the spatial frequency components measured simultaneously. In the process of making those measurements, all information about the phase of the complex-valued Fourier transform is lost, and only the magnitude of the Fourier

transform (often referred to as the visibility function) is measured. This limited information is insufficient to compute an image in a straightforward manner. However with iterative phase retrieval algorithms, developed under this effort, a diffraction-limited image can be reconstructed. Aberrations then have no effect on the reconstructed image, and so fine resolution can be obtained despite warping of the mirror or, if present, atmospheric turbulence.

In this report is described an investigation using phase retrieval algorithms to reconstruct fine-resolution images from an aberrated system (the MAAI) for the SDI midcourse discrimination scenario. Section 1.2 gives a brief overview of the accomplishments that are described in detail in the rest of the report. Section 1.3 gives recommendations for future effort. Section 2 describes the basic theory behind the MAAI. Section 3 shows the performance of data estimation and image reconstruction for low light levels. Section 4 describes an analysis of the imaging performance that would be expected for future SDI experiments. Section 5 discusses the reconstruction of images for the case of partially-filled apertures as would occur if the telescope has a central obscuration. Section 6 describes alternative geometries within the MAAI that would enable it to measure low spatial frequencies despite a central obscuration, which would be useful for ground-based experiments. Section 7 describes an alternative new phase retrieval algorithm based on a blind deconvolution algorithm. Section 8 explores the probability that an image reconstructed by a phase retrieval algorithm is not unique. Section 9 shows the computational requirements for phase retrieval algorithms. Section 10 mentions plans towards reconstruction of images from MAAI data gathered in the laboratory. Additional details are given in several appendices. References are found at the end of each section.

1.2 OVERVIEW OF ACCOMPLISHMENTS

In this section the principal results of the program are briefly summarized. They are reported in detail in the sections and appendices that follow.

The basic theory of the MAAI was derived. This is explained in Section 2.

A signal and noise model for the MAAI was developed and analyzed. Several estimators for the object's Fourier magnitude from the measured data were derived, and the variance of the estimate was calculated as a function of detected photons and visibility magnitude. This leads to an optimum way to process the raw data prior to phase retrieval. Digital simulation and reconstruction experiments were performed to show the quality of imagery that would be reconstructed at different light levels and for different types of objects. This is described in Section 3.

For parameters of actual field experiments that were to be performed, the data was simulated and images were reconstructed. The scenario that was simulated was the imaging of the first Firefly exercise (piggybacking on the MIT Lincoln Laboratory laser radar experiment) launched from Wallops Island as would be viewed by the MAAI attached to the 48-inch telescope at Goddard Space Flight Center. Light levels received by the MAAI assuming sun illumination of the target, were computed, the detected data was simulated, and images were reconstructed. The results predicted that the images produced from the MAAI data from the Goddard 48-inch telescope would be of poor quality. A limiting factor was that the Goddard 48-inch telescope has a large central obscuration, preventing the measurement of the low-to-mid spatial frequencies, where most of the information resides. However, if the low spatial frequencies were measured, then it was shown that

good quality imagery could be reconstructed. This could be accomplished by changes in the MAAI (which will be described later) or by using a telescope which has a small central obscuration, such as the 24-inch at the Innovative Science and Technology Experimental Facility (ISTEF). Then for the same scenario, high-quality images would be reconstructed with resolution several times better than that ordinarily allowed by atmospheric turbulence. Furthermore, if the same experiment were performed in a space-borne MAAI at the same range, then excellent results would be obtained, even with shorter integration times. This is described in Section 4.

For the case of partially-filled aperture, including central obscurations or multiple-mirror telescopes, portions of the spatial frequency domain are not measured. Then the reconstruction algorithm must simultaneously interpolate the phase and magnitude values where they are missing while retrieving the phase where the magnitude is measured. This is a particularly difficult task if the lower spatial frequencies are missing because of a central obscuration of the telescope, since the visibility magnitude at lower spatial frequencies is typically much larger than at the higher spatial frequencies. Algorithms we developed to overcome this problem are described in Section 5.

Another way to get around the problem of a telescope with a central obscuration is to change the way that the aperture is sheared by the interferometer so that it measures the lower spatial frequencies. When this is done the highest spatial frequencies are lost, but the net image quality can be far higher than what would be obtained with the traditional method of shearing the wavefront. This is important for ground based experiments using existing telescopes, although it would probably not be a problem for an eventual space-based system for which a second small telescope could fill the need for the low spatial frequencies. This is described in Section 6.

An alternative to the iterative transform phase retrieval algorithm (which was the workhorse algorithm for most of this effort) was developed. It is a version of the Ayers-Dainty blind deconvolution algorithm modified to solve the phase retrieval problem, using support and nonnegativity constraints. This is described in Section 7.

A question that always arises for image reconstruction by phase retrieval is whether the image obtained is unique. If it were likely that other images were also consistent with the data and constraints, then the method would not be reliable. A new methodology of quantifying the uniqueness of the solution was developed and exercised. The subspace of all ambiguous solutions was analytically derived for the case of small (2 x 3 pixels) images. Monte Carlo experiments were conducted to determine the probability that a random image would lie within a certain distance of this subspace. The computation was performed for several different cases. This is reported in Section 8.

The computational requirements for phase retrieval were analyzed. Versions of the algorithm were also sent to other researchers to implement on particular computer architectures, such as the Carnegie-Mellon Warp. These results are described in Section 9.

Laboratory experiments were initiated, including preparation of target objects and porting software to a computer at the University of Maryland, as described in Section 10.

Publications arising from this effort are given below.

"Image Reconstruction for an Aberrated Amplitude Interferometer with a Partially-Filled Aperture," J.R. Fienup and J.D. Gorman, Proceedings of the NOAO-ESO Conference on High-Resolution Imaging by Interferometry, 15-18 March 1988, Garching bei Munchen, West Germany.

"Estimation and Reconstruction from Aberrated Amplitude Interferometer Measurements," J.D. Gorman and J.R. Fienup, in D.M. Alloin and J.-M. Mariotti, eds., Diffraction-Limited Imaging with Very Large Telescopes, (Kluwer Academic Publishers, Boston, 1989) pp. 405-414.

"Phase-Retrieval Imaging for SDI Applications," J.R. Fienup, Proceedings of the SDIO/IST Workshop on Sensor Signal Processing, 25-27 April, 1989, Leesburg, VA.

"Numerical Investigation of Phase Retrieval Uniqueness," J.H. Seldin and J.R. Fienup, in Signal Recovery and Synthesis III, digest of papers (Optical Society America, 1989), 14-16 June 1989, N. Falmouth, MA, pp. 120-123.

"Numerical Investigation of the Uniqueness of Phase Retrieval," J.H. Seldin and J.R. Fienup, J. Opt. Soc. Am. A 7, pp. 412-427, March 1990.

"Phase Retrieval Using Ayers/Dainty Deconvolution," J.H. Seldin and J.R. Fienup in Signal Recovery and Synthesis III, digest of papers (O.S.A., 1989), 14-16 June 1989, N. Falmouth, MA, pp. 124-127.

"Iterative Blind Deconvolution Algorithm Applied to Phase Retrieval," J.H. Seldin and J.R. Fienup, J. Opt. Soc. Am. A 7, pp. 428-433, March 1990.

"Lower Bounds on Parametric Estimators with Constraints," J.D. Gorman and A.O. Hero, Fourth Annual ASSP Workshop on Spectrum Estimation and Modeling, August 1988.

"Lower Bounds for Parametric Estimation with Constraints," J.D. Gorman and A.O. Hero, IEEE Trans. Inform. Theory 36, 1285-1301 (1990).

1.3 RECOMMENDATIONS

Phase retrieval has been shown via computer simulations to be a means of obtaining fine-resolution images, important for discriminating targets from decoys, from a badly-aberrated large-aperture telescope employing an amplitude interferometer. This will enable the generation of fine-resolution images from an imaging system that is much cheaper, simpler, and lighter in weight than what would otherwise be possible with competing technologies such as adaptive optics. It is recommended that phase retrieval be used in future imaging experiments to demonstrate its capabilities in the real world, that it be further developed to increase its speed and reliability, and that it be automated. The analysis of the uniqueness of the reconstructed image should be extended to include the case of larger, more realistic

images. Further analysis should be performed to determine which of the many known imaging modalities is best suited to the SDI midcourse discrimination problem. Phase retrieval can also be used to improve the images obtained with other types of imaging modalities.

2.0 AMPLITUDE INTERFEROMETER THEORY

2.1 OVERVIEW OF THE INTERFEROMETER

In this section we describe the basic theory behind the amplitude interferometer and discuss alternative ways to arrive at an estimate of the magnitude of the coherence function from it.

The multi-aperture amplitude interferometer [2.1,2.2,2.3] is essentially a highly parallel, multichannel, Michelson stellar interferometer [2.4] that uses a pair of measurements in an optimized measurement scheme. It can also be viewed as a dual-channel rotational shearing interferometer [2.5,2.6] with a 180° angle of rotation. It is presently under development by a group at the University of Maryland headed by D.G. Currie. A full description of the multiaperture amplitude interferometer has not appeared in the literature, and the description that follows was arrived at from a combination of the references cited above, conversations with the University of Maryland group, and our own analysis.

From the data collected by the amplitude interferometer we can compute the two-dimensional modulus (magnitude) of the complex coherence function of an astronomical object. If the conditions for the validity of the van-Cittert Zernike theorem are satisfied, then the complex coherence function is proportional to the Fourier transform of the two-dimensional intensity (brightness) distribution of the object under measurement. If both the modulus and phase of the complex coherence function could be computed, then one could obtain an image of the object by Fourier transformation. However, atmospheric turbulence and/or telescope aberrations severely distort the phase, allowing the determination of only the modulus of the complex coherence function, which is known as the visibility function.

In the amplitude interferometer, the incoming field is split into two halves, one of which is rotated by 180° with respect to the other. The two halves are then interfered and detected. The beamsplitter in the interferometer causes the interference pattern to appear simultaneously in two different planes. Both of these interference patterns, which are similar to one another yet different in a useful way, are detected. From them the modulus of the complex coherence function can be computed. The amplitude interferometer has an advantage over the rotational-shearing interferometer. The measurement of the pair of interference patterns largely allows for the correction of the effects of scintillation [2.1].

From the squared modulus of the coherence function we can compute the autocorrelation function of the object. Reconstruction of an image of the object requires the retrieval of the phase of the complex coherence function, which can be accomplished using a phase retrieval algorithm [2.7,2.8]. By this means an image can be obtained that has several times finer resolution than what could ordinarily be obtained through the turbulent atmosphere or through an aberrated telescope.

2.2 THE AMPLITUDE INTERFEROMETER

We make the standard assumptions that the object of interest radiates incoherently, the interferometer is in the far-field of the object, and the detected light is quasi-monochromatic. Under these conditions the van Cittert-Zernike theorem, which states that the object brightness distribution is the Fourier transform of the complex coherence function, is valid [2.9]. We also assume isoplanatism: that the effects of the aberrations are modeled by a random phase-amplitude screen appearing at the entrance pupil of the interferometer, and its aberrating effects are space-invariant.

The amplitude interferometer was originally designed to measure stellar diameters by making one-dimensional measurements of the modulus of the complex coherence function. This one-dimensional interferometer receives recollimated light from a telescope, and consists of a Koster's prism, spectral filters, and photomultiplier tubes at each output arm of the prism. This arrangement allowed the measurement of the interference between a pair of pinholes with variable separation. This type of measurement was sufficient for stellar diameter measurements. In the current amplitude interferometer, the multi-aperture amplitude interferometer (MAAI), which is illustrated in Figure 2-1, the photomultiplier tubes have been replaced by 2-D CCD arrays and additional optics have been incorporated between the collimator and the Koster's prism, making it capable of making two-dimensional measurements. These measurements are made in a plane that is a demagnified version of the aperture (pupil) plane.

The key optical component of the amplitude interferometer is a Koster's prism. The prism acts as a beamsplitter, combining two incident optical fields. If an intensity detector is placed at an output of the prism, what is measured includes a term proportional to the coherence function of the incident field. This principle is used to measure the modulus of the complex coherence function of the object. In our discussion we assume an ideal Koster's prism. Liewer [2.3] discusses the effects of a nonideal prism.

A complex-valued optical field $U(x,y,t)$ enters the interferometer from a telescope and is split into half fields. One half field passes through two mirror reflections and into one side of the Koster's prism. The other half passes through three mirror reflections and into the other side of the prism (Figure 2-1). The mirrors between the telescope and the prism act to invert one of the halves about the horizontal axis, making it $U(x,-y,t)$, while the other half remains unchanged. These two halves are combined with the beamsplitting action

- Amplitude interferometer (Currie 1967, 1974)

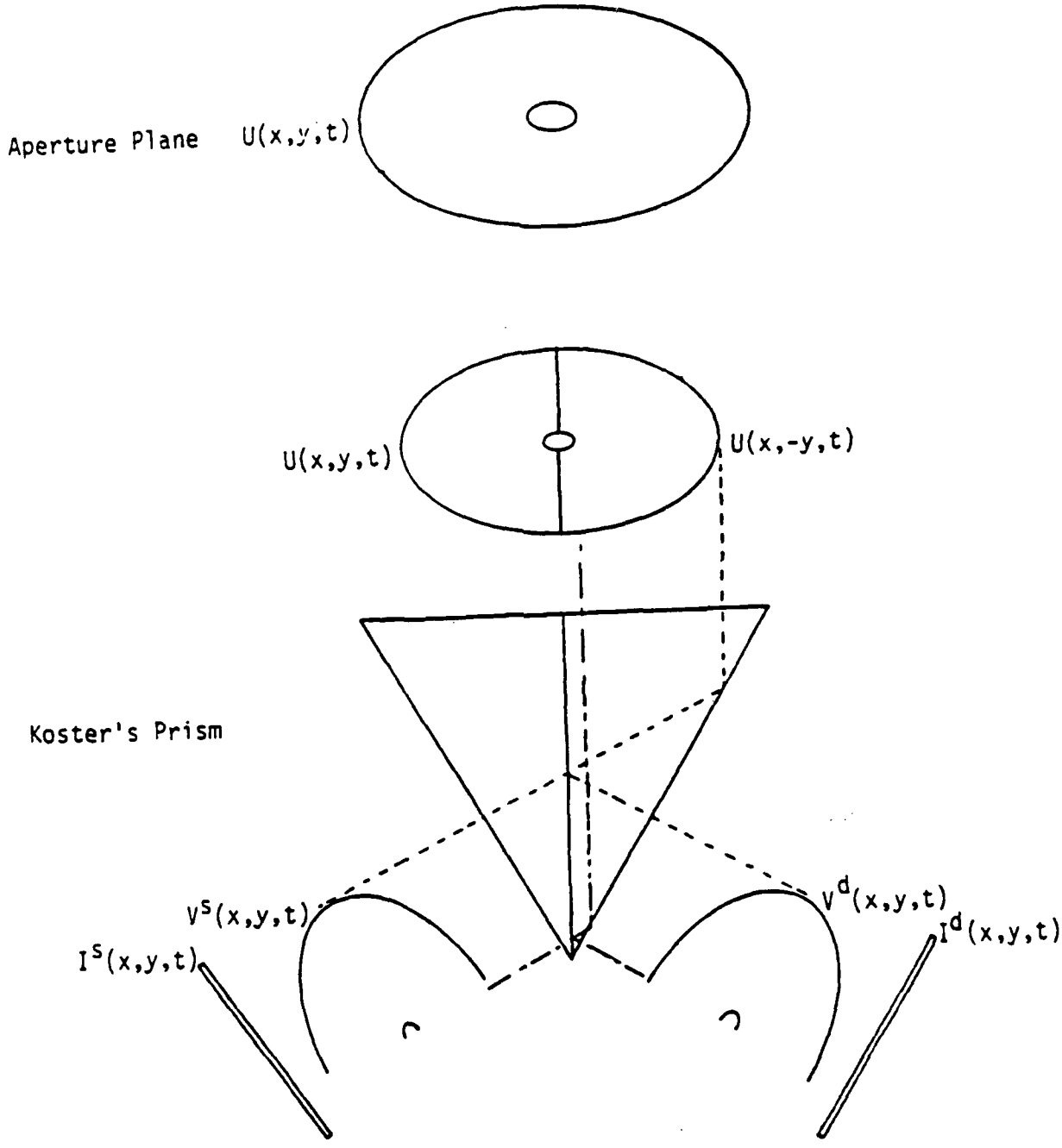


Figure 2-1. Functional Diagram of the Multi-Aperture Amplitude Interferometer.

of the prism. A simple ray-tracing argument can be used to show that the transmitted beam undergoes a constant phase shift of θ_T and an inversion about the vertical axis while the reflected beam undergoes a constant phase shift of θ_R. Assume that $U(x,y,t)$ enters the left side of the prism and the inverted beam $U(x,-y,t)$ enters the right side of the prism. Then the complex field of the beam output on the left side of the prism, denoted as beam 1, is

$$V_1(x,y,t) = \frac{1}{\sqrt{2}} \left\{ U(x,y,t)\, e^{i\theta_R} + U(-x,-y,t)\, e^{i\theta_T} \right\} . \qquad (2\text{-}1)$$

where the $1/\sqrt{2}$ factor is required for energy conservation. Similarly, the output complex field on the right side of the prism is

$$V_2(x,y,t) = \frac{1}{\sqrt{2}} \left\{ U(-x,y,t)\, e^{i\theta_T} + U(x,-y,t)\, e^{i\theta_R} \right\} . \qquad (2\text{-}2)$$

Let $\langle \bullet \rangle_\tau$ denote a time average over the interval $[t,t+\tau]$; that is,

$$\langle f(t) \rangle_\tau = \frac{1}{\tau} \int_t^{t+\tau} f(t')\, dt' . \qquad (2\text{-}3)$$

In the context of our model, τ represents the single-frame integration time of the CCD array, which would typically be on the order of 1 msec to 10 msec for the case of atmospheric turbulence. Then the detected intensity of beam 1 is

$$I_1(x,y,t) = \langle |V_1(x,y,t)|^2 \rangle_\tau$$

$$= \frac{1}{2} \left\{ \langle |U(x,y,t)|^2 \rangle_\tau + \langle |U(-x,-y,t)|^2 \rangle_\tau \right.$$

$$\left. + \langle U(x,y,t)\, U^*(-x,-y,t) \rangle_\tau\, e^{i\delta} + \text{c.c.} \right\}$$

$$= \frac{1}{2}\left\{I(x,y,t) + I(-x,-y,t) + \langle U(x,y,t)\, U^*(-x,-y,t)\rangle_\tau\, e^{i\delta} + \text{c.c.}\right\} \tag{2-4}$$

where $\delta = \theta_R - \theta_T$, $I(x,y,t) = \langle |U(x,y,t)|^2 \rangle_\tau$, and c.c. denotes the complex conjugate of the preceding term. For an ideal beamsplitter $\delta = \pi/2$.

The optical field in the aperture plane is assumed to be given by

$$U(x,y,t) = U_0(x,y,t)\, \exp[\alpha(x,y,t) + i\beta(x,y,t)] \tag{2-5}$$

where

$$U_0(x,y,t) = U_0(x,y)\, \exp(i\omega t) \tag{2-6}$$

is the quasimonochromatic optical field of wavelength $\lambda = 2\pi c/\omega$ due to the object in the absence of atmospheric effects, $\alpha(x,y,t)$ is the intensity-modulating effect (scintillation) of atmospheric turbulence (the log-amplitude function) [2.9, pp. 398, 404], $\beta(x,y,t)$ is the phase error induced by atmospheric turbulence or aberrated optics, and c is the speed of light.

We assume that the integration time τ is many times the coherence time of the optical field, which is approximately the reciprocal of the bandwidth, $\Delta\nu = \Delta\omega/2\pi$, of the radiation. Thus the mutual intensity of the incident optical field due to the object is given by

$$\begin{aligned} \Gamma(\Delta x,\Delta y) &= |\Gamma(\Delta x,\Delta y)|\, \exp[i\phi(\Delta x,\Delta y)] \\ &= \langle U_0(x,y,t)\, U_0^*(x - \Delta x,\, y - \Delta y,\, t)\rangle_\tau \\ &= I_0\, \gamma(\Delta x,\Delta y) \end{aligned} \tag{2-7}$$

where $I_o = \Gamma(0,0) = \langle |U_o(x,y,t)|^2 \rangle_\tau$ is the average intensity and the normalized quantity $\gamma(\Delta x,\Delta y)$ is the complex coherence function. ($\gamma(\Delta x,\Delta y)$ is usually denoted by $\mu_{12} = \gamma_{12}(0)$ [2.9, p. 183]; however we use the symbol γ to be consistent with the notation of earlier publications on the amplitude interferometer.)

Inserting Eqs. (2-5) to (2-7) into Eq. (2-4), and assuming that $\alpha(x,y,t)$ and $\beta(x,y,t)$ are constant over the time interval τ, yields

$$I_1(x,y,t) = (I_o/2)\ \{\exp[2\alpha(x,y,t)] + \exp[2\alpha(-x,-y,t)] + \gamma(2x,2y)\ \exp[\alpha(x,y,t) + \alpha(-x,-y,t) + i\beta(x,y,t) - i\beta(-x,-y,t)]\ \exp(i\delta) + c.c.\}. \quad (2\text{-}8)$$

For an ideal beamsplitter, with $\delta = \theta_R - \theta_T = \pi/2$, this becomes

$$I_1(x,y,t) = (I_o/2)\ \{\exp[2\alpha(x,y,t)] + \exp[2\alpha(-x,-y,t)] - 2\ \exp[\alpha(x,y,t) + \alpha(-x,-y,t)]\ |\gamma(2x,2y)|\ \sin[\phi(2x,2y) + \beta(x,y,t) - \beta(-x,-y,t)]\} \quad . \quad (2\text{-}9)$$

$|\gamma|$ is the visibility (contrast) of the sinusoidal fringe that was seen by Michelson. Similarly

$$I_2(-x,y,t) = (I_o/2)\ \{\exp[2\alpha(x,y,t)] + \exp[2\alpha(-x,-y,t)] + 2\ \exp[\alpha(x,y,t) + \alpha(-x,-y,t)]\ |\gamma(2x,2y)|\ \sin[\phi(2x,2y) + \beta(x,y,t) - \beta(-x,-y,t)]\} \quad . \quad (2\text{-}10)$$

A function related to the fringe visibility function is given by

$$\frac{I_2 - I_1}{I_2 + I_1} = \frac{I_2(-x,y,t) - I_1(x,y,t)}{I_2(-x,y,t) + I_1(x,y,t)}$$

$$= \frac{2\exp[\alpha(x,y,t) + \alpha(-x,-y,t)]\,|\gamma(2x,2y)|\,\sin[\phi(2x,2y) + \beta(x,y,t) - \beta(-x,-y,t)]}{\exp[2\alpha(x,y,t)] + \exp[2\alpha(-x,-y,t)]}$$

$$= \frac{|\gamma(2x,2y)|\,\sin[\phi(2x,2y) + \beta(x,y,t) - \beta(-x,-y,t)]}{\cosh[\alpha(x,y,t) - \alpha(-x,-y,t)]} \quad . \qquad (2\text{-}11)$$

One of the major advantages of the amplitude interferometer over other rotational shearing interferometers is the suppression of the effects of the scintillation, $\alpha(x,y,t)$, by the cosh[ ] function in Eq. (2-11).

In the absence of phase errors, $(I_2 - I_1)/(I_2 + I_1)$ of Eq. (2-11) yields $|\gamma(2x,2y)|\,\sin[\phi(2x,2y)]$, which is the imaginary part of $\gamma(2x,2y)$. Under this condition, if the object were to be positioned to one side of the optical axis, then it could easily be reconstructed by Fourier transforming the imaginary part of $\gamma(2x,2y)$ and discarding one of the resulting twin images. However, the phase errors $\beta(x,y,t)$ prevent us from doing this when imaging through the aberrations. Averaging over a time long compared with the fluctuation time of $\beta(x,y,t)$ just causes $(I_2 - I_1)/(I_2 + I_1)$ to average out to zero.

Suppose we gather M short exposures (frames), each of duration τ, separated by time Δt. Further suppose that the total collection time, $T = M\Delta t$, is many times the correlation time of the phase error. Then one way to extract desired quantity, $|\gamma(2x,2y)|$, from Eq. (2-11) is as follows. Ignoring $\alpha(x,y,t)$, we can square Eq. (2-11) and obtain

$$\left[\frac{I_2 - I_1}{I_2 + I_1}\right]^2 = |\gamma(2x,2y)|^2 \sin^2[\phi(2x,2y) + \beta(x,y,t) - \beta(-x,-y,t)]. \qquad (2\text{-}12)$$

Averaging this quantity over the M frames gives

$$\left\langle\left[\frac{I_2 - I_1}{I_2 + I_1}\right]\right\rangle_T = |\gamma(2x,2y)|^2 \left\langle\sin^2[\phi(2x,2y) + \beta(x,y,t) - \beta(-x,-y,t)]\right\rangle_T$$

$$\equiv |\gamma(2x,2y)|^2 M^{-1} \sum_{m=1}^{M} \sin^2[\phi(2x,2y) + \beta(x,y,m\Delta t) - \beta(-x,-y,m\Delta t)]$$

$$\simeq |\gamma(2x,2y)|^2/2 \quad , \tag{2-13}$$

where it is assumed that the phase error β varies with time and is uniformly distributed over $(-\pi,\pi)$ over the time interval T. Therefore a reasonable estimator for $|\gamma(2x,2y)|^2$ is

$$|\hat{\gamma}(2x,2y)|^2 = 2 \left\langle \left[\frac{I_2 - I_1}{I_2 + I_1}\right]^2 \right\rangle_T \tag{2-14}$$

Currie [2.1,2.2] proposed using the quantity

$$|\hat{\gamma}(2x,2y)| = \sqrt{2} \left[\frac{AC - CC}{AC + CC}\right]^{1/2} \tag{2-15}$$

where

$$AC = \langle I_1^2 + I_2^2 \rangle_T \tag{2-16}$$

and

$$CC = 2\langle I_1 I_2 \rangle_T \quad . \tag{2-17}$$

Inserting Eqs. (2-16) and (2-17) into Eq. (2-15) reveals that this yields

$$|\hat{\gamma}(2x, 2y)|^2 = 2 \frac{\langle (I_2 - I_1)^2 \rangle_T}{\langle (I_2 + I_1)^2 \rangle_T} \tag{2-18}$$

which is similar to the estimator given in Eq. (2-14) but changes the order of the time averaging operation and the division operation. However, as will be seen later, the performance of the estimator in Eq.

(2-14) can be shown to be significantly better for the case of low light levels.

Alternatively if the phase error $\beta(x,y,t)$ is constant during the total integration time T, then the fluctuations in β cannot be employed to cause the average of the $\sin^2[\ ]$ term to be 1/2. Then we can achieve the same effect by introducing a phase plate, with spatially uniform phase $\theta(t)$, which can change with time, in front of one half of the Koster's prism. Then Eq. (2-12) is replaced by

$$\left[\frac{I_2 - I_1}{I_2 + I_1}\right]^2 = |\gamma(2x,2y)|^2 \sin^2[\phi(2x,2y) + \beta(x,y,t) - \beta(-x,-y,t) - \theta(t)] \ . \tag{2-19}$$

One choice of $\theta(t)$ would be 0 for half the time and $\pi/2$ for the other half the time. Since $\sin^2(\theta_o + 0) + \sin^2(\theta_o - \pi/2) = \sin^2(\theta_o) + \cos^2(\theta_o) = 1$, then

$$\left\langle\left[\frac{I_2 - I_1}{I_2 + I_1}\right]^2\right\rangle_T = |\gamma(2x,2y)|^2 \ . \tag{2-20}$$

This scheme has the great advantage that only two frames of data need taken to estimate $|\gamma|^2$, and this maximizes the signal-to-noise ratio for a given total number of photons, as will be seen later. Another possible choice for $\theta(t)$ is the discrete values $\{0, \pi/2, \pi, 3\pi/2\}$. Another is to vary $\theta(t)$ continuously between 0 and 2π radians, while integrating over an integer number of frames during each 0 to 2π cycle.

Additional estimators of $|\gamma|^2$ can be obtained by averaging then dividing, i.e. $\langle(I_2 - I_2)^2\rangle_T / \langle(I_2 + I_1)^2\rangle_T$, rather than dividing then averaging as was assumed above.

The section that follows treats the case of measurements limited by photon noise in which case different estimators of $|\gamma|^2$ can have significantly different variances.

REFERENCES

[2.1] D.G. Currie, Appendix II of "Woods Hole Summer Study on Synthetic Aperture Optics," Vol 2, Natl. Acad. of Sci. - Nat. Res. Council, Wash. D.C., 1967.

[2.2] D.G. Currie, S.L. Knapp and K.M. Liewer, "Four Stellar-Diameter Measurements by a New Technique: Amplitude Interferometry," Astrophys. J. 187, 131-134 (1974).

[2.3] K.M. Liewer, Ph.D. Thesis, University of Maryland, 1974.

[2.4] A.A. Michelson and F.G. Peace, "Measurement of the Diameter of Alpha Orionis with the Interferometer," Astrophys. J. 53, 249-259 (1921).

[2.5] J.B. Breckinridge, "Obtaining Information Through the Atmosphere at the Diffraction Limit of a Large Aperture," J. Opt. Soc. Am. 65, 2996-2998 (1975).

[2.6] F. Roddier, C. Roddier, and J. Demarcq, "A Rotation Shearing Interferometer with Phase-Compensated Roof Prisms," J. Optics (Paris) 9, 1978.

[2.7] J.R. Fienup, "Reconstruction of an Object from the Modulus of Its Fourier Transform," Opt. Lett. 3, 27-29 (1978).

[2.8] J.C. Dainty and J.R. Fienup, "Phase Retrieval and Image Reconstruction for Astronomy," Chapter 7 in H. Stark, ed., Image Recovery: Theory and Application (Academic Press, 1987), pp. 231-275.

[2.9] J.W. Goodman, Statistical Optics, J. Wiley & Sons, New York, 1985.

3.0 PERFORMANCE AT LOW LIGHT LEVELS

In this section we examine the performance of the amplitude interferometer at low light levels, both analytically and through computer simulation. Continuing the development in Section 2, we provide a statistical model of the amplitude interferometer and discuss a method for obtaining diffraction-limited imagery from aberrated, low light-level measurements of the mutual coherence function. Our basic approach is to perform a sequence of measurements from which samples of the modulus of the mutual coherence can be estimated and then to perform phase retrieval to recover the complex mutual coherence function. The recovered samples of the coherence function are then Fourier transformed to yield an image of the object intensity.

The organization of Section 3 is as follows. In Section 3.1, we present a statistical model for the amplitude interferometer and discuss three methods for estimating the modulus of the mutual coherence from low light level amplitude interferometer measurements in the presence of aberrations. The first two methods, [3.3], which are suitable for applications in which the aberration in slowly varying, require a modification of the amplitude interferometer as shown in Figure 3-1. The third method, proposed by Currie, [3.4,3.5], can be used in situations where the aberrations are rapidly-varying such as aberrations caused by atmospheric turbulence. In Section 3.2, we develop a lower bound on the mean-squared error in estimating the object intensity from amplitude interferometer measurements, using the statistical model of Section 3.1. Finally, Section 3.3 contains results from several digital simulations and image-reconstruction experiments. As one might expect, the quality of the reconstruction depends not only on the light level, but also on the content of the image. The more specular or point-like the object is, the better the reconstruction; diffuse objects are the most difficult to reconstruct. These observations are confirmed by the digital simulations in Section 3.3.

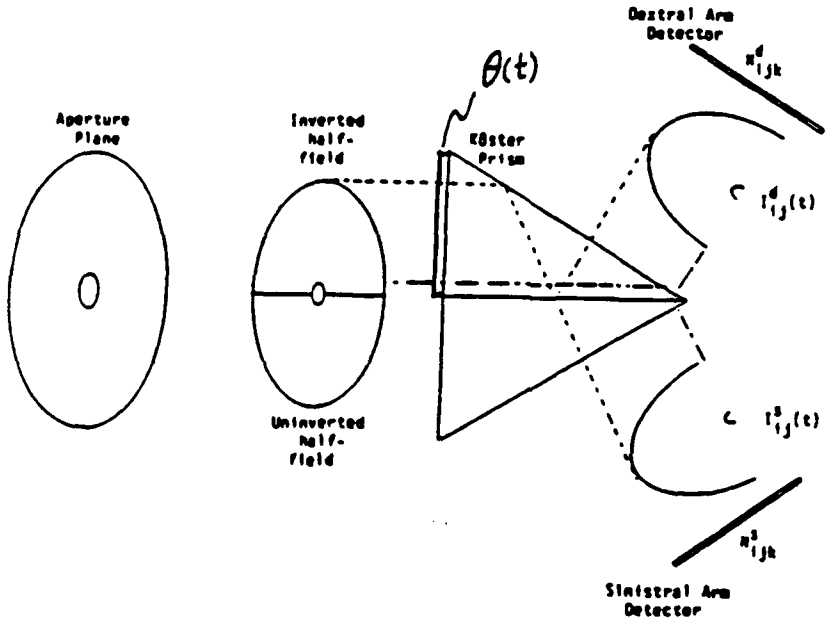


Figure 3-1. Schematic Diagram of a Modified Amplitude Interferometer. A variable phase plate has been added to allow the introduction of phase term $\theta(t)$ into the measurements.

3.1 MEASUREMENT MODEL

The amplitude interferometer measurements are assumed to consist of a sequence of pairs of two-dimensional video frames which are the outputs of the CCD arrays at each of the two output arms of the interferometer. We denote these measurements as (N^1_{ijk}, N^2_{ijk}), where N^1_{ijk} and N^2_{ijk} respectively are the detected output energy at the (i,j)th detector element and the kth frame of the left and right output arms of the interferometer. At low light levels and with ideal detectors, N^1_{ijk} and N^2_{ijk} consist of the number of photon events detected over each detector element (i,j) and over each time frame k. The counts are well-modeled as Poisson-distributed random variables [3.13,3.15] with mean values

$$\Lambda^1_{ijk} = \int_{t_{k-1}}^{t_k} \left[I^1_{ij}(t) + I_B\right] dt, \quad \text{and} \quad \Lambda^2_{ijk} = \int_{t_{k-1}}^{t_k} \left[I^2_{ij}(t) + I_B\right] dt \qquad (3\text{-}1)$$

respectively, where I_B models contributions due to background light and the dark current of the CCD arrays and $[t_{k-1}, t_k]$ denotes the detector integration interval for the kth frame. $I^1_{ij}(t)$ and $I^2_{ij}(t)$ denote the respective instantaneous intensities at the output of the two interferometer arms.

Expressions for $I^1_{ij}(t)$ and $I^2_{ij}(t)$ were previously derived in Section 2. Here we use the subscript notation to emphasize the fact that the output intensities $I_1(x,y,t)$ and $I_2(-x,y,t)$ of Eqs. (2-9) and (2-10) are sampled:

$$I^1_{ij}(t) = \int_{\Delta x} \int_{\Delta y} I_1(x - x_i, y - y_j, t)\, dx\, dy \ , \qquad (3\text{-}2)$$

where (x_i, y_j) denotes the center of the (i,j)th detector element and $(\Delta x, \Delta y)$ is its area. A similar relationship holds for $I^2_{ij}(t)$. In our notation for the discretized mutual coherence function, we suppress the fact that γ is sampled at <u>half</u> the rate that the output intensities I_1 and I_2 are:

$$\gamma_{ij} \simeq \int_{\Delta x} \int_{\Delta y} \gamma(2x - 2x_i, 2y - 2y_j)\, dx\, dy \quad . \tag{3-3}$$

This reduction in sampling rate results from the fact that incident field components (x_i, y_j) and $(-x_i, -y_j)$ are interfered to obtain the mutual coherence component at $(2x_i, 2y_j)$. This difference in sampling rates is not important for the discussion in this section, however, it plays an important factor in the determination of the appropriate sampling rates in the Firefly simulation discussed in Section 4.3. For simplicity, we assume the integration interval $\Delta t = t_k - t_{k-1}$ is the same for each frame. We also assume that I_B is explicitly known and, for simplicity, that it is constant in time and over the entire aperture plane. The intensity parameters Λ^1_{ijk} and Λ^2_{ijk} are possibly random variables due to the stochastic nature of the phase term $\phi^1_{ij}(t)$. Therefore, processes such as N^1_{ijk} and N^2_{ijk} are typically called doubly-stochastic Poisson processes [3.15]. By this we mean that, conditioned on Λ^1_{ijk}, N^1_{ijk} is a discrete random variable with the probability mass function:

$$\Pr\left\{N^1_{ijk} = n \middle| \Lambda^1_{ijk}\right\} = \frac{\left(\Lambda^1_{ijk}\right)^n \exp\left\{-\Lambda^1_{ijk}\right\}}{n!} \quad . \tag{3-4}$$

A similar relationship holds for N^2_{ijk}.

Assuming an ideal beamsplitter ($\delta = \pi/2$ in Eq. 2-4) and ignoring the effects of scintillation, the instantaneous output intensities, $I^1_{ij}(t)$ and $I^2_{ij}(t)$, can be reexpressed as:

$$I_{ij}^{1}(t) = I_o\{1 - |\gamma_{ij}| \sin[\arg \gamma_{ij} + \Psi_{ij}(t)]\}$$

$$I_{ij}^{2}(t) = I_o\{1 + |\gamma_{ij}| \sin[\arg \gamma_{ij} + \Psi_{ij}(t)]\} \quad , \qquad (3\text{-}5)$$

where subscripts 1 and 2 denote the left and right output arms of the interferometer, γ_{ij} denotes the (discretized) normalized complex mutual coherence function of the incident field, I_o denotes the average instantaneous detected energy in photons per second, and $\Psi_{ij}(t)$ denotes the phase difference between the input arms of the interferometer and can include both random and non-random contributions from fixed or varying system aberrations and atmospheric turbulence.

We assume that I_o is known or can be accurately determined from the measurements. This is not an unrealistic assumption since, by Eq. (3-1) I_o can be estimated by forming the sum,

$$\hat{I}_o = \frac{1}{2N^2 K \Delta t} \sum_{ijk} \left[N_{ijk}^{1} + N_{ijk}^{2}\right] - I_B \quad . \qquad (3\text{-}6)$$

Here, N^2 denotes the total number of pixels in each of K pairs of frames in the data collection. $\hat{I}_o$ is based upon N^2K independent measurements and its variance decreases as the number of frames or pixels increase.

Our approach to image reconstruction from amplitude interferometer measurements will be to form an estimate of the modulus of the mutual coherence function $|\gamma_{ij}|$ and perform phase retrieval to recover the phase of the coherence function, $\gamma_{ij} = |\gamma_{ij}| \exp\{\arg \gamma_{ij}\}$, from its modulus. A reconstructed image is then formed by inverse Fourier transformation of the coherence function. An estimator for $|\gamma_{ij}|$ can be determined given the model described above. A reasonable estimate is to choose the values $|\hat{\gamma}_{ij}|$ which are most likely to have resulted in

the measurements (N^1_{ijk}, N^2_{ijk}). Such an estimate is obtained by maximizing the logarithm of the probability of the measurements, $p(N^1_{ijk}, N^2_{ijk})$, with respect to $|\gamma_{ij}|$. This approach, called maximum-likelihood estimation, has several desirable features which are mentioned in [3.15]. For Poisson-distributed random variables, the logarithm of the probability distribution, denoted $L(\gamma)$, is

$$L(\gamma) = -\sum_{ijk} \left[\Lambda^1_{ijk} + \Lambda^2_{ijk}\right] + \sum_{ijk} \log\left[\Lambda^1_{ijk}\right] N^1_{ijk} + \sum_{ijk} \log\left[\Lambda^2_{ijk}\right] N^2_{ijk} + C \tag{3-7}$$

where C is a constant which is independent of $|\gamma_{ij}|^2$. The maximum-likelihood estimator for $|\gamma_{ij}|$, if it exists, is then a solution of the *equation*

$$0 = \frac{\partial L}{\partial |\gamma_{ij}|} = \sum_k \frac{N^1_{ijk}}{\Lambda^1_{ijk}} \frac{\partial \Lambda^1_{ijk}}{\partial |\gamma_{ij}|} + \sum_k \frac{N^2_{ijk}}{\Lambda^2_{ijk}} \frac{\partial \Lambda^2_{ijk}}{\partial |\gamma_{ij}|} \quad . \tag{3-8}$$

Equation (3-8) is nonlinear in $|\gamma_{ij}|$ and is generally difficult to solve. Moreover, no information has been specified about $\Psi_{ij}(t)$. In the subsequent discussion, we examine three estimators for $|\gamma_{ij}|$ for the cases where:

1. $\Psi_{ij}(t)$ is constant over each of K intervals
2. $\Psi_{ij}(t)$ varies linearly over the collection period, and
3. $\Psi_{ij}(t)$ contains a phase term due to atmospheric turbulence and changes rapidly over the collection period.

In the first two cases, we assume that the phase term $\Psi_{ij}(t)$ is given by

$$\Psi_{ij}(t) = \Delta\beta_{ij}(t) + \theta(t) \quad , \tag{3-9}$$

where $\Delta\beta_{ij}(t) = \beta(x,y,t) - \beta(-x,-y,t)$ from Eq. (2-9), and $\theta(t)$ is a user-controlled phase term introduced into the amplitude interferometer. One method of incorporating such a phase term is to place a variable-phase plate over one of the input arms of the interferometer as shown in Figure 3-1. In the third case, which is discussed somewhat at the end of Section 2, we assume that $\Psi_{ij}(t)$ is given by

$$\Psi_{ij}(t) = \Delta\beta_{ij}(t) \quad , \tag{3-10}$$

where $\Delta\beta_{ij}(t)$ is the phase difference introduced by atmospheric turbulence as described in Eq. (2-9). In the discussion to follow we assume that the phase term $\Psi_{ij}(t)$ is constant during any integration interval $[t_{k-1}, t_k]$ and denote it as Ψ_{ijk}.

3.1.1 Discrete Stepped-Phase Systems

Consider a stepped-phase system in which $\theta(t)$ in Eq. (3-9) is constant over each of K intervals of length $\Delta t = T/K$, where T is the total collection period, and denote its value by θ_k, $k = 1, \ldots, K$. Here we assume that $\Delta\beta_{ij}(t)$ is constant over T: $\Delta\beta_{ij}(t) = \Delta\beta_{ijk}$, $t \epsilon [t_{k-1}, t_k]$. Define

$$g_{ijk} = |\gamma_{ij}| \sin[\arg \gamma_{ij} + \Delta\beta_{ij} + \theta_k] \quad . \tag{3-11}$$

Then A^1_{ijk} and A^2_{ijk} become, using (3-1), (3-5), and (3-9),

$$A^1_{ijk} = \frac{I_o T}{K} \{c - g_{ijk}\} \quad \text{and} \quad A^2_{ijk} = \frac{I_o T}{K} \{c + g_{ijk}\} \quad , \tag{3-12}$$

where

$$c = 1 + I_B/I_o \quad . \tag{3-13}$$

If θ_k is chosen to satisfy

$$\frac{1}{K}\sum_{k=1}^{K} \sin[\arg \gamma_{ij} + \Delta\beta_{ij} + \theta_k] = 0 \quad \text{for} \quad K > 2,$$

and

$$\frac{1}{K}\sum_{k=1}^{K} \sin^2[\arg \gamma_{ij} + \Delta\beta_{ij} + \theta_k] = \frac{1}{2} \quad , \tag{3-14}$$

we have that

$$|\gamma_{ij}|^2 = \frac{2}{K}\sum_{k} g_{ijk}^2 \quad . \tag{3-15}$$

Thus the motivation for introducing the controllable phase term θ_k into the interferometer is that for a suitable sequence θ_k, $k = 1, \ldots, K$, one can determine $|\gamma|^2$ from (A_{ijk}^1, A_{ijk}^2) regardless of the aberration $\Delta\beta_{ij}$.

One could consider the two-step process of first computing the maximum-likelihood estimate of g_{ijk}^2 and then estimating $|\gamma_{ij}|^2$ from g_{ijk}^2, using the above equation. Maximization of Eq. (3-7) with respect to g_{ijk}^2 is much simpler and the maximum-likelihood estimate of g_{ijk}^2 is given by

$$\hat{g}_{ijk}^2 = c^2 \left[\frac{N_{ijk}^2 - N_{ijk}^1}{N_{ijk}^2 + N_{ijk}^1}\right]^2 \quad . \tag{3-16}$$

where c is given by Eq. (3-13). The resulting estimator for the squared modulus is then

$$|\hat{\gamma}_{ij}|^2 = \frac{2}{K} \sum_k \hat{g}^2_{ijk} = \frac{2c^2}{K} \sum_k \left[\frac{N^2_{ijk} - N^1_{ijk}}{N^2_{ijk} + N^1_{ijk}} \right]^2 \quad . \tag{3-17}$$

We refer to this as "discrete estimator 1" (D1). For each pixel (i,j) and each frame k, the quantity $N^2_{ijk} - N^1_{ijk}$ is normalized by the total number of counts detected within the pixel and frame, $N^2_{ijk} + N^1_{ijk}$. One might also consider performing the normalization operation after frame averaging; this results in two other estimators which we refer to as "discrete estimator 2" (D2):

$$|\hat{\gamma}_{ij}|^2 = \frac{1}{2K(\hat{I}_o T)^2} \sum_k \left[N^2_{ijk} - N^1_{ijk} \right]^2 \quad , \tag{3-18}$$

and "discrete estimator 3" (D3):

$$|\hat{\gamma}_{ij}|^2 = 2c^2 \frac{\sum_k \left[N^2_{ijk} - N^1_{ijk} \right]^2}{\sum_k \left[N^2_{ijk} + N^1_{ijk} \right]^2} \quad . \tag{3-19}$$

At extremely low light levels there is a bias term proportional to $1/I_o$ which is present in all three estimators. To account for this bias, correction terms can be incorporated into the estimators. Bias-corrected (BC) versions of these estimators are given by

D1-BC:

$$|\hat{\gamma}_{ij}|^2 = \frac{2c^2}{K} \sum_k \frac{\left[N^2_{ijk} - N^1_{ijk} \right]^2 - \left[N^2_{ijk} + N^1_{ijk} \right]}{\left[N^2_{ijk} + N^1_{ijk} \right]^2 - \left[N^2_{ijk} + N^1_{ijk} \right]} \quad , \tag{3-20}$$

D2-BC:

$$|\hat{\gamma}_{ij}|^2 = \frac{1}{2K(\hat{I}_o T)^2} \sum_k \left[\left(N^2_{ijk} - N^1_{ijk}\right)^2 - \left(N^2_{ijk} + N^1_{ijk}\right)\right] \quad , \quad (3\text{-}21)$$

where $\hat{I}_o$ is given by (3-6)

and D3-BC:

$$|\hat{\gamma}_{ij}|^2 = 2c^2 \frac{\sum_k \left(N^2_{ijk} - N^1_{ijk}\right)^2 - \left(N^2_{ijk} + N^1_{ijk}\right)}{\sum_k \left(N^2_{ijk} + N^1_{ijk}\right)^2 - \left(N^2_{ijk} + N^1_{ijk}\right)} \quad . \quad (3\text{-}22)$$

3.1.2 Continuous-Phase Systems

Another possibility is that the controlled phase term $\theta(t)$ of Eq. (3-9) varies linearly from 0 to 2π as t goes from t_o to $t_o + T$: $\theta(t) = 2\pi(t - t_o)/T$, $t_o \le t \le t_o + T$. In this case, $|\gamma_{ij}|^2$ can be recovered from a sequence of four frames, each with integration time equal to T/K, $K = 4$. Let

$$A_{ij} = \frac{1}{I_o} \int_{t_o}^{t_o+T/4} \left[I^2_{ij}(t) - I^1_{ij}(t)\right] dt, \quad B_{ij} = \frac{1}{I_o} \int_{t_o+T/4}^{t_o+T/2} \left[I^2_{ij}(t) - I^1_{ij}(t)\right] dt,$$

$$C_{ij} = \frac{1}{I_o} \int_{t_o+T/2}^{t_o+3T/4} \left[I^2_{ij}(t) - I^1_{ij}(t)\right] dt, \quad D_{ij} = \frac{1}{I_o} \int_{t_o+3T/4}^{t_o+T} \left[I^2_{ij}(t) - I^1_{ij}(t)\right] dt. \quad (3\text{-}23)$$

Then, from Eqs. (3-1) and (3-5),

$$I_o^2 |\gamma_{ij}|^2 = (A_{ij} - C_{ij})^2 + (B_{ij} - D_{ij})^2 \quad . \quad (3\text{-}24)$$

Arguments similar to those of Section 3.1.1 can be used to derive another two estimators. Let

$$\tilde{A}_{ij} = c\,\frac{N^2_{ij1} - N^1_{ij1}}{N^2_{ij1} + N^1_{ij1}}\ , \quad \tilde{B}_{ij} = c\,\frac{N^2_{ij2} - N^1_{ij2}}{N^2_{ij2} + N^1_{ij2}}\ ,$$

$$\tilde{C}_{ij} = c\,\frac{N^2_{ij3} - N^1_{ij3}}{N^2_{ij3} + N^1_{ij3}}\ , \quad \tilde{D}_{ij} = c\,\frac{N^2_{ij4} - N^1_{ij4}}{N^2_{ij4} + N^1_{ij4}}\ . \qquad (3\text{-}25)$$

Then "continuous estimator 1" (C1) is

$$|\hat{\gamma}_{ij}|^2 = \left(\tilde{A}_{ij} - \tilde{C}_{ij}\right)^2 + \left(\tilde{B}_{ij} - \tilde{D}_{ij}\right)^2\ . \qquad (3\text{-}26)$$

Similarly, let

$$\bar{A}_{ij} = \frac{N^2_{ij1} - N^1_{ij1}}{2\hat{I}_o}\ , \quad \bar{B}_{ij} = \frac{N^2_{ij2} - N^1_{ij2}}{2\hat{I}_o}\ ,$$

$$\bar{C}_{ij} = \frac{N^2_{ij3} - N^1_{ij3}}{2\hat{I}_o}\ , \quad \bar{D}_{ij} = \frac{N^2_{ij4} - N^1_{ij4}}{2\hat{I}_o}\ . \qquad (3\text{-}27)$$

Then "continuous estimator 2" (C2) is

$$|\hat{\gamma}_{ij}|^2 = \left(\bar{A}_{ij} - \bar{C}_{ij}\right)^2 + \left(\bar{B}_{ij} - \bar{D}_{ij}\right)^2\ . \qquad (3\text{-}28)$$

At low light levels both of these estimators have a bias term which is proportional to $1/I_o$. As in the previous section, bias correction terms can be added to reduce the bias of these estimators.

3.1.3 Phase Diversity From Atmospheric Turbulence

The third possibility we consider is that the phase error term is $\Psi_{ij}(t)$ caused by atmospheric turbulence. Here the frame integration time Δt is assumed to be short enough that the phase errors within each frame are essentially constant. As discussed earlier in Section 2, this requirement limits Δt to be less than or equal to the coherence time of the atmosphere. Typical values for the coherence time of the atmosphere in the optical regime vary between 5 and 20 ms. Assuming that the phase error is constant over each frame, $\Psi_{ij}(t) = \Psi_{ijk}$ for $t \epsilon [t_{k-1}, t_k]$, we can use the discrete-phase estimators discussed in Section 3.1.1. When Ψ_{ijk}, k=1, ... K is uniformly distributed over the interval $[-\pi, \pi]$ we then have, in the limit for large K,

$$\frac{1}{K} \sum_{k=1}^{K} \sin(\arg \gamma_{ij} + \Psi_{ijk}) \simeq 0$$

$$\frac{1}{K} \sum_{k=1}^{K} \sin^2(\arg \gamma_{ij} + \Psi_{ijk}) \simeq \frac{1}{2} \quad . \tag{3-29}$$

and conditions (3-14) are satisfied.

3.2 ESTIMATOR PERFORMANCE

Here we examine the performance of the estimators described in Section 3.1. An important measure of performance which we focus on is the root mean-squared error. In Section 3.2.1, we derive asymptotic expressions for the bias and squared error which are valid at moderate to high light levels. The low light level performance of the estimators is determined by the use of Monte Carlo simulation. In Section 3.2.2, we derive a lower bound on the expected image-reconstruction error. An important feature of the bound is that it accounts for the object support constraint which is imposed in the reconstruction algorithm.

3.2.1 Estimator Bias and Squared Error

Asymptotic expressions for the normalized bias (NB), normalized standard deviation (NSD) and normalized root mean-squared error (NRMSE) of several of the squared-modulus estimators were derived with the aid of the symbolic-computation program MAPLE [3.16]. For a given squared-modulus sample $|\gamma_{ij}|^2$, the NB and NSD are defined as

$$NB = E\left\{|\hat{\gamma}_{ij}|^2 - |\gamma_{ij}|^2\right\}/|\gamma_{ij}|^2 \qquad (3\text{-}30a)$$

and

$$NSD = \left[E\left\{\left(|\hat{\gamma}_{ij}|^2 - E\left\{|\hat{\gamma}_{ij}|^2\right\}\right)^2\right\}\right]^{1/2}/|\gamma_{ij}|^2 \quad , \qquad (3\text{-}30b)$$

where $E\{\bullet\}$ denotes expectation. The NRMSE can be computed from the NB and NSD as

$$NRMSE = \sqrt{NB^2 + NSD^2} \quad . \qquad (3\text{-}30c)$$

The expressions for the NB, NSD and NRMSE of each of the four estimators are complicated functions of the parameters $|\gamma_{ij}|$, I_o, T, K, and I_B, and are therefore omitted here. The expressions for the unnormalized versions of these quantities and details of their derivation can be found in Appendix A. We plot the NB, NSD and NRMSE as a function of I_oT, since I_oT is the average number of photons detected during time T in a single detector element (i,j) at the output of one of the output arms of the interferometer. The estimate $|\hat{\gamma}_{ij}|^2$, however, is based upon an average of $2I_oT$ photons, since it is based upon the counts detected in both arms of the interferometer.

It is also of interest to consider the mean-squared error of the modulus $|\hat{\gamma}_{ij}|$. Considering only the leading terms in the mean-squared error given in Eqs. (A-3) and (A-8) in Appendix A (i.e., moderately

high light level), c=1 (i.e. no bias exposure) and K=2 frames, the mean-squared error of $|\hat{\gamma}_{ij}|^2$ is $2|\gamma_{ij}|^2/I_o$. By algebraic manipulation it can be shown that this implies that the mean-squared error of $|\hat{\gamma}_{ij}|$ is

$$\text{MSE}\{|\hat{\gamma}_{ij}|\} = \frac{1}{2I_o} \quad . \tag{3-31}$$

Note that this first-order approximation to the mean-squared error of $|\hat{\gamma}_{ij}|$ is independent of the value of $|\gamma_{ij}|$.

Plots of the expressions for NB and NSD in Figures 3-2a and 3-2b show the relative contributions to the NRMS error due to bias and standard deviation respectively, for each of the four estimators with $|\gamma_{ij}| = 0.2$, $I_B = 0.2\ I_o$ and I_oT varying from 10 to 1000 photons. For the D1 and D2 estimators, the photon collection was divided into two frames, with $\theta_1 = 0$ and $\theta_2 = \pi/2$, whereas for the C1 and C2 estimators, four frames were required. As expected, the estimator performance improves as I_oT, the average total number of photons collected per detector element, increases. The bias and standard deviation of the D1 and D2 estimators are nearly identical. A similar trend is observed for the C1 and C2 estimators. The D1 and D2 estimators, which were based on the discrete-phase system, perform better than the continuous-phase system C1 and C2 estimators. For all four estimators, however, the NRMS error is dominated by the standard deviation of the estimator, which has a strong dependence on $|\gamma_{ij}|^2$. This is due to the fact that the estimators are trying to determine the squared difference between the means of the two Poisson random variables, N^1_{ijk} and N^2_{ijk}. This difference is directly proportional to $|\gamma_{ij}|$ [see Equations (3-1) and (3-5)], and as the value of $|\gamma_{ij}|$ decreases, the average difference between N^1_{ijk} and N^2_{ijk} diminishes, causing the standard deviation of the estimate to rise dramatically. Thus, although bias corrections can be easily incorporated into the estimators, they will improve the estimator performance only slightly.

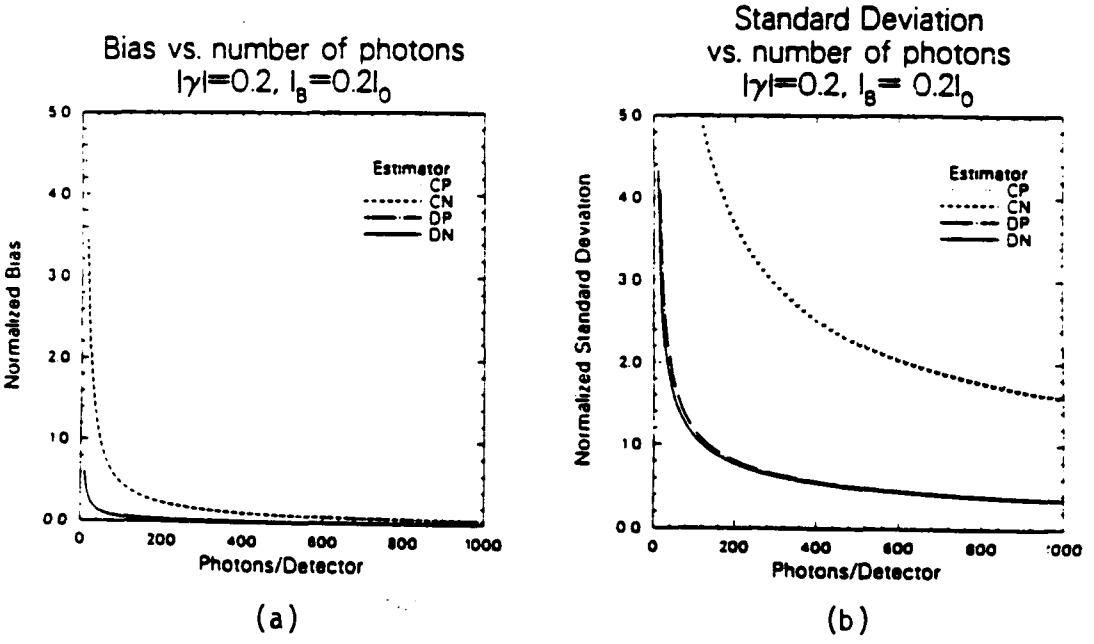


Figure 3-2. (a) Bias and (b) Standard Deviation of the Squared-Modulus Estimators as a Function of the Average Number of Photons per Detector Element (I_0T) for $|\gamma| = 0.2$ and $I_B = 0.2I_0$.

Another measure of performance is the number of photons required to achieve a specified NRMS error in estimating a given squared-modulus sample, $|\gamma_{ij}|^2$. This is illustrated in Figure 3-3 for the D2 estimator with K = 2 frames and $I_B = 0.2I_o$. To achieve a NRMSE of 0.1 when $|\gamma_{ij}| = 0.25$, it would require, on average, 7700 photons per detector. To achieve a NRMSE of 0.5 when $|\gamma_{ij}| = 0.5$, however, requires only 80 photons per detector. On the other hand, if an average of 2000 photons is collected in each detector element, then the NRMS error in estimating modulus values which are greater than 0.5 is less than 10 percent, while the error in estimating modulus values which are less than 0.1 is greater than 50 percent. In general, this would imply that the performance is better for objects which consist of a small collection of points, where the mutual coherence modulus samples are relatively large, than on extended objects, for which the mutual coherence values are small at higher spatial frequencies. At extremely low light levels, the expressions derived for NB and NRMSE are *not* accurate since they are based on low-order asymptotic expansions in $1/I_oT$. Investigations of the estimator performance in the low light regime, $I_o\Delta t \leq 10$ photons, were carried out by the use of Monte Carlo simulation. At each light level, $I_o\Delta t$, and visibility level, γ, 1,000 realizations of the output of a single pair of detectors (N^1_{ijk}, N^2_{ijk}), k=1, ..., K was simulated. Each of the three discrete estimators, D1, D2, and D3 was applied. Then the estimator bias and squared error were then determined from the sample-mean and sample variance of the estimates.

Two scenarios were considered. In the first scenario, a spaced-based interferometer was assumed and a K=2 frame data collection ($\theta_1 = 0$ and $\theta_2 = \pi/2$ in Eq. 3-11) was simulated. In the second scenario, a ground-based interferometer was assumed and a K=20,000 frame collection with a uniformly-distributed phase error term was simulated. Figure 3-4 shows the RMS error in the modulus estimate $|\hat{\gamma}_{ij}|$ (i.e., the square root of $|\hat{\gamma}_{ij}|^2$) for the two-frame

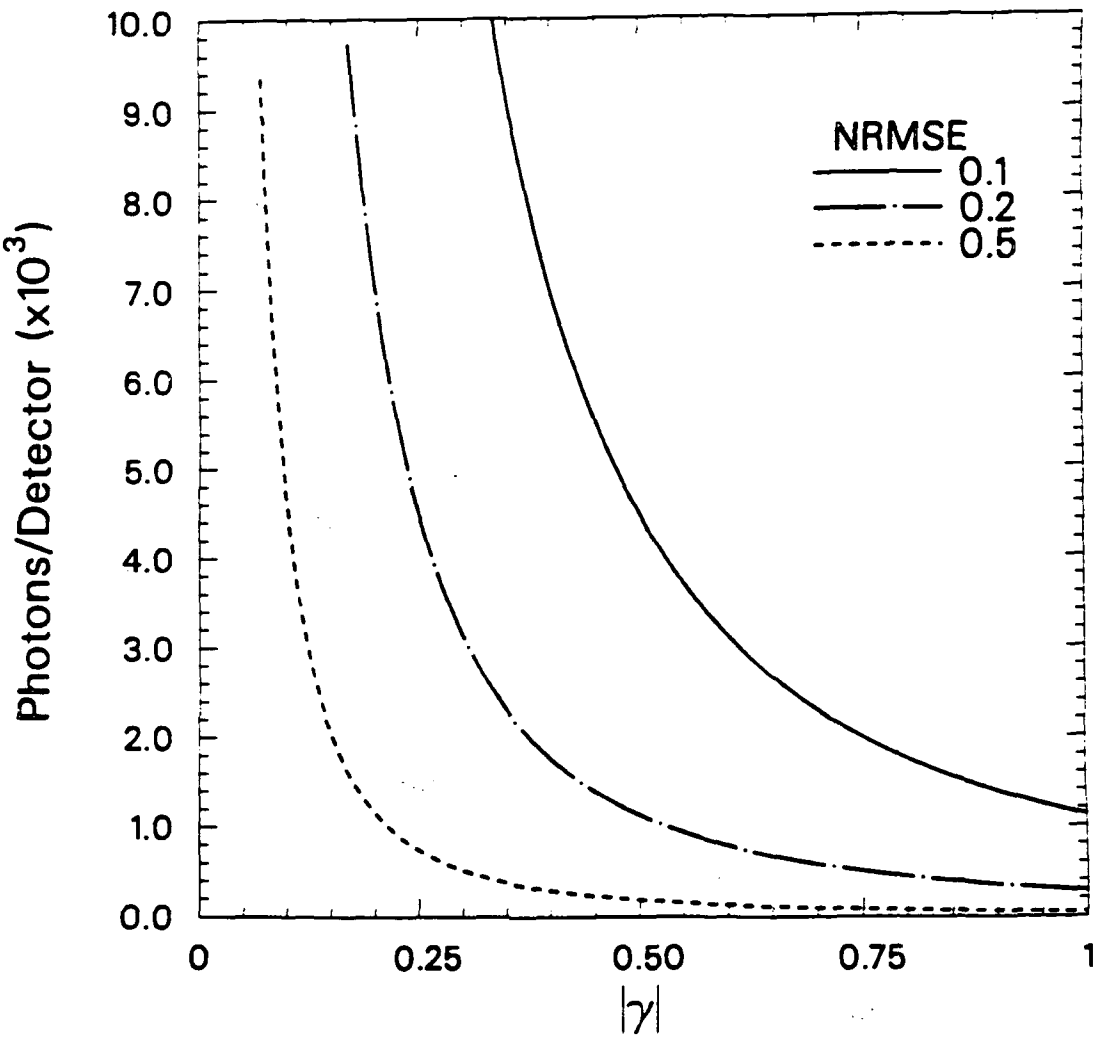


Figure 3-3. The Number of Photons Required per Detector Element to Achieve a Specified NRMSE for the DP Estimator, K = 2 Frames, $I_B = 0.2I_o$ and NRMSE = 0.1, 0.2, 0.5.

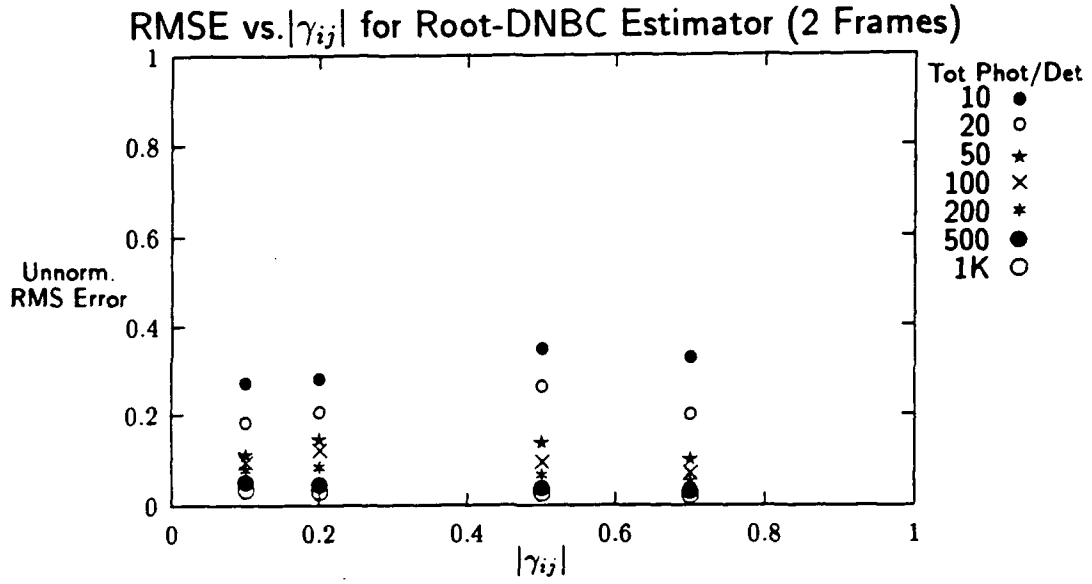


Figure 3-4. RMS Error of the Modulus Estimate $|\hat{\gamma}_{ij}|$ for a Two Frame Collection.

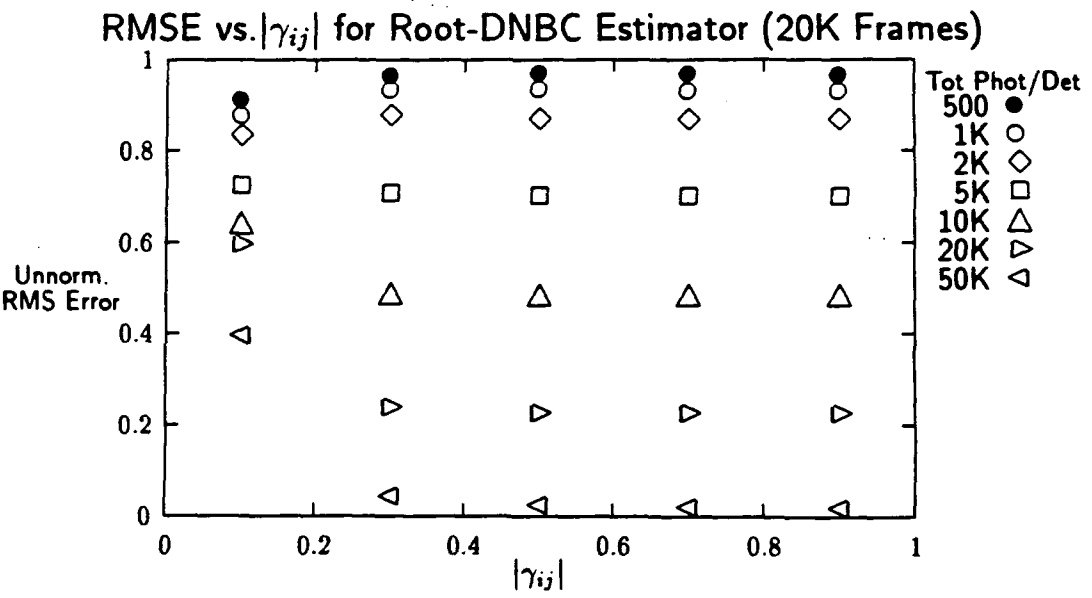


Figure 3-5. RMS Error of the Modulus Estimate $|\hat{\gamma}_{ij}|$ for a 20,000 Frame Collection.

collection at a range of light levels $KI_0\Delta t$ and fringe visibilities $|\gamma|$. Note that the RMS error of $|\hat{\gamma}_{ij}|$ is nearly independent of $|\gamma_{ij}|$, approximately $1/\sqrt{2I_0}$, as predicted. In Figure 3-5 the RMS error of $|\hat{\gamma}_{ij}|$ is shown for the K = 20,000-frame collection. Comparing the two cases, we see that about three orders of magnitude more photons are required in the 20,000-frame collection to achieve a performance comparable to that of the two-frame collection.

3.2.2 Lower Bounds on Image Reconstruction Error

Asymptotic expansions and Monte Carlo simulations were used in Section 3.3.1 to derive explicit expressions and plots of the error in estimating the Fourier intensity components $|\hat{\gamma}_{ij}|^2$. In assessing the performance of the image reconstruction algorithm described in Section 3.3, this approach is not feasible since the algorithm is iterative and nonlinear. Our approach here is to lower bound the image reconstruction error. In this Subsection we present lower bounds on the image reconstruction error for the case of image reconstruction from amplitude interferometer measurements. The bounds derived here are independent of the procedure used to reconstruct the image and thus represent the best possible performance of any such estimator. These bound allow a means of comparing a wide variety of reconstruction algorithms against some "best possible" performance standard.

We will denote the object intensity as $f(x,y)$, where we assume that $f(x,y) \geq 0$ for all x and y in the field of view and that f has finite support. This allows f to be described by samples of its Fourier transform, which we represent in this case by γ_{ij}. By the use of Parseval's theorem, we can then represent the squared error between f and an estimate, say $\hat{f}$, as a function of γ_{ij} and its estimates $\hat{\gamma}_{ij}$:

$$\int\int |\hat{f}(x, y) - f(x, y)|^2 \, dx \, dy = \sum_{ij} |\hat{\gamma}_{ij} - \gamma_{ij}|^2 \quad . \qquad (3\text{-}32)$$

Our strategy is to develop lower bounds on the error term $\Sigma|\hat{\gamma}_{ij} - \gamma_{ij}|^2$.

Appendix B contains a derivation of a Cramer-Rao (CR) type lower bound [3.15,3.21] which incorporates side information. In the present application the side information incorporated into the reconstruction algorithm is the support of the object and its nonnegativity; both are incorporated into the algorithm described in Section 3.3.

Let $\gamma_{ij} = \gamma_{ij}^R + i\gamma_{ij}^I$, where the non-subscripted $i = \sqrt{-1}$. For convenience, we will represent the complex mutual coherence samples γ_{ij}, $i, j = 1, \ldots, N$, by the $2N^2$-length real vector

$$\boldsymbol{\gamma} = \left[\gamma_{00}^R, \gamma_{00}^I, \gamma_{01}^R, \gamma_{01}^I, \cdots \gamma_{ij}^R, \gamma_{ij}^R, \ldots\right]^T . \qquad (3\text{-}33)$$

Denote the estimate by $\hat{\boldsymbol{\gamma}}$. For simplicity we assume $\hat{\boldsymbol{\gamma}}$ is unbiased. In Appendix B a more general result is derived for biased estimators. The CR bound of Appendix B can be expressed as (c.f. Theorem 1 of Appendix B)

$$\begin{aligned} E\{(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma})(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma})^T\} &\geq P(PJ_{\boldsymbol{\gamma}}^{-1} P)^+ P \\ &= Q\, J_{\boldsymbol{\gamma}}^{-1} \end{aligned} \qquad (3\text{-}34)$$

where $J_{\boldsymbol{\gamma}}$ is the Fisher information matrix of $\boldsymbol{\gamma}$, defined by Eq. (13) of Appendix B, T denotes matrix transposition, (+) denotes the Moore-Penrose pseudo inverse (c.f. Eq. (9), Appendix B) and P and Q are projection matrices which depend on the object support (c.f. Eqs. (38) and (50) in Appendix B). In (3-34), Q reflects the amount of improvement afforded by the use of the support constraint. A bound on the total or absolute mean squared error of the image reconstruction can then be found by

$$\sum_{ij} |\hat{\gamma}_{ij} - \gamma_{ij}|^2 = \mathrm{tr}[E\{(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma})\ (\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma})^T\}]$$

$$\geq \mathrm{tr}\left\{Q\ J_{\boldsymbol{\gamma}}^{-1}\right\}\ , \qquad (3\text{-}35)$$

where tr[•] denotes the matrix trace operation.

The bound in (3-35) is directly applicable in the case where the aberration Ψ_{ijk}, i, j = 1, ..., N, k = 1, ..., K, is fixed and nonrandom. If Ψ_{ijk} are unknown or random they are referred to as nuisance parameters. When nuisance parameters are present, calculation of an error lower bound is more difficult. One approach for the case of random nuisance parameters is to determine the minimum lower bound for the worst case nuisance parameters; such a bound is called a minmax lower bound. Another approach which is available when the distribution of the nuisance parameters is known is to derive the Fisher information matrix $\mathfrak{J}$ of the augmented vector $(\boldsymbol{\gamma}, \boldsymbol{\Psi})$, where $\boldsymbol{\Psi}$ is the lexicographical ordering of Ψ_{ijk} into a real-valued KN^2 length vector as in (3-33). One can then form a bound similar to (3-35) based on $\mathfrak{J}$. $\mathfrak{J}$ has the form

$$\mathfrak{J} = \begin{bmatrix} J_{\boldsymbol{\gamma}} & J_{\boldsymbol{\gamma}\boldsymbol{\Psi}} \\ J_{\boldsymbol{\Psi}\boldsymbol{\gamma}} & J_{\boldsymbol{\Psi}} \end{bmatrix}\ , \qquad (3\text{-}36)$$

where for instance $J_{\boldsymbol{\Psi}}$ is the Fisher information associated with the nuisance parameters. The lower bound then takes the form

$$E\{(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma})\ (\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma})^T\} \geq P\left[P\left[J_{\boldsymbol{\gamma}} - J_{\boldsymbol{\gamma}\boldsymbol{\Psi}}\ J_{\boldsymbol{\Psi}}^{-1}\ J_{\boldsymbol{\Psi}\boldsymbol{\gamma}}\right]P\right]^+ P\ . \qquad (3\text{-}37)$$

Equation (3-35) can then be interpreted as the first term in a series expansion of (3-37). Note that (3-35) and (3-37) are equivalent when $J_{\gamma\Psi} = J_{\Psi\gamma}^T = 0$; this is the case when the nuisance parameters are orthogonal to the parameters of interest γ.

We derive the bound of (3-35) for the amplitude interferometer image reconstruction. In light of the discussion above, this bound may be overly optimistic, but it should give an indication of the order of magnitude of the expected image reconstruction error. A straightforward calculation using Eq. (13) of Appendix B for the Fisher information matrix and Eq. (3-7) for the likelihood function yields

$$J_\gamma = (I_o \Delta t) \operatorname*{diag}_{ij}\{B_{ij}\} \tag{3-38}$$

where

$$B_{ij} = \begin{bmatrix} b_{ij}^{11} & b_{ij}^{12} \\ b_{ij}^{21} & b_{ij}^{22} \end{bmatrix} . \tag{3-39}$$

$$b_{ij}^{11} = E\left\{ \sum_k \frac{\sin^2(\Psi_{ijk})}{c - |\gamma_{ij}|^2 \sin^2(\arg \gamma_{ij} + \Psi_{ijk})} \right\}$$

$$b_{ij}^{12} = b_{ij}^{21} = E\left\{ \sum_k \frac{\cos(\Psi_{ijk}) \sin(\Psi_{ijk})}{c - |\gamma_{ij}|^2 \sin^2(\arg \gamma_{ij} + \Psi_{ijk})} \right\} \tag{3-40}$$

$$b_{ij}^{22} = E\left\{ \sum_k \frac{\cos^2(\Psi_{ijk})}{c - |\gamma_{ij}|^2 \sin^2(\arg \gamma_{ij} + \Psi_{ijk})} \right\} \tag{3-41}$$

Here diag$\{B_{ij}\}$ indicates a diagonal matrix with blocks B_{ij} along its diagonal. Also, recall that $c = 1 + I_o/I_B$ and Ψ_{ijk} is given by either (3-9) or (3-10). Calculation of Q is also straightforward but we omit the details here. Let $\hat{S}$ be the Fourier transform of the support constraint. Then

$$\gamma_{ij} = \sum_{i'j'} \hat{S}_{i-i',j-j'} \gamma_{i'j'} \quad . \tag{3-42}$$

This relationship is expressed more compactly as

$$[I - C] \boldsymbol{\gamma} = R\boldsymbol{\gamma} = 0 \quad , \tag{3-43}$$

where I is the $2N^2 \times 2N^2$ identity matrix and C is a symmetric block-circulant matrix with entries $\hat{S}_{i-i',j-j'}$. Q then becomes

$$Q = I - J_{\boldsymbol{\gamma}}^{-1} R\left[R\, J_{\boldsymbol{\gamma}}^{-1} R\right]^{+} R \tag{3-44}$$

and the right-hand side of (3-34) is

$$Q\, J_{\boldsymbol{\gamma}}^{-1} = J_{\boldsymbol{\gamma}}^{-1} - J_{\boldsymbol{\gamma}}^{-1} R\left[R\, J_{\boldsymbol{\gamma}}^{-1} R\right]^{+} R\, J_{\boldsymbol{\gamma}}^{-1} \quad . \tag{3-45}$$

Calculation of the squared-error lower bound of (3-35) requires the evaluation of (3-40) through (3-45) which can be accomplished numerically. As a simple example though, consider the case where no support constraint is in use, Q = I, one frame is collected, K = 1, and where Ψ_{ijk} takes on the values 0 and $\pi/2$ with equal probability. Then $b_{ij}^{12} = b_{ij}^{21} = 0$ in (3-40) and

$$b_{ij}^{11} = \frac{1}{2} \frac{1}{c^2 - |\gamma_{ij}|^2 \cos^2(\arg \gamma_{ij})} = \frac{1}{2} \frac{1}{c^2 - \left(\gamma_{ij}^R\right)^2}$$

$$b_{ij}^{22} = \frac{1}{2} \frac{1}{c^2 - |\gamma_{ij}|^2 \sin^2(\arg \gamma_{ij})} = \frac{1}{2} \frac{1}{c^2 - \left(\gamma_{ij}^I\right)^2} \qquad (3\text{-}46)$$

Substituting (3-46) into (3-39), (3-40), (3-35) and (3-32) results in

$$\int\int |\hat{f}(x, y) - f(x, y)|^2 \, dx \, dy \geq \sum_{ij} \frac{4}{I_o \Delta t} \left[c^2 - |\gamma_{ij}|^2\right] \qquad (3\text{-}47)$$

We see that the absolute squared error is inversely proportional to the average light level per collection frame, $I_o \Delta t$, and is directly proportional to the difference

$$c^2 - |\gamma_{ij}|^2 = 1 + I_B/I_o - |\gamma_{ij}|^2 \qquad (3\text{-}48)$$

This bound increases as the background light level I_B increases or as the squared modulus $|\gamma_{ij}|^2$ decreases: either change causes a decrease in the measurable fringe contrast. Related error behavior is seen in the digital simulations in Section 3.3. In Section 3.3.2 we observe that diffuse objects, those which have smaller fringe visibility values $|\gamma_{ij}|$, are more difficult to reconstruct than objects which contain specular or glinty components.

3.3 DIGITAL SIMULATION EXPERIMENTS

Once the squared-modulus of the mutual coherence has been estimated, an image of the object intensity can be determined by using the fact that the mutual coherence is just the Fourier transform of the object image intensity. Therefore, reconstruction of the object intensity from the squared modulus of the mutual coherence function requires the retrieval of the phase of the mutual coherence function. This phase retrieval can be accomplished with the iterative Fourier transform (IFT) algorithm [3.6,3.7,3.8] using positivity and support

constraints. The IFT algorithm is closely related to the Gerchberg-Saxton algorithm [3.11]. Estimates of the object support are formed from the estimate of $|\gamma_{ij}|^2$ as follows: (i) $|\hat{\gamma}_{ij}|^2$ is inverse Fourier transformed to provide an estimate of the autocorrelation of the object, (ii) the autocorrelation estimate is then thresholded to provide an estimate of the support of the autocorrelation of the object, (iii) an initial estimate of the object support is formed from the autocorrelation support by using a triple-intersection rule [3.2,3.9]. This initial object support depends on thresholded values and thus may exclude parts of the actual object. Hence as the iterations progress, the support constraint is enlarged by including neighboring pixels, thus ensuring that the whole object is eventually contained within the support constraint. Each iteration of the IFT algorithm consists of the following four steps, as illustrated in Figure 3-6: (i) the current object intensity estimate is Fourier transformed to produce an estimate of the Fourier transform of the object, (ii) the modulus of the Fourier transform is replaced by the estimate of $|\gamma_{ij}|$; (iii) the result is inverse Fourier transformed; (iv) the object-domain constraints of positivity and support are enforced using the hybrid input-output algorithm in conjunction with the error-reduction algorithm [3.6,3.7,3.8].

We performed a number of simulation experiments to determine the performance of the IFT algorithm for image reconstruction from low light levels. Three distinct series of simulation experiments were performed. Initially, a series of simple simulations was performed to determine the robustness of the IFT algorithm with respect to Fourier modulus error. Independent and identically distributed Gaussian noise was added to each Fourier modulus sample to approximate the type of measurement error that might occur with the amplitude interferometer. It was found that, for the diffuse object used in the simulation, a useful reconstruction was obtained even at noise levels which gave a Fourier modulus error of 25%. This is described in Section 3.3.1. The

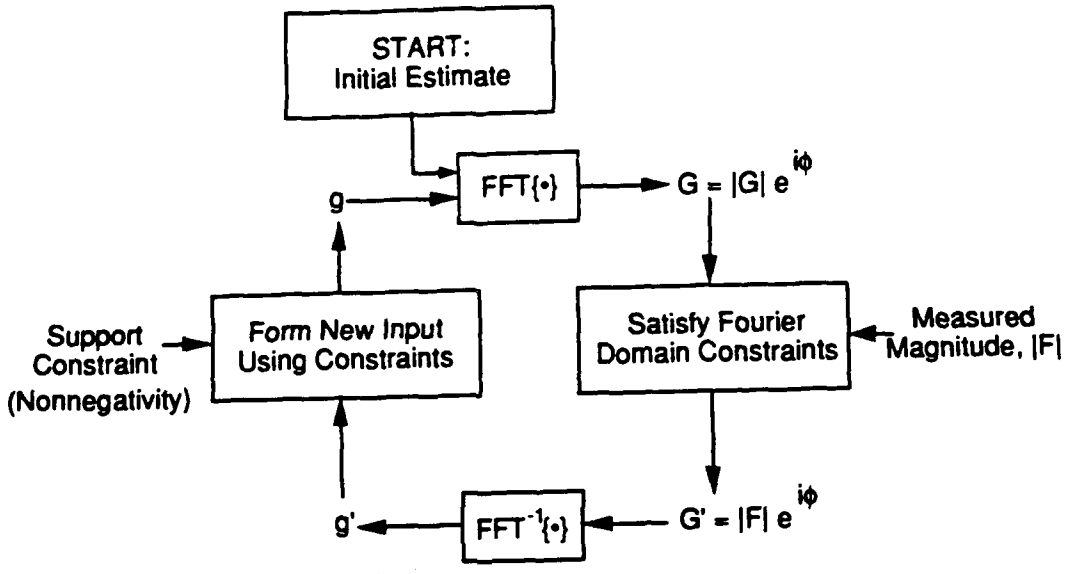


Figure 3-6. Block Diagram of the Iterative Fourier Transform Algorithm.

second series of simulation experiments was performed to demonstrate the performance of the discrete stepped-phase system described in Section 3.1.1 for the case of a two-frame collection, one frame with $\theta(t) = 0$, $t\epsilon[0, T/2]$ the other with $\theta(t) = \pi/2$ $t\epsilon[T/2, T]$. To demonstrate the object-dependent performance of the imaging system, three distinct objects were used: a simple object consisting of four equally-bright points, one being four times the area of the other three; a satellite which had both specular and diffuse components; and a completely diffuse image of a simulated post-boost vehicle (BUS) with several attached re-entry vehicles (RV's) and one detached RV. The general trend we observed was that the specular objects were easier to reconstruct and that reasonable reconstructions were obtained with much less light for specular objects than for diffuse objects. This is described in Section 3.3.2. In the final series of simulation experiments we simulated a ground-based amplitude interferometer which used the effects of turbulent atmosphere to provide phase diversity as described in Section 3.1.3. The goal was to demonstrate the performance of the amplitude interferometer imaging system for the Firefly experiment. Simulations were performed for two cases: a collection using the 48" Cassegrain telescope facility at Goddard and a collection using the ISTEF 24" Cessegrain telescope. This is described later in Section 4.3.

3.3.1 Initial Simulations with Noisy Modulus Data

Clearly, the quality of the reconstructed image has a strong dependence on the accuracy of the squared modulus estimate, $|\hat{\gamma}_{ij}|^2$. Since the IFT algorithm is iterative and highly nonlinear, it is difficult to derive analytically the performance of the IFT as a function of error in the modulus estimate. Empirical simulation studies have shown, however, that the algorithm is robust under certain types of Fourier modulus error [3.5].

As an initial assessment of the viability of the IFT algorithm for image reconstruction from noisy Fourier modulus data, we performed a series of simulations in which Gaussian noise of varying intensities was added to the Fourier modulus of a diffuse object. This was done to approximate the types of Fourier modulus error one could expect from the estimators discussed in Section 3.1. The IFT algorithm was then used to reconstruct an image from each simulated noisy Fourier modulus data and the normalized root mean-squared error of the reconstruction was compared to the error in the Fourier modulus data which was induced by the added Gaussian noise. Figure 3-7 shows the sequence of reconstructed images along with the original object used in the simulation. Figure 3-8 shows the corresponding sequence of Fourier modulus data. Gaussian noise with variances of 400, 1K, 2.5K, 10K, 40K, 100K, 300K, 1M, 3M, 10M, and 30M was added to the modulus data to obtain the Fourier moduli shown in Figure 3-8 (b) through (l). For reference, the peak of the Fourier modulus at DC was 187,793. Figure 3-9 shows a plot of the reconstructed image NRMSE versus the NRMS Fourier modulus error. The image reconstruction error appears to be linear with the Fourier modulus error with an error of approximately 4.5% for the case where no noise was added to the modulus. The small image reconstruction error which occurs at zero Fourier modulus error is most likely due to the "stripe artifact" discussed in Ref. [3.8]. Above 25% NRMS Fourier modulus error, the reconstruction had an error of more than 35%, and the object was barely discernible.

3.3.2 Simulations of a Space-Based Amplitude Interferometer

In the second series of simulation experiments we investigated the performance of the amplitude interferometer assuming the discrete stepped-phase system described in Section 3.1.1, and the estimator D3-BC of Eq. (3-22). The number of frames collected was K = 2. Such a system would be appropriate where the aberration or phase errors are fixed or slowly varying. Here, we assumed that the aberrations were fixed over the collection time.

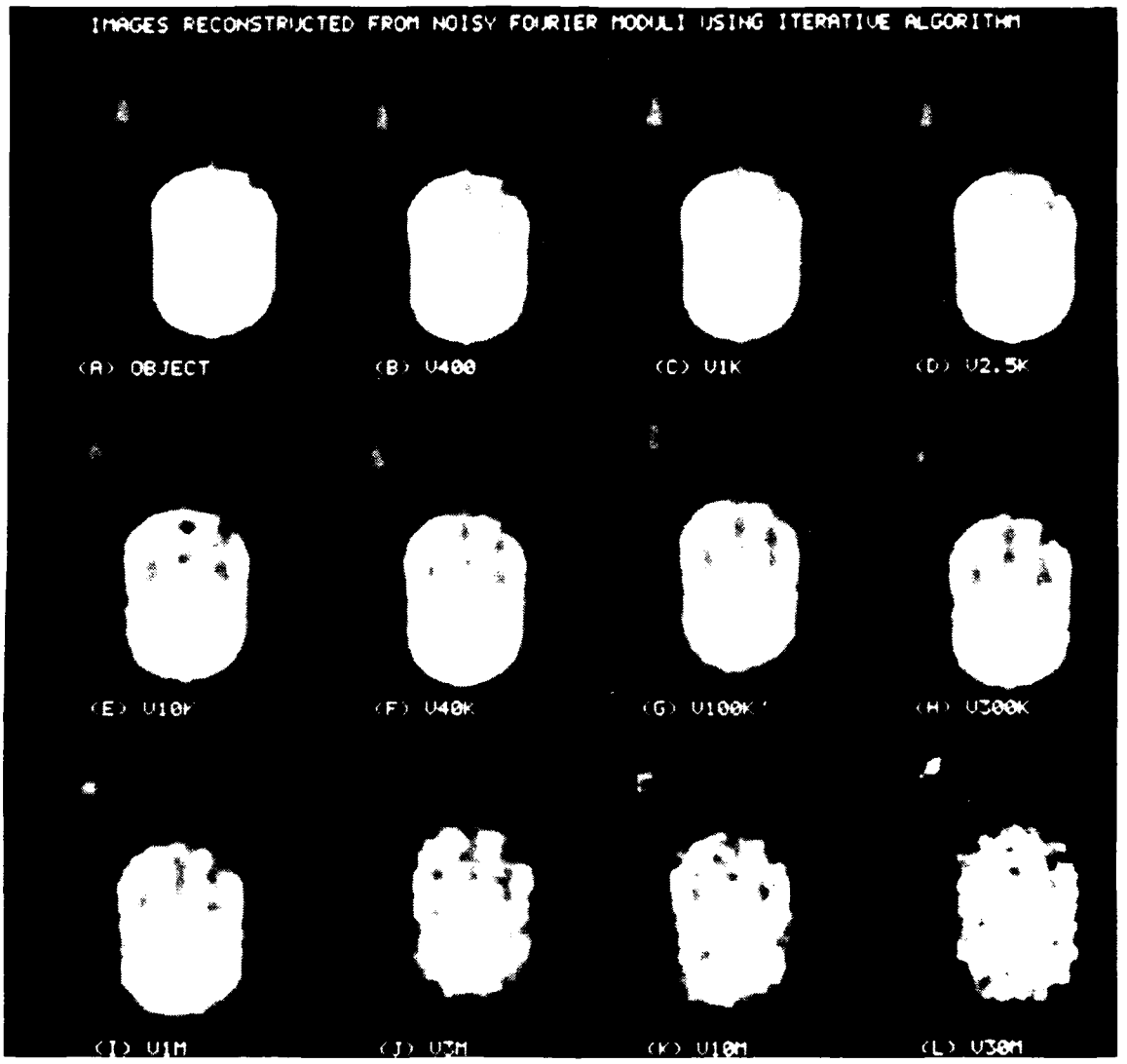


Figure 3-7. Phase Retrieval Image Reconstructions from Noisy Fourier Modulus Data.

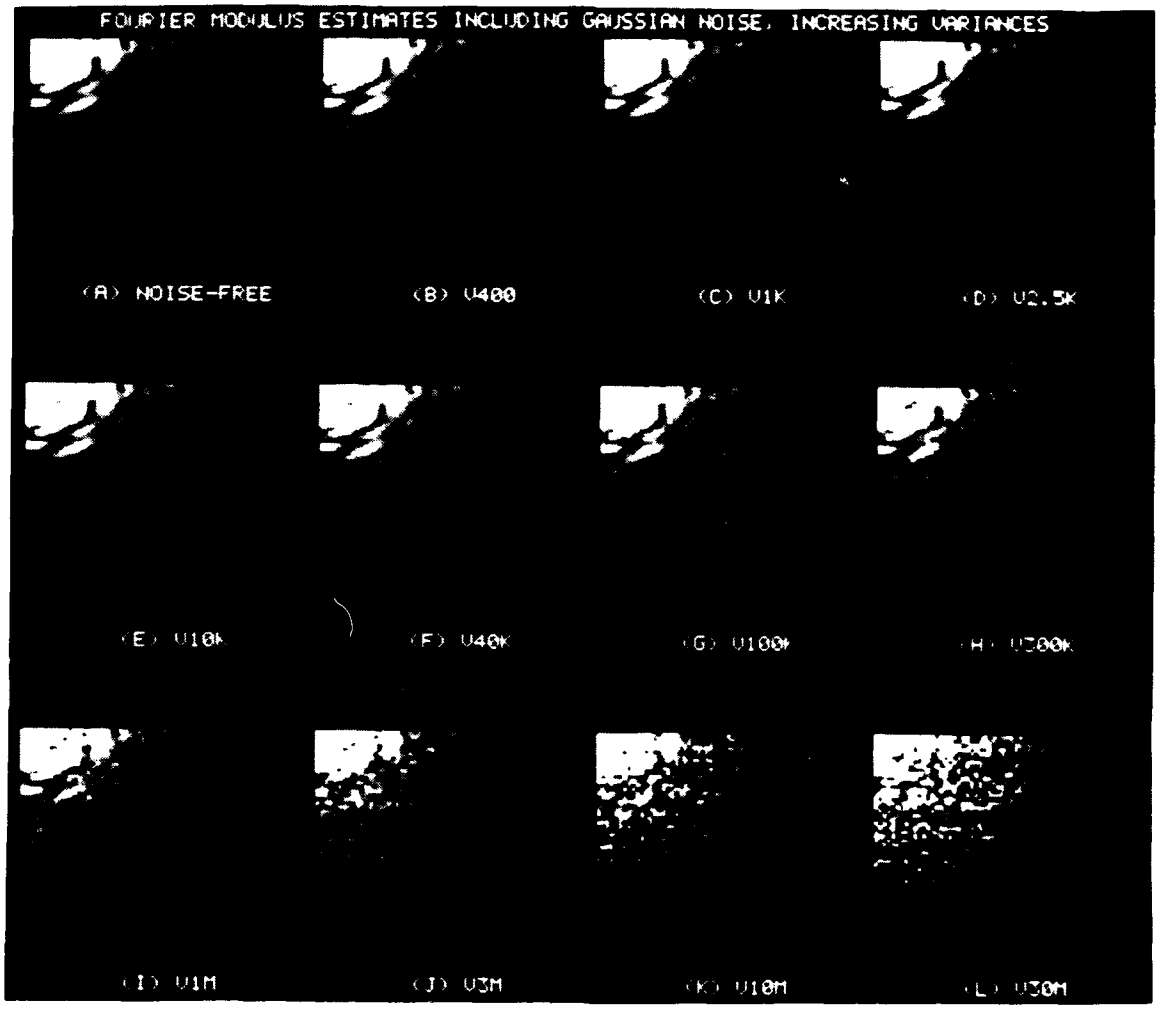


Figure 3-8. Noisy Fourier Modulus Data used in the Reconstructions Shown in Figure 3-7.

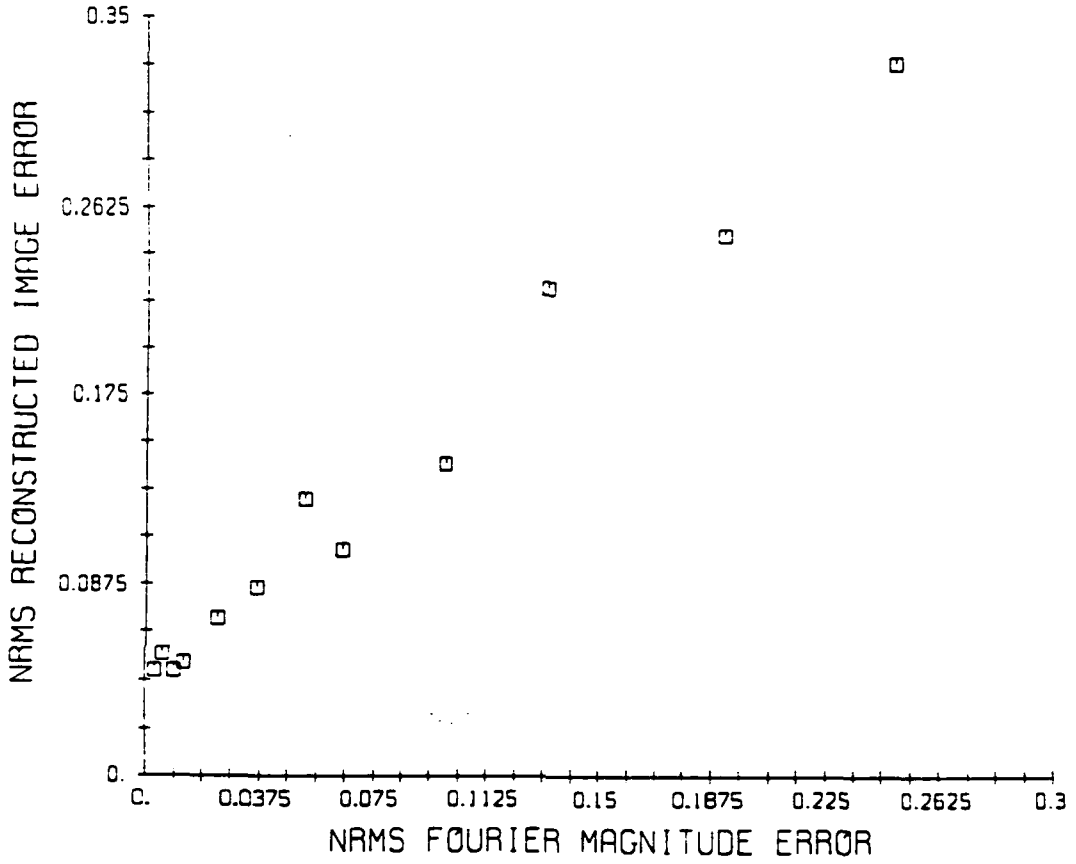


Figure 3-9. Plot of the Absolute Error of the Reconstructions in Figure 3-7 as a Function of Fourier Modulus Error.

Three different objects of increasing complexity were used in the simulation to demonstrate the overall performance of the combined modulus estimation/image reconstruction algorithm as a function of image content. Figures 3-11(a), 3-12(a) and 3-13(a) show the three objects used. Figure 3-10 shows cuts through the spin-averaged Fourier modulus of each object. Each of the objects fits within a 64 x 64 pixel square, and a 128 x 128 array was used in the reconstructions. The object of Figure 3-11(a), called "four points," consists of three equally-bright unresolved points and a fourth part being a 2 x 4 rectangle. Figure 3-12(a), referred to as "satellite," is a model of a communications satellite, and the object of Figure 3-13(a), "Bus/RV," is a simulated post-boost vehicle with several attached re-entry vehicles (RV's) and one detached RV. As shown in Figure 3-10, the Fourier modulus of the "four points" object drops off slowly, while the moduli of the "satellite" and "Bus/RV" objects drop off more rapidly.

In each of these simulation experiments, one realization of a two-frame collection was simulated and an estimate of $|\gamma|^2$ was formed using Eq. (3-22). Next, a reconstruction of the complex mutual coherence (and hence the object itself) was performed by using the iterative Fourier transform (IFT) algorithm [3.5-3.10], using positivity and support constraints.

After the object reconstruction was performed, the absolute squared error between the reconstruction and the original object was measured to provide a quantitative measure of algorithm performance. Since the location of the object within the field of view of the interferometer is not uniquely determined from the modulus estimate, the reconstruction can be translated with respect to the original object. Also, both the object and its 180° rotation have the same Fourier modulus, so the reconstruction can appear rotated by 180° with respect to the original. Therefore the object and reconstruction must be registered before the absolute difference can be calculated. This registration is done by using the procedure described in [3.8].

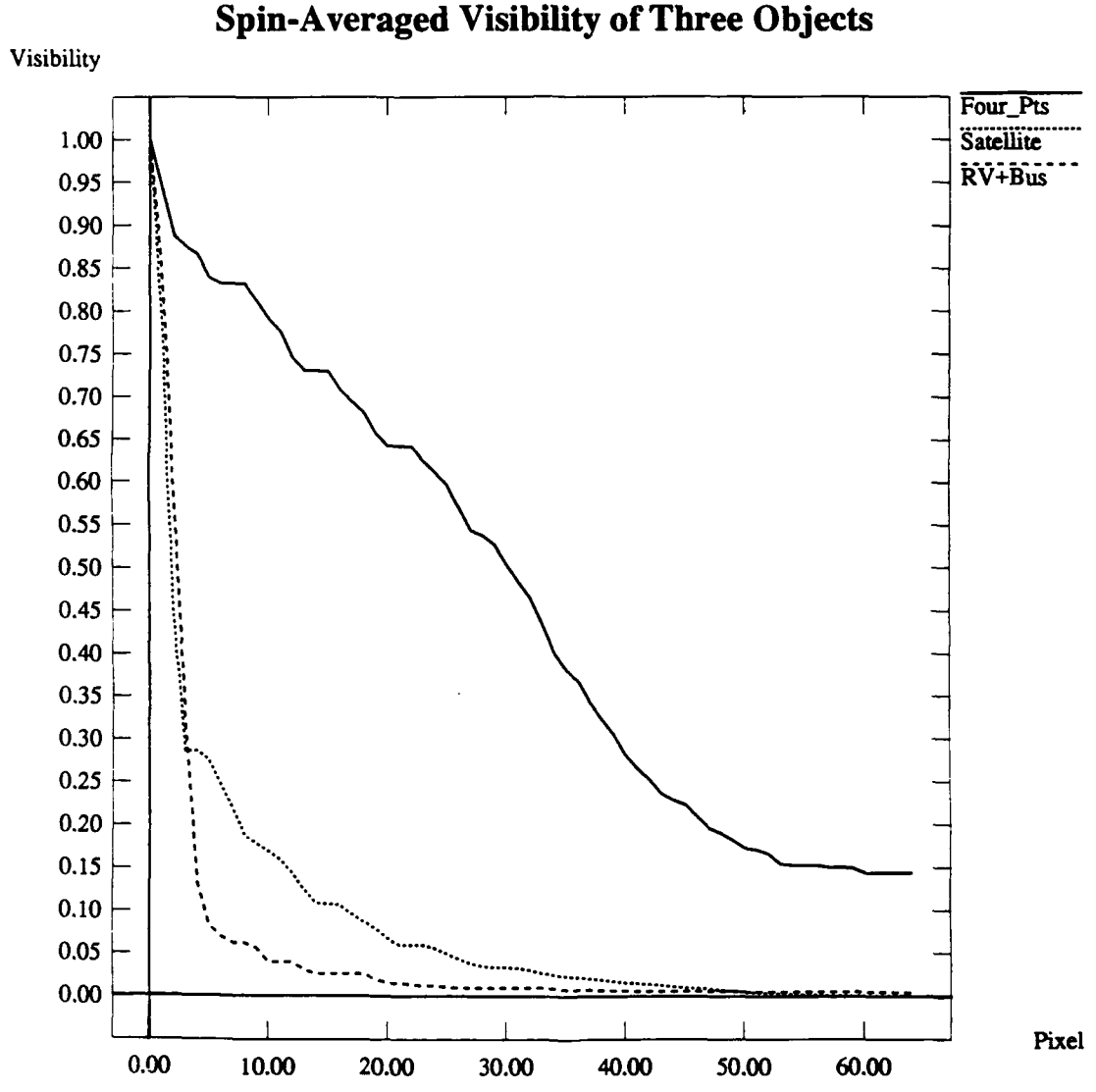


Figure 3-10. A Plot of Cuts through the Spin-Averaged Fourier Moduli of the "Four Points," "Satellite," and "Bus/RV" Objects.

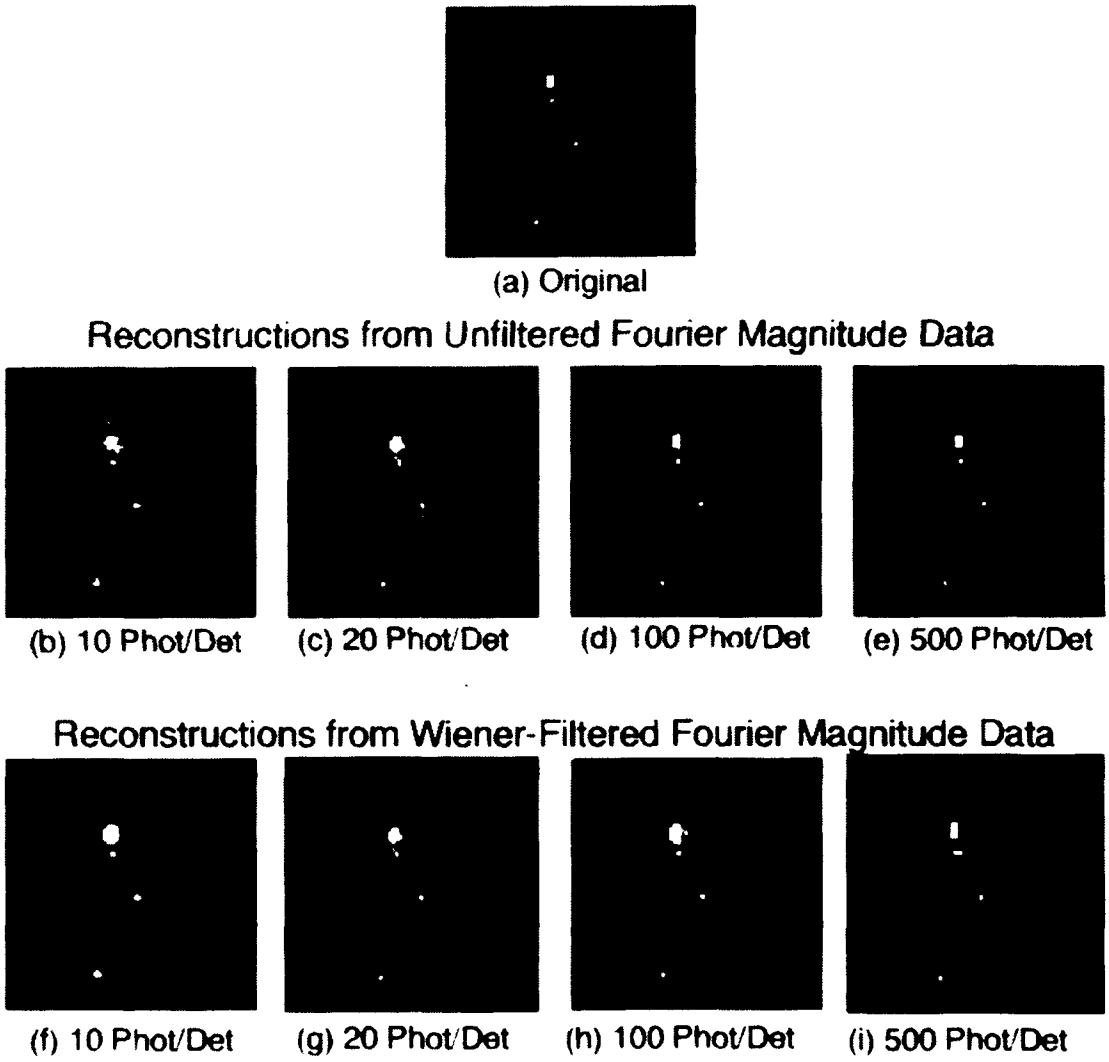


Figure 3-11. Images Reconstructed from Simulated Amplitude Interferometer Measurements of the "Four Points" Object. (a) Object; (b)-(e) images reconstructed from unfiltered Fourier modulus data; (f)-(i) images reconstructed from Wiener filtered Fourier modulus data.

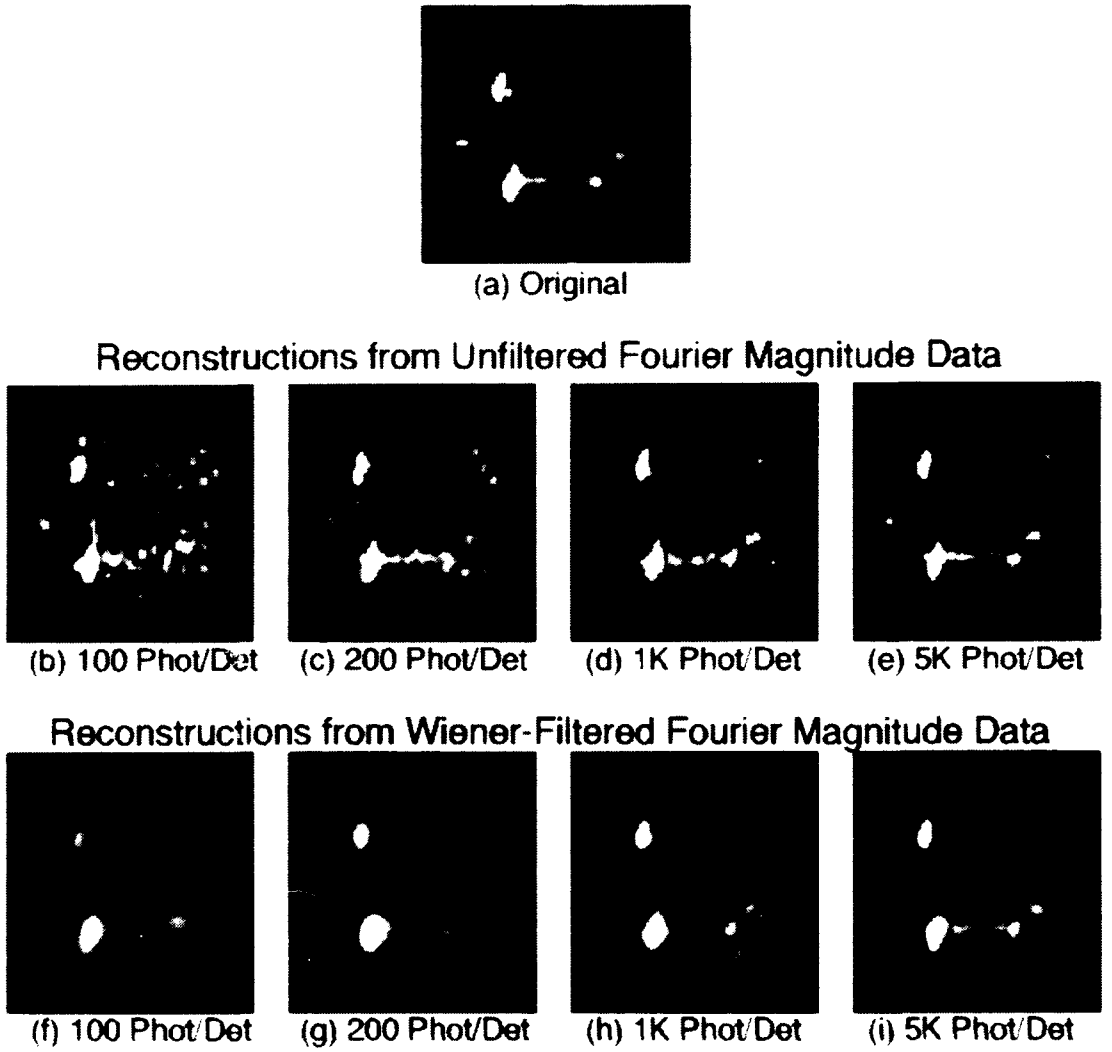


Figure 3-12. Images Reconstructed from Simulated Amplitude Interferometer Measurements of the "Satellite" Object. (a) Object; (b)-(e) images reconstructed from unfiltered Fourier modulus data; (f)-(i) images reconstructed from Wiener filtered Fourier modulus data.

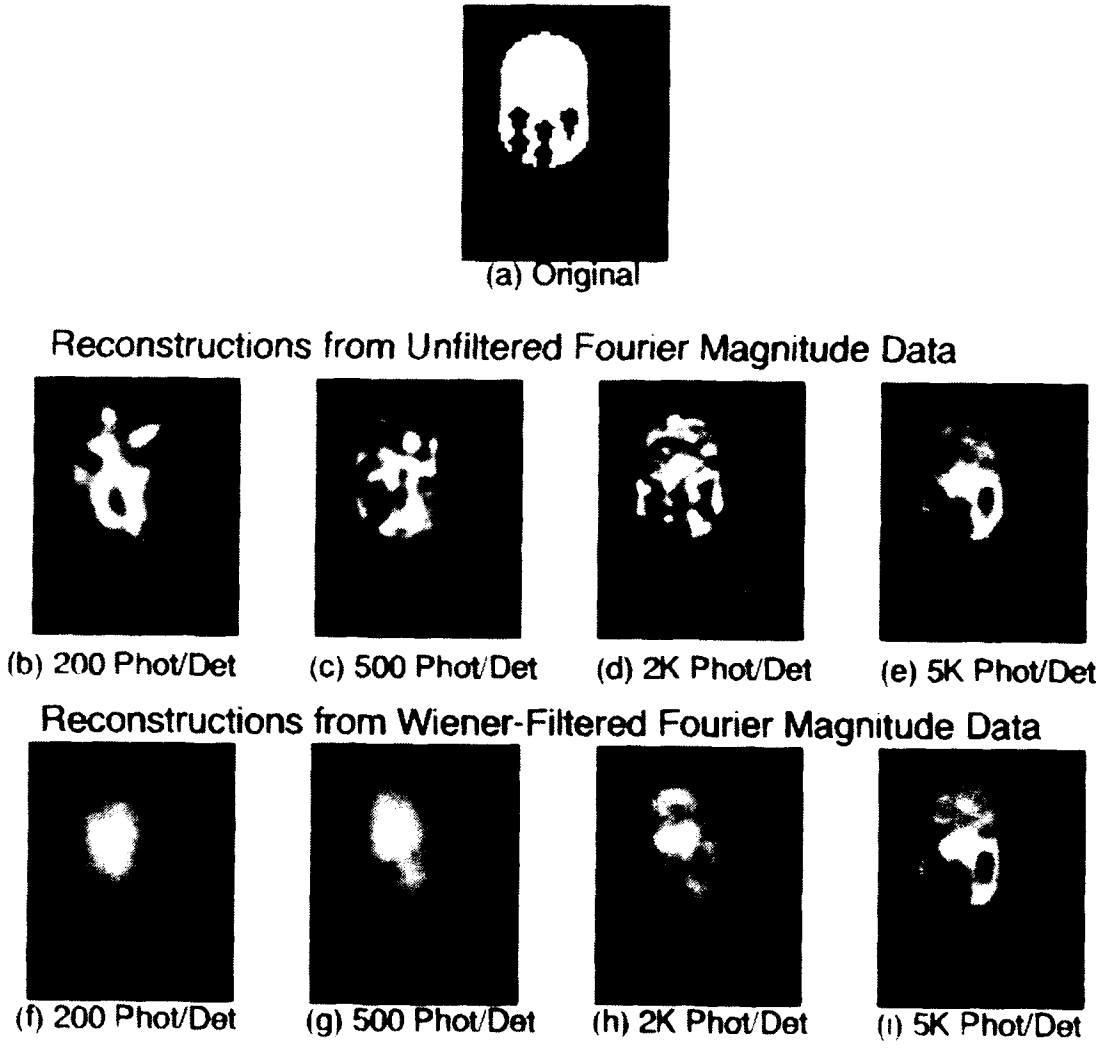


Figure 3-13. Images Reconstructed from Simulated Amplitude Interferometer Measurements of the "Bus/RV" Object. (a) Object; (b)-(e) images reconstructed from unfiltered Fourier modulus data; (f)-(i) images reconstructed from Wiener filtered Fourier modulus data.

Figures 3-11(b)-(d) show reconstructions from simulated two-frame measurements of the "four points" object for the case of $I_0\Delta t$ = 10, 20, 100, and 500 photons per detector per frame. Figures 3-12(b)-(d) show reconstructions of the "satellite" for the case of 100, 200, 1000, and 5000 photons per detectors per frame. Figures 3-13(b)-(d) show reconstructions of the "Bus/RV" for simulations of 200, 500, 2000, and 5000 photons per detector per frame. What we see is that the simpler "four points" object requires far fewer photons for a reasonable reconstruction then the "Bus/RV" object. The locations of the four points can be seen with as few as 10 photons per detector per frame. The "satellite" object, which contains glints, also reconstructs with recognizable features down to 100 photons per detector per frame.

The impact of Wiener filtering the Fourier modulus estimates before reconstruction was also investigated. The Wiener filter has been shown to be the optimal filter in the restoration of images degraded by additive Gaussian noise [3.17] but it also plays a role in iterative image reconstruction algorithms [3.18-3.20]. In the current context, we use the Wiener filter to reduce noise artifacts in the reconstructions which arise from poor estimates of the high spatial-frequency components in the modulus. The proper Wiener filter W requires the squared-modulus $|F|^2$ of the original object and has the form

$$W(i, j) = \frac{|F_{ij}|^2}{|F_{ij}|^2 + \sigma^2} \quad , \tag{3-49}$$

where F_{ij} denotes a sample of the Fourier transform F and σ^2 is the variance of the estimate $|\hat{\gamma}_{ij}|$. Note that, to a first-order approximation, $\sigma^2 = 1/(2I_0)$ independent of $|\gamma_{ij}|$. However, $|F|^2$ is unavailable. As a first pass, we formed a Wiener filter based upon the spin-average $|\bar{F}|^2$ of the Fourier modulus:

$$\bar{W}_{(i,j)} = \frac{|F_{ij}|^2}{|F_{ij}|^2 + \sigma^2} \quad . \tag{3-50}$$

Figures 3-11(f)-(i), 3-12(f)-(i) and 3-13(f)-(i) show the corresponding reconstructions from Wiener-filtered modulus estimates for the three objects. As one would expect, the high-frequency noise artifacts present in the reconstructions from the Wiener-filtered data are greatly diminished, but some of the resolution has also been sacrificed.

A plot of the absolute root mean-squared error of the various reconstructions as a function of the number of simulated photons per detector per frame is shown in Figure 3-14. The Bus/RV object requires two orders of magnitude greater photons than the four points object to get roughly the same image quality. On the other hand, the Bus/RV object has two orders of magnitude more illuminated resolved points than the four points object. Consequently, image quality was similar for the same number of photons per detector per illuminated resolved point on the target.

3.4 SUMMARY

Our proposed method for reconstructing an image from aberrated low-light level aperture-plane amplitude interferometer measurements is to first form an estimate of the squared modulus of the mutual coherence and then to reconstruct a diffraction-limited image by using phase retrieval.

Two amplitude interferometer systems were analyzed in which a controllable phase term $\theta(t)$ was introduced in order to allow measurement of the squared modulus and aberrated phase of samples of the discretized mutual coherence function: one in which $\theta(t)$ took on

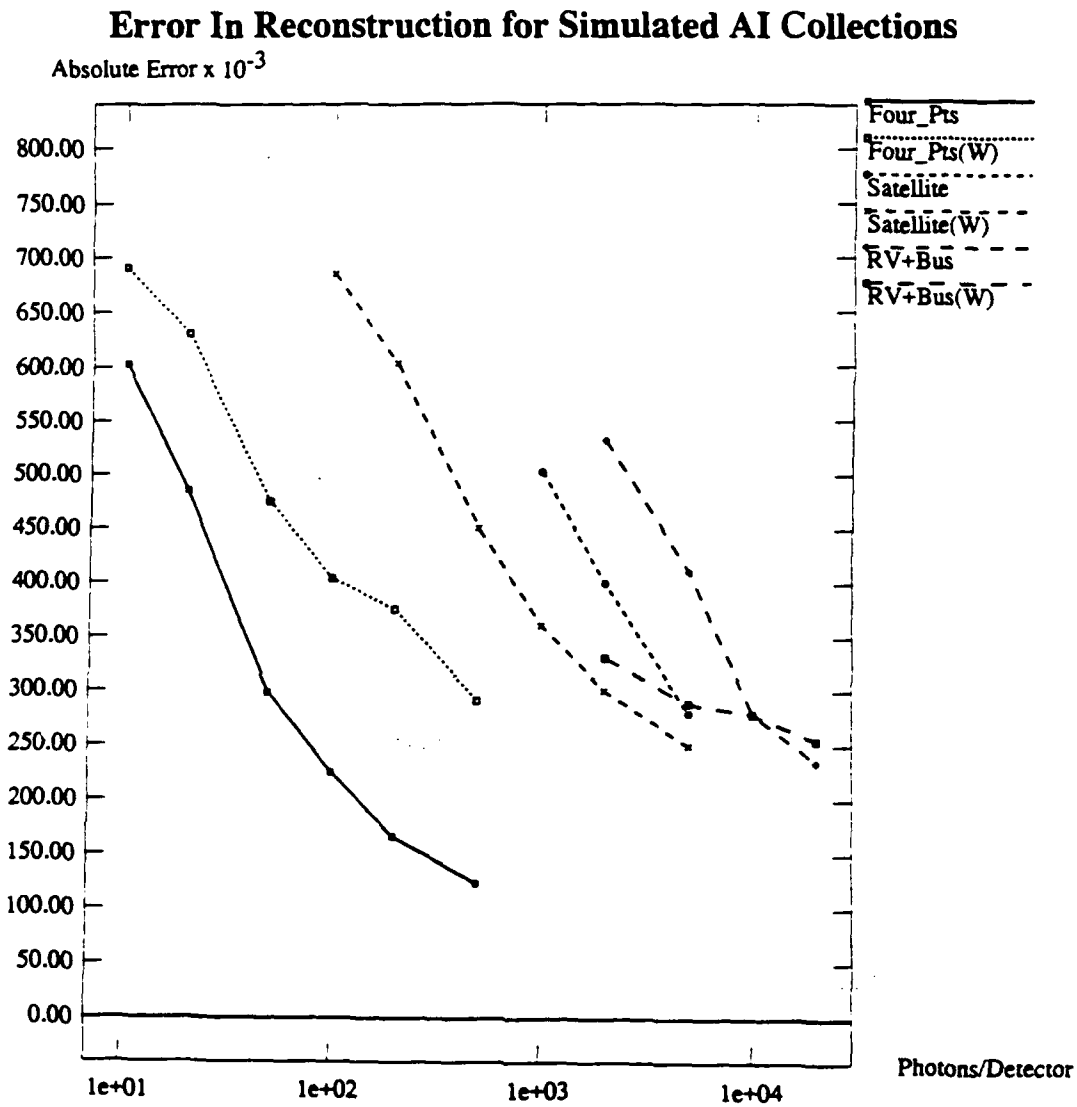


Figure 3-14. Plots of the Absolute RMS Error of the Reconstructed Images Shown in Figures 3-11 through 3-13.

discrete values 0 and $\pi/2$, and the other in which $\theta(t)$ varied linearly over $[0,2\pi]$. It was found that squared-modulus estimators for the discrete-phase system perform better than the estimators for the continuous-phase system. It was also found that the accuracy of squared-modulus estimates has a strong dependence on the value of the squared-modulus, as illustrated in Figs. 3-4 and 3-5, and that the dominant source of error was the standard deviation of the estimator. This standard deviation results from the fact that the estimate relies on the squared difference between the two Poisson random variables, N^2_{ijk} and N^1_{ijk}. The dependence of the performance on the value of the squared-modulus of the coherence functions also results in the performance being much better for point-like objects, for which the coherence function decreases slowly with increasing spatial frequency, than for diffuse, extended objects, for which the coherence function drops rapidly with increasing spatial frequencies.

REFERENCES

[3.1] J.B. Breckinridge, J. Opt. Soc. Am. 65, 755-759 (1975).

[3.2] T.R. Crimmins, J.R. Fienup, and B.J. Thelen, "Improved Object Support Reconstruction from Autocorrelation Support and Application to Phase Retrieval," J. Opt. Soc. Am. A, 7, 3-13 (1990).

[3.3] D.G. Currie, Appendices II and III of "Woods Hole Summer Study on Synthetic Aperture Optics," (vol. 2), Nat. Acad. Sci., Wash. D.C. (1967).

[3.4] D.G. Currie, S.L. Knapp, and K.M. Liewer, Ap. J. 187, 131-144 (1974).

[3.5] G.B. Feldkamp and J.R. Fienup, in 1980 International Optical Computing Conference, W.T. Rhodes, ed., Proc. SPIE 231, 84-93 (1980).

[3.6] J.R. Fienup, Opt. Lett. 3, 27-29 (1978).

[3.7] J.R. Fienup, in Transformations in Optical Signal Processing, W.T. Rhodes, J.R. Fienup, and B.E.A. Saleh, eds., Proc. SPIE 373, 147-160 (1981).

[3.8] J.R. Fienup and C.C. Wackerman, J. Opt. Soc. Am. A 3, 1897-1907 (1986).

[3.9] T.R. Crimmins, J.R. Fienup and B.J. Thelen, J. Opt. Soc. Am. A 7, 3-13 (1990).

[3.10] J.R. Fienup and J.D. Gorman, in Proc. NOAO/ESO Conf., High Resolution Imaging by Interferometry, ESO, Garching bei Munchen, FRG, 15-18 March 1988 (1988).

[3.11] R.W. Gerchberg and W.O. Saxton, Optik 35, 237-246 (1972).

[3.12] K.L. Liewer, PhD thesis, University of Maryland (1975).

[3.13] R. Loudon, The Quantum Theory of Light, 2nd ed, Oxford U. Press, New York (1983).

[3.14] C. Roddier, in Diffraction-Limited Imaging with Very Large Telescopes, ed. D.M. Alloin and J.-M. Mariotti, NATO ASI Series, Kluwer Academic Publishers, The Netherlands (1989).

[3.15] D.L. Snyder, Random Point Processes, J. Wiley and Sons, New York (1975).

[3.16] B.W. Char, K.O. Geddes, G.H. Gonnet, M.B. Monagan and S.M. Watt, MAPLE Reference Manual, 5th ed., WATCOM Publications Ltd., Ontario (1988).

[3.17] C.W. Helstrom, J. Opt. Soc. Am. 57, 297-303 (1967).

[3.18] G.R. Ayers and J.C. Dainty, Opt. Lett. 13, 547-549 (1988).

[3.19] B.L.K. Davey, R.G. Lane and R.H.T. Bates, Opt. Commun. 69, 353-356 (1989).

[3.20] J.H. Seldin and J.R. Fienup, J. Opt. Soc. Am. A 7, 428-433 (1990).

[3.21] Van Trees, H.L. Detection, Estimation, and Modulation Theory, Part I, J. Wiley and Sons, New York (1968).

4.0 PREDICTION OF IMAGE QUALITY FOR FUTURE EXPERIMENTS

In this section we describe analysis, simulation, and reconstruction results that would predict the quality of the imagery that can be expected to be reconstructed from future field experiments. The scenario that was simulated was the imaging of the first Firefly exercise (piggybacking on the MIT Lincoln Laboratory laser radar experiment) launched from Wallops Island as would be viewed by the MAAI attached to the 48-inch telescope at Goddard Space Flight Center. Light levels received by the MAAI, assuming sun illumination of the target, were computed, the detected data was simulated, and images were reconstructed from the simulated data. The simulation results predict that the images produced from the MAAI data from the Goddard 48-inch telescope would be of poor quality. A limiting factor was that the Goddard 48-inch telescope has a large central obscuration, preventing the measurement of the low-to-mid spatial frequencies, where most of the information resides. However, if the low spatial frequencies were measured, then it was shown that good quality imagery could be reconstructed. This could be accomplished by changes in the MAAI (which will be described later) or by using a telescope, such as the ISTEF 24-inch, which has a small central obscuration. Then for the same scenario, images would be reconstructed with resolution far exceeding that ordinarily allowed by atmospheric turbulence. Furthermore, if the same experiment were performed in a space-borne MAAI at the same range, then excellent results would be obtained, even with shorter integration times.

In Section 4.1 we derive expressions for received light levels for the cases of (1) blackbody emission by the target, (2) sunlight reflected by the target and (3) laser illumination reflected by the target. Then we predict the reflected sunlight levels that would be obtained for the Firefly experiment in Section 4.2. In Section 4.3, we comment on the undersampling problem that could occur in the

experiment. In Section 4.4 we show digital simulation and reconstruction experiments that demonstrate the image quality that would be obtained under various assumptions.

4.1 LIGHT LEVEL ESTIMATION - GENERAL CASE

4.1.1 Energy Scattered or Radiated by the Object

There are three cases of interest: objects emitting in the infrared, objects scattering sunlight in the visible or infrared, and objects scattering laser illumination that is of sufficiently short spatial and/or temporal coherence to be effectively incoherent. In the first two cases, the energy must be weighted by the spectral filter which determines the wavelength band to be detected. The total energy is determined by the detector integration time.

Using a blackbody model for infrared emission, the spectral radiance L_s (energy emitted per unit time per unit area per unit solid angle per unit wavelength) of an object is:

$$L_s = \frac{hc^2\epsilon}{\lambda^5[\exp(hc/\lambda kT) - 1]} \qquad (4\text{-}1)$$

where h is the Planck constant, c is the speed of light, ϵ is the object emissivity, λ is the wavelength, k is the Boltzmann constant, and T is the object temperature. Ideally, an integration is required over the surface of the object, including the effects of the angle θ between the local surface normal on the object and the line of sight to the sensor and of variations in the emissivity and temperature, to compute total energy.

For sunlight illumination, the spectral radiance of an object is given by the product of (1) the solar spectral irradiance at the object's altitude, (2) factors depending on the angles between the object's surface normal and (a) the solar illumination direction and (b) the line of sight to the sensor, and (3) the object reflectivity. (Ideally, an integration is required over the surface of the object.) Solar spectral irradiance tables can be found in The Infrared Handbook, Section 3.4 [4.1].

For laser illumination, the energy scattered per unit solid angle is the product of the transmitted laser energy, one way transmittance losses (e.g., due to atmospheric propagation), the ratio of the object cross-sectional area to the laser beam area at the object (including the effect of nonuniform beam intensity), the object reflectivity (again, including nonuniform effects), and the reciprocal of the scattering solid angle. For rough objects, the scattering solid angle can approach 4π steradians. However, for smooth flat objects, the solid angle can be so small as to give a glint, so some care must be taken in estimating this solid angle.

4.1.2 Transmittance Losses

Transmittance losses could be due to propagation through the atmosphere, transmission through the receiver optics, and use of a polarizer.

For pulsed laser illumination, there is an additional loss. The amplitude interferometer can collect data for the entire object only during the time interval over which light is arriving from all parts of the object. For a pulse of length L_p and an object of depth ΔR (along the line of sight to the amplitude interferometer), the fraction of the pulse which may be used (i.e., the pulse utilization efficiency) is $(L_p - 2\Delta R)/L_p$. This factor is unity for emissive or continuously illuminated objects.

4.1.3 Receiver Collection Solid Angle

For fixed image resolution, the collection solid angle of each detector pixel, i.e., the solid angle it subtends with respect to the object plane, is $(d_a/R)^2 = (\lambda/ad_{om})^2$ where d_a^2 is the area of a detector pixel, R is the range to the target, λ is the mean wavelength, a is the desired detector oversampling factor, and d_{om} is the maximum object diameter. For minimum sampling of amplitude interferometer data, a = 2.

This result may be derived as follows. For resolution Δd at the object, the receiver aperture must be of diameter $D = \lambda R/\Delta d$. For an instantaneous field-of-view of diameter (at the object) ad_{om}, where d_{om} is the object's diameter, the Nyquist sample spacing at the aperture plane is $\lambda R/ad_{om}$. Assuming detector elements of width equal to the detector spacing, the solid angle of a detector element is therefore $(\lambda/ad_{om})^2$. There are $D/(\lambda R/ad_{om}) = ad_{om}/\Delta d$ detectors across the aperture.

4.1.4 Parametric Formulas

For thermal emission, the energy per detector E_{det} (i.e., the product of the factors discussed above) is:

$$E_{det} = \frac{hc^2\ \epsilon\ A \cos\theta\ \Delta t\ \Delta\lambda\ \tau_{atm}\ \tau_{opt}\ \tau_{pol}\ \lambda^2}{\lambda^5\ [\exp(hc/\lambda kT) - 1]\ (ad_{om})^2} \qquad (4\text{-}2)$$

where

ϵ is object emissivity
T is object temperature
d_{om} is maximum object diameter
A is object cross-sectional area

λ is mean wavelength
$\Delta\lambda$ is wavelength band
Δt is detector integration time
τ_{atm} is atmospheric transmittance
τ_{opt} is receiver optics transmittance
τ_{pol} is polarizer transmittance
a is the desired oversampling
θ is the angle between the object surface normal and the direction to the sensor

and

h is the Planck constant
k is the Boltzmann constant
c is the speed of light.

Note that all integrations over spatial and wavelength variations have been approximated. For a = 2 (the minimum allowable), $A = \pi(d_{om}/2)^2$, and $\cos\theta = 1$, the formula becomes:

$$E_{det} = \frac{\pi\, hc^2\, \epsilon\, \Delta t\, \Delta\lambda\, \tau_{atm}\, \tau_{opt}\, \tau_{pol}}{16\, \lambda^3\, [\exp(hc/\lambda kT) - 1]} \quad . \tag{4-3}$$

It should be noted that for determination of detected signal-to-noise ratio, the background light level must also be determined and the detectivity D^* of the detector determined.

For sunlight illumination, the energy per detector E_{det} is:

$$E_{det} = \frac{E_\lambda\, A \cos\theta_i \cos\theta_o\, r_{obj}\, \Delta t\, \Delta\lambda\, \tau_{atm}\, \tau_{opt}\, \tau_{pol}\, \lambda^2}{\pi(ad_{om})^2} \tag{4-4}$$

where

E_λ is solar spectral irradiance
d_{om} is maximum object diameter
A is object cross-sectional area
r_{obj} is object reflectivity
λ is mean wavelength
$\Delta\lambda$ is wavelength band
Δt is detector integration time
τ_{atm} is atmospheric transmittance
τ_{opt} is receiver optics transmittance
τ_{pol} is polarizer transmittance
a is desired oversampling
θ_i is the angle between the object surface normal and the solar illumination direction
θ_o is the angle between the object surface normal and the direction to the sensor

and it has been assumed that the object is a Lambertian scatterer. All integrations over spatial and wavelength variations have been approximated. For a = 2, A = $\pi(d_{om}/2)^2$, and $\phi = \theta = 45°$, the formula becomes

$$E_{det} = \frac{E_\lambda \, r_{obj} \, \Delta t \, \Delta\lambda \, \tau_{atm} \, \tau_{opt} \, \tau_{pol} \, \lambda^2}{32} \quad . \tag{4-5}$$

For laser illumination, the energy per detector E_{det} is:

$$E_{det} = \frac{E \, \tau_{atm}^2 \, r_{area} \, r_{obj} \, \eta_{pulse} \, \tau_{opt} \, \tau_{pol} \, \lambda^2}{\Omega \, (a d_{om})^2} \tag{4-6}$$

where

E is the transmitted laser energy
τ_{atm} is the one way atmospheric transmittance
r_{area} is the ratio of object to beam area
r_{obj} is the object reflectivity
Ω is the scattering solid angle
η_{pulse} is the pulse utilization efficiency, $(L - 2\Delta R)/L$
L is the laser pulse length
ΔR is the object depth
d_{om} is the object diameter
τ_{opt} is the receiver optics transmittance
τ_{pol} is the polarizer transmittance
λ is the wavelength
a is the desired oversampling.

Note again that any integrations have been approximated.

4.1.5 Example Calculations

For thermal emission, the energy per detector is 1.2×10^{-15} Joule or 6×10^{4} photons for

$\epsilon = 0.5$
$T = 300°K$ (sun illuminated)
$d_{om} = 5$ meters
$\lambda = 10\ \mu m$ (near blackbody peak)
$\Delta\lambda = 0.5\ \mu m$
$\Delta t = 1$ ms
$\tau_{atm} = 1.0$
$\tau_{opt} = 0.1$
$\tau_{pol} = 0.5$
$a = 2$
$\theta = 0°$.

Note that

$$\frac{hc}{\lambda kT} = 4.81 ,$$

$$\exp (hc/\lambda kT) - 1 = 121 ,$$

and

$$\frac{hc}{\lambda} = 2 \times 10^{-20} \text{ J} .$$

For sunlight illumination, the energy per detector is 0.55×10^{-17} Joule or 15 photons for

$E_\lambda = 1942 \text{ W/m}^2 \ \mu\text{m}$ (exo atmospheric)
$r_{obj} = 0.1$
$\Delta t = 10$ ms
$\lambda = 0.5 \ \mu\text{m}$
$\Delta\lambda = 0.03 \ \mu\text{m}$
$\tau_{atm} = 1.0$
$\tau_{opt} = 0.1$
$\tau_{pol} = 0.5$
$a = 2$
$\phi = 45°$
$\theta = 45°$.

Note that

$$\frac{hc}{\lambda} = 4 \times 10^{-19} \text{ J}.$$

For laser illumination, the energy per detector per pulse is 1.2×10^{-18} Joule or 6 photons for

$E = 1$ Joule/pulse
$\tau_{atm} = 1.0$

$r_{area} = 0.1$
$r_{obj} = 0.1$
$\Omega = 2\pi$
$\eta_{pulse} = 0.75$ ($\Delta R = 5$ m, L = 40 m or 130 nsec)
$\tau_{opt} = 0.1$
$\tau_{pol} = 1.0$ (no polarizer)
$a = 2$
$d_{om} = 5$ meters
$\lambda = 1\ \mu m.$

Note that

$$\frac{hc}{\lambda} = 2 \times 10^{-19}\ J\ .$$

In the above,

$h = 6.63 \times 10^{-34}$ Joule sec
$c = 3 \times 10^{8}$ m/sec
$hc = 1.99 \times 10^{-25}$ Joule m
$k = 1.38 \times 10^{-23}$ Joule/°K
$\frac{hc}{k} = 0.0144$ m°K .

4.2 LIGHT LEVEL ESTIMATION - FIREFLY EXPERIMENTS

In this section we estimate the light level expected from the first Firefly experiment when imaging the large cylindrical object.

For sunlight illumination, the number of detected photons (photoelectrons) per detector per frame is, for a general object,

$$N_{pe} = \eta_q E_{det}/(hc/\lambda) \qquad (4\text{-}7a)$$

$$= \eta_q E_\lambda \Delta\lambda \Delta t(\lambda/hc) r_{obj} (A \cos\theta_i \cos\theta_o/\pi) \tau_{atm} \tau_{opt} \tau_{pol} (d_a^2 \eta_d/R^2) \quad (4\text{-}7b)$$

$$= \eta_q E_\lambda \Delta\lambda \Delta t(\lambda/hc) r_{obj}[(Ld_o/2\pi) V(\Psi_i + \Psi_o)] \tau_{atm} \tau_{opt} \tau_{pol}(d_a^2 \eta_d/R^2). \quad (4\text{-}7c)$$

The parameters in this expression and their estimated values for the Firefly experiment are listed in Table 4-1. Equation (4-7b) was obtained from Eq. (4-4) by replacing the oversampling ratio, a, by

$$a = \frac{\lambda R}{d_a d_{om}} \quad (4\text{-}8)$$

where d_a is the detector spacing and $\eta_d \; d_a^2$ is the area per detector element. Equation (4-7c) is obtained making the further substitution of $(Ld_o/2\pi) V(\Psi_i + \Psi_o)$ for $(A \cos \theta_i \cos \theta_o)/\pi$ for the cylindrical target in the Firefly experiment. The object is assumed to be a cylinder of length L and diameter d_o. In Appendix C the theory of a reflecting cylinder is worked out in detail, and the energy reflected by the cylinder, assumed to be a Lambertian reflector, is proportional to

$$V(\Psi_i + \Psi_o) = (1/2) [\sin(\Psi_i + \Psi_o) - (\Psi_i + \Psi_o) \cos(\Psi_i + \Psi_o)] \quad (4\text{-}9)$$

where Ψ_i is the angle of the sun below the horizontal and Ψ_o is the angle of the receiver below the horizontal, as seen from the target. For the Firefly cylinder, $V(\Psi_i - \Psi_o)$ is about 0.292, as compared with a maximum possible value of $\pi/2 = 1.57$ for illumination from the same angle as the sensor views the object (i.e., for the sun behind the sensor).

Table 4-1
Parameters for Firefly

Parameter Value	Parameter Symbol	Parameter Name
0.10	η_q	detector quantum efficiency
1942 $W/m^2/\mu m$	E_λ	solar spectral irradiance
0.50 μm	λ	mean wavelength
4×10^{-19} J	hc/λ	energy per photon
0.03 μm	$\Delta\lambda$	wavelength band
10 msec	Δt	detector frame integration time
2.4 m	$L = d_{om}$	maximum object diameter
varies	θ_i	the angle between the object surface normal and the solar illumination direction
varies	θ_o	the angle between the object surface normal and the direction to the sensor
0.8	τ_{atm}	atmospheric transmittance
0.056	$\tau_{opt}\ \tau_{pol}$	receiver optics transmittance
600 km	R	range to target
0.4	η_d	fractional active detector area
(3 cm x 4 cm)	$d_{au} \times d_{av}$	detector element center-to-center spacing
0.0009 m^2	$\eta_d d_a^2$	area per detector element
For the cylindrical object:		
0.4 m	d_o	cylinder diameter
2.4 m	L	cylinder length
10°	Ψ_i	solar angle below horizon
55° + 8°	Ψ_o	sensor angle below object plane
107°	$180°-\Psi_i-\Psi_o$	bistatic angle
73°	$\Psi_i + \Psi_o$	180° - (bistatic angle)
0.292	$V(\Psi_i + \Psi_o)$	reflectivity factor (Appendix C)

For the parameters listed in Table 4-1, N_{pe} = 0.1 photo-electrons per detector (in 10 msec). In one second this would be 10 photo-electrons, and in 150 sec of observing time 1500 photo-electrons would be detected. As seen in Table 4-2, 150 sec would be available between times 200 sec and 350 sec from launch, during which period the target would appear to be relatively stationary as viewed from the Goddard site. At most, 3200 photons could be detected during the 320 seconds between times 130 and 450 seconds from launch.

At 3,200 photo-electrons per detector, one can achieve a normalized mean-squared error (NRMSE) of 0.1 (suitable for phase retrieval) for $|\gamma|$ down to 0.5, and one can achieve a NRMSE of 0.5 (suitable for parameter estimation from the Fourier modulus) for $|\gamma|$ down to about 0.1.

4.3 SAMPLING REQUIREMENTS

For the parameters listed in Table 4-1, a = 2.08 if the 3 cm detector spacing direction is oriented along the long axis of the cylinder, but a = 1.56 if it is oriented in the opposite way. Recall that a = 2.0 is required for Nyquist sampling of $|\gamma|^2$. Since this opposite orientation was contemplated, serious problems could arise. For this reason it is worthwhile to review the basis for this sampling requirement.

For a shear of Δu, $\gamma(\Delta u)$ requires a sample spacing of

$$\Delta u \leq \frac{\lambda R}{L} \tag{4-10}$$

in order to avoid aliasing and satisfy the Nyquist criterion, where L is the length of the target. Recall from Section 2 that the interferometer measures $|\gamma(2\Delta u, 2\Delta v)|^2$ for a detector at location

Table 4-2

Firefly Launch Parameters as Viewed from Goddard

Time (sec)	Range (km)	Elevation (deg)	Bistatic angle (deg)	Comment
130	379	50.3	112.5	Rising fast
200	506	55.6	107.5	
350	664	55.4	108.3	Release cannister
450	685	50.7	113.4	Dropping fast

$(\Delta u, \Delta v)$. Therefore a doubling of the sampling rate is required because of the squaring operation (a function squared has twice the bandwidth as the original function), and another doubling of the sampling rate is required because of the 180° rotational shear giving the spacing $(2\Delta u, 2\Delta v)$. Therefore the detector spacing must be

$$d_a \leq \frac{\lambda R}{4L} \tag{4-11}$$

which is 3.1 cm for R = 600 km, $\lambda = 0.5 \times 10^{-6}$ m and L = 2.4 m.

4.4 DIGITAL SIMULATION EXPERIMENTS

A model of the Firefly payload is shown in Figure 4-1. In this case we are imaging the cylindrical object (which later separates into two parts) 2.4 m long and 0.4 m diameter with a nozzle at one end. (The simulated reentry vehicle was judged to be too small and dim for an initial demonstration of amplitude interferometry.) Because of the oblique illumination angle, it would not be realistic to use a digitized version of this photograph as the object for our digital experiments. So instead, we fashioned a three-dimensional shape from wood and painted it white with a black stripe. Shown in the CCD-camera image in Figure 4-2(a), it has features that are similar to those of the Firefly object. Figure 4-2(b) is a photograph of the same object illuminated from below and behind at an angle approximating the one at which the sun would be shining at the Firefly object. At the nearly grazing angle involved, a weak glint on the left half of the object appeared despite the fact that the paint used (Liquid Paper white-out) was not glossy.

Figure 4-2(c) shows the image as would be seen from a diffraction-limited phase-measuring amplitude interferometer (as though there were such a thing) of aperture diameter 1.2 m (48 in), operating at a

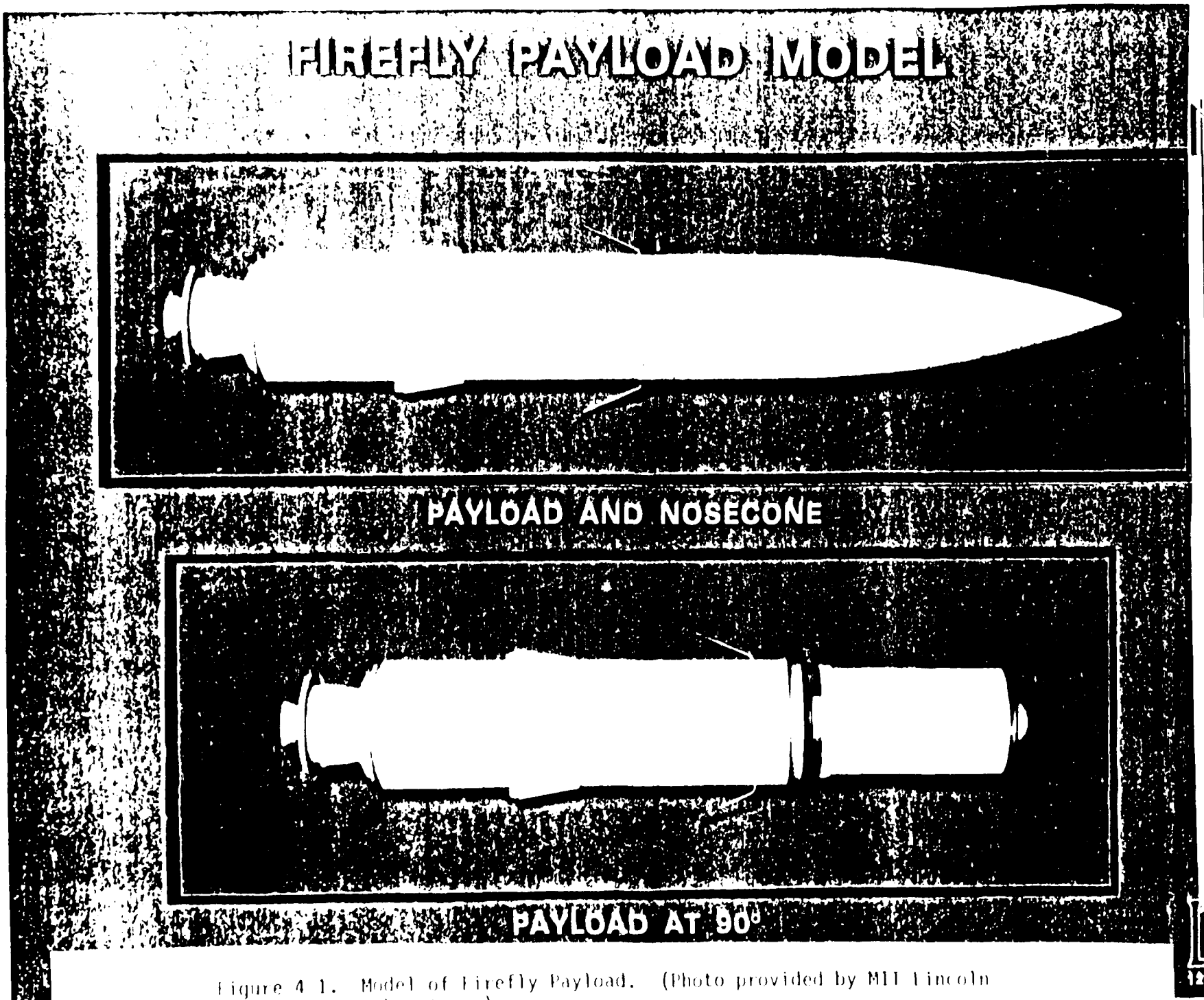


Figure 4 1. Model of Firefly Payload. (Photo provided by MIT Lincoln Laboratory.)

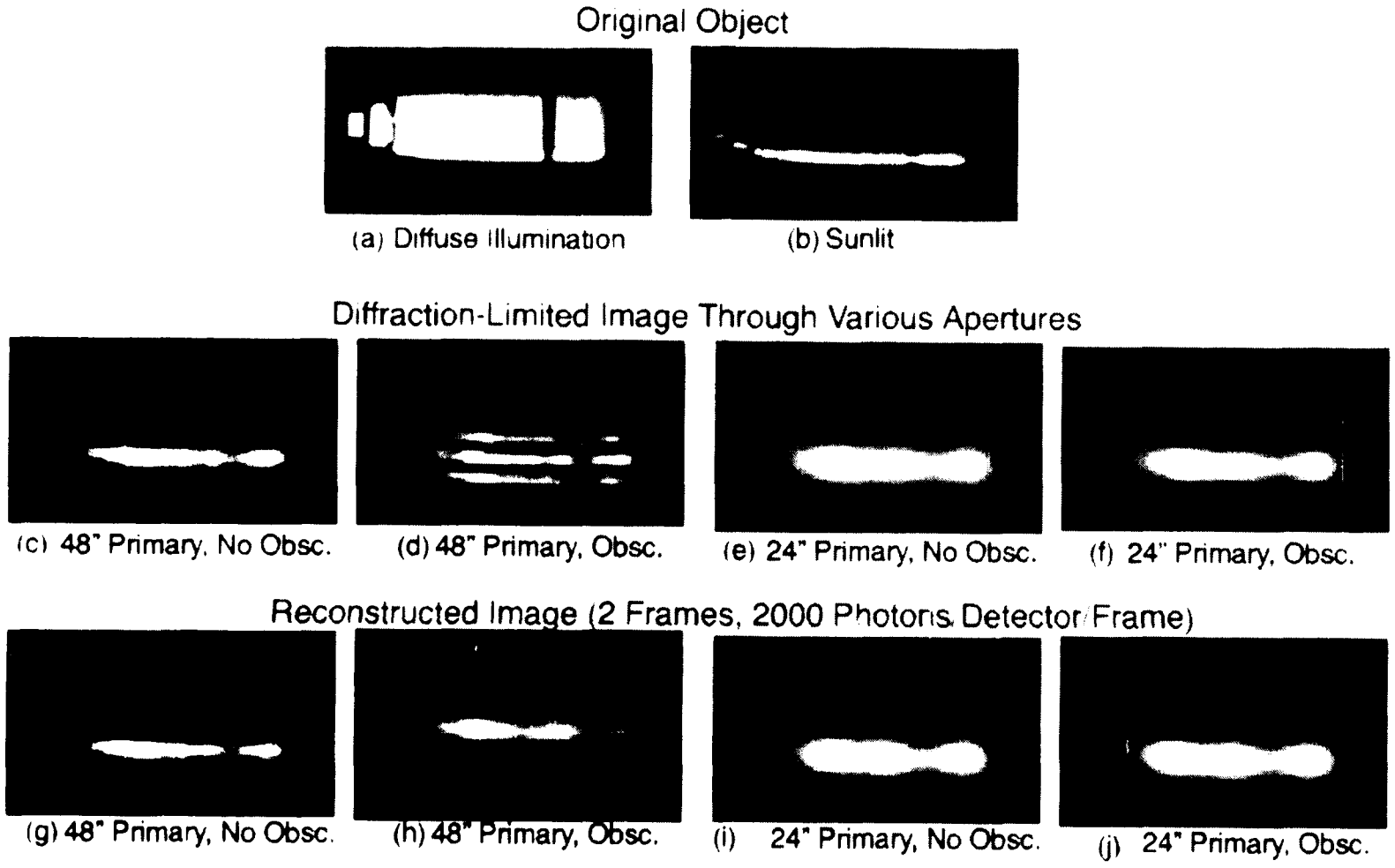


Figure 4-2. Object and Reconstructed Images for Simulation of Space-Based Imaging with the Amplitude Interferometer. (a) ERIM model used for experiments, diffusely illuminated; (b) model illuminated by a spotlight at an angle 107° from the direction of the camera; diffraction-limited images as would be seen through (c) an unobscured 48-inch aperture telescope, (d) the Goddard 48-inch telescope, (e) an unobscured 24-inch telescope, and (f) the ISTEF 24-inch telescope; images reconstructed from two frames of MAAI data with 2,000 photons per detector per frame (as would be appropriate for operation in space), (g) for an unobscured 48-inch telescope, (h) for the Goddard 48-inch telescope, (i) for an unobscured 24-inch telescope, and (j) for the ISTEF 24-inch telescope.

wavelength of 0.5 μm at a range of 600 km, with no noise. This image was obtained by Fourier transforming the object shown in Figure 4-2(b), multiplying by a circular aperture in the Fourier plane of the appropriate size, and inverse Fourier transforming. From this we see that for the large cylindrical object under this illumination condition, even under the most ideal conditions the best that could ever be done with a 1.2 m telescope is to see a thin line that curves upward at one end (where it is thinner at the nozzle) and has a barely discernible dark band near the other end. This illustrates the need for very large apertures for discrimination.

Figure 4-2(d) shows an image that would be obtained from a diffraction-limited phase-measuring MAAI using a 1.2 m aperture having a 0.6 m central obscuration, like the Goddard 48-inch telescope has. Because of the large central obscuration, all the low-to-mid spatial frequencies are not measured -- only the high spatial frequencies are measured, and the result is a high-pass filtered version of the image shown in Figure 4-2(b). The same image features are seen, but very large ringing artifacts dominate the image. The narrow width of the image can no longer be reliably estimated. Discrimination would be difficult with this aperture even with ideal imaging with the phase. To get an image comparable to that shown in Figure 4-2(c), the Fourier data would have to be interpolated from the high spatial frequencies into the mid and low spatial frequencies.

Figure 4-2(e) shows the image that would be obtained from a diffraction-limited phase-measuring MAAI using a 0.6 m (24 inch) filled aperture, and Figure 4-2(f) shows the image that would be obtained from a diffraction-limited phase-measuring MAAI using a 0.6 m aperture with a 0.1 m central obscuration, like a telescope that is available at the Innovative Science and Technology Experimental Facility (ISTEF) on Cape Canaveral. The image is lower in resolution by a factor of two, as expected, but the ringing artifacts are much less pronounced than for

the Goddard 48-inch, since the ISTEF 24-inch has a very small central obscuration.

If the telescope were being operated in space, and if the aberrations were unknown but were slowly varying over the integration time, then the method using only two frames, described in Section 3, could be used. As discussed in Section 4.2, for the first Firefly experiment with the Goddard 48-inch telescope and the then-current implementation of the MAAI, about 1,500 to 3,200 photons per detector could be obtained during the integration time. Data was simulated with 2,000 photons per detector over two frames for each of the four apertures described above. The iterative transform algorithm was used to retrieve the phase over the aperture and, for the annular apertures, simultaneously interpolate the complex values into the mid and low spatial frequencies where no data would be measured. (Section 5.0 and Appendix D describe the algorithm in more detail.) The reconstructed images, shown in Figure 4-2(g)-(j), are comparable in quality to the diffraction-limited images from the filled apertures. In fact, for the 48-inch Goddard annular aperture, the reconstructed image is actually better than a diffraction-limited image with a phase-measuring MAAI [compare Figure 4-2(h) with 4-2(d)]. This results from the success of the interpolation of the mid and low spatial frequencies that would otherwise be lost. This is a remarkable success for the phase retrieval/interpolation algorithm operating on MAAI data.

We also performed experiments with lower numbers of photons, corresponding to proportionally shorter integration times. For only 400 total photons per detector over the two frames, which is 1/5 the light level expected for the Firefly experiment, the major features of the object are still seen in the reconstructed image, although the image is noticeably noisier than the one for 2,000 photons per detector.

For an earth-bound telescope, atmospheric turbulence limits the integration time for a single frame to about 10 msec. Therefore during a 200 sec total integration time, one must collect 20,000 frames of data of exposure time 10 msec each. We simulated 4,000 total photons per detector over the 20,000 frames. Note that this is equivalent to an average of 1/5 photon per detector per frame. That is, most detectors would receive zero photons in a given frame. This data is extremely noisy, to say the least. By summing over 20,000 frames the signal-to-noise ratio is built up. The image reconstructed from this simulated data for the Goddard 48-inch and ISTEF 24-inch annular apertures are shown in Figure 4-3(e) and (h) respectively. For the Goddard 48-inch aperture, large amounts of noise fill the support constraint used during the iterations. A hint of the long, thin object is seen in the image, but the high level of noise would cause one to have little confidence in it. This illustrates the fact that, even if a large number of photons are collected, if they are spread over too many frames, they are not as effective as the same number of photons spread over a small number of frames. The interpolation, which worked well for the case of 2 frames for a space-based sensor, work poorly here since the coherence function squared-modulus estimate is so much noisier. As shown in Figure 4-3(h), the image reconstructed from the same number of photons per detector and the same number of frames, but for the ISTEF 24-inch aperture, is much less noisy and clearly shows the major features of the object although at only half the resolution. This greatly improved result is due to the fact that the much-smaller central obscuration requires far less interpolation. Then the interpolation task is much easier and the image quality is limited only by the aperture size and the performance of the phase retrieval algorithm.

Since the atmospheric "seeing" can be expected to have a correlation distance of about 0.05 meters under these circumstances, the ISTEF 24-inch (0.6 m) image shown in Figure 4-2(h) has resolution

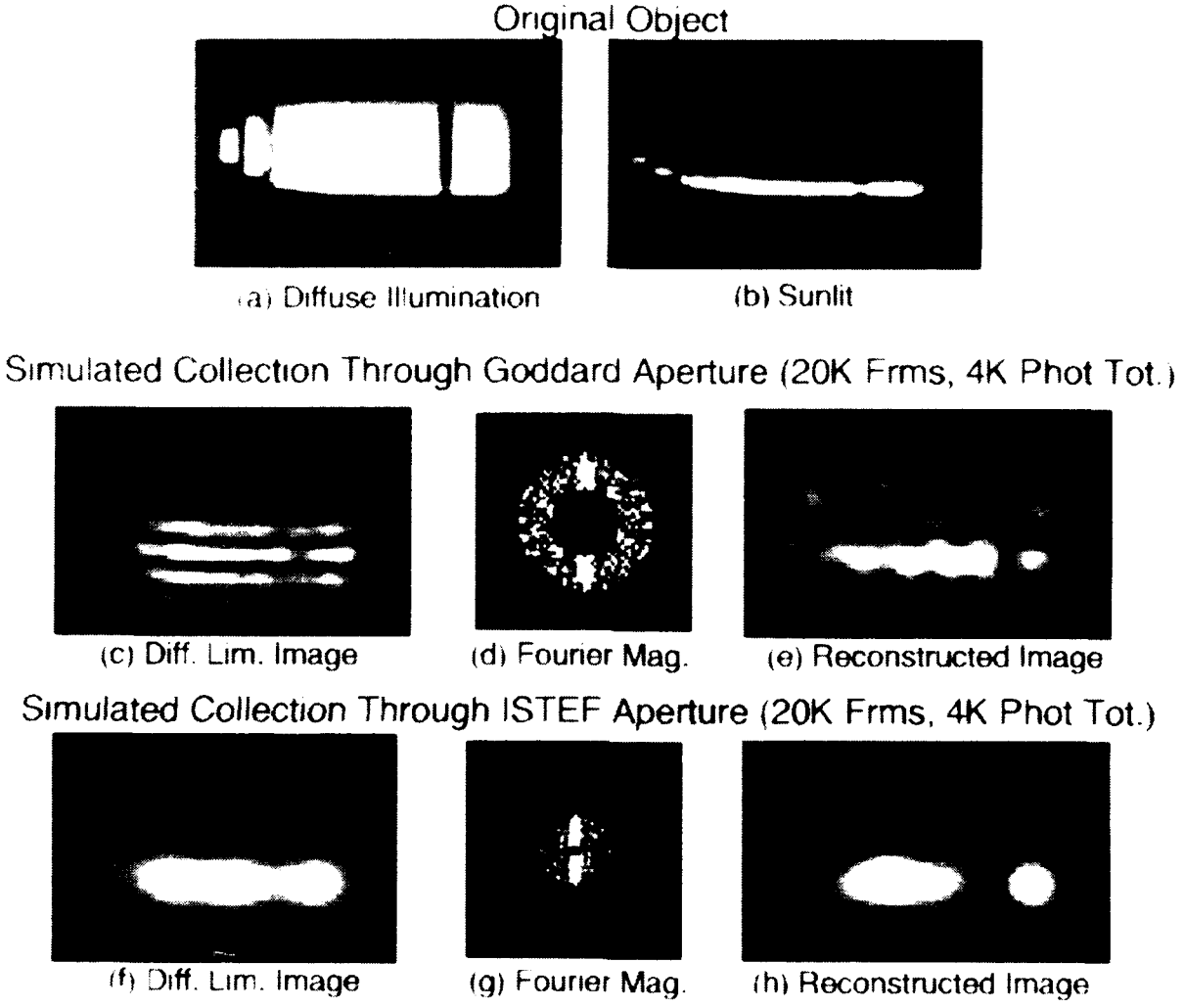


Figure 4-3. Object and Reconstructed Images for Simulation of Ground-Based Imaging through Atmospheric Turbulence with the Amplitude Interferometer. (a) Model diffusely illuminated; (b) model illuminated by spotlight; for the Goddard 48-inch aperture: (c) diffraction-limited image, (d) Fourier modulus, (e) reconstructed image; for the ISTEF 24-inch aperture: (f) diffraction-limited image, (g) Fourier modulus, (h) reconstructed image.

about (0.6m/0.05m) = 12 times better than what would be seen with a diffraction-limited telescope viewing the same object through the same turbulent atmosphere. In fact, the blur circle for atmospheric-limited imaging in this case would be $\lambda R/r_o$ = 6 m, which is 2.5 times wider than the length of the target. Therefore an image of this target from a conventional diffraction-limited telescope would be a large blob showing no detail whatsoever, whereas the image from the MAAI operating with the ISTEF 24-inch would show recognizable features of the object. This demonstrates the tremendous advantage of using the MAAI under the right circumstances.

Figures 4-2(d) and 4-2(g) show the MAAI data (squared-modulus of the coherence function) simulated over the 48-inch Goddard aperture and the 24-inch ISTEF aperture for the ground-based case. The vertical streak down their centers is due to the fact that the target is long and thin in the opposite direction. The holes in the centers are due to the central obscurations of the telescopes. Note that in the horizontal dimension, in which the target is resolved, the signal-to-noise ratio rapidly drops away from the center. This helps to explain why the Goddard aperture worked so poorly. The central obscuration of the Goddard 48-inch is about the same size as the entire 24-inch ISTEF aperture. That is, the annulus of data gathered by the Goddard 48-inch would only start beyond the outer diameter of the ISTEF 24-inch. Since at this point the data has become quite noisy, we see that the Goddard 48-inch would miss the data where the signal-to-noise ratio is good and measure it where the signal-to-noise ratio is primarily poor. For this reason it is important to change the way that the MAAI measures data with telescopes like the Goddard 48-inch -- modifications are necessary to measure the low spatial frequencies, even if it means missing some of the highest spatial frequencies. This is described in Section 6.

REFERENCE

[4.1] W.L. Wolfe and G.J. Zissis, The Infrared Handbook, Revised Edition (Environmental Research Institute of Michigan, 1985), Table 3-17, p. 3-35.

5.0 IMAGE RECONSTRUCTION WITH A PARTIALLY-FILLED APERTURE

For the case of partially-filled aperture, including central obscurations or multiple-mirror telescopes, portions of the spatial frequency domain are not measured. One way to get around this problem is to change the way that the aperture is sheared by the interferometer so that it measures the lower spatial frequencies. When this is done the highest spatial frequencies are lost, but the net image quality can be far higher than what would be obtained with the traditional method of shearing the wavefront. This alternative shearing approach is described in Section 6. If the alternative shearing approach is not taken, then the reconstruction algorithm must simultaneously interpolate the missing phase and modulus values where they are missing while retrieving the phase where the modulus is measured. This is necessary because the impulse response of a partially-filled aperture usually has large sidelobes that go both positive and negative, which interferes with both the support constraint and the nonnegativity constraint used by the phase retrieval algorithm. This is a particularly difficult task if the lower spatial frequencies are missing because of a central obscuration of the telescope, since the visibility modulus at lower spatial frequencies is typically much larger than at the higher spatial frequencies. How we accomplished this and the results are briefly summarized below. A detailed description is given in Appendix D.

The method of simultaneous phase retrieval and interpolation is a modification of the standard iterative transform algorithm. One iteration consists of the usual four steps, but with the following change in the second step in the Fourier domain: where the Fourier modulus is measured, the computed Fourier modulus is replaced by the measured modulus; where the Fourier modulus is not measured but is within the area that would have been occupied by a filled aperture of the same diameter, the Fourier modulus is unchanged; and beyond the area that would have been occupied by the filled aperture, the Fourier modulus is set to zero. If any phase information has been measured in any region, then in that region the computed phase is replaced by the measured phase.

We found that for filled apertures with no phase information, the iterative transform algorithm usually converges reasonably quickly to the correct solution. For a partially filled aperture with no phase information, for which both phase retrieval and interpolation are required, successful reconstructions were obtained, but only when the central obscuration was small. This was for the case of a very extended object. As was seen in Section 4, for a simpler object, reconstructions of this type are also possible with a larger central obscuration if the signal-to-noise ratio (light level) is very high.

We also experimented with interpolation when the phase is measured. Problems with nonunique solutions were encountered if the missing region was large. Therefore the difficulty with combined phase retrieval and interpolation may be limited more by the interpolation than by the phase retrieval in some circumstances.

6.0 ALTERNATIVE AMPLITUDE INTERFEROMETER FOR GROUND-BASED EXPERIMENTS

In order to avoid the problems with the reconstruction algorithms that occur when the telescope has a central obscuration, the way that the aperture is sheared by the interferometer can be changed so that it measures the lower spatial frequencies. When this is done the highest spatial frequencies are lost, but the net image quality can be far higher than what would be obtained with the traditional method of shearing the wavefront. This is important for ground based experiments using existing telescopes, although it would probably not be a problem for an eventual space-based system for which a second small telescope could fill the need for the low spatial frequencies.

The usual geometry for the 180° rotational shear and the detectors is shown in Figure 6-1(A). Only the right half of Figures 6-1(A), (B) and (C) get through one side of the Koster's prism. The annular aperture is rotate 180° about its center and interfered, so that, for a symmetric aperture, the sheared and combined fields occupy the same area as the original aperture. The detector array (shown shaded), on one of the two sides of the Koster's prism, covers only half of the aperture, but that is all that is needed since the coherence function is symmetric about the origin. The low to mid spatial frequencies surrounding the origin in spatial frequency space, indicated by a dot in the figure, are all missing. The low to mid spatial frequencies are measured by either of the alternative geometries shown in Figure 6-1(B) and (C). In these cases the fields are translated horizontally (B) or vertically (C) prior to rotation by 180° so they are rotated about points other than the center of the aperture. For the cases shown in Figure 6-1(B) and (C), the rotations are about points half way between the inner and outer radii of the annulus. That point is the location of the origin of spatial frequency space, and all the low to mid spatial frequencies around it are measured. This can be accomplished simply by shifting the optical axis of the interferometer making it offset with respect to the optical axis of the telescope. For a ratio of radii of 2:1, for the geometry of Figure 6-1(B) in the horizontal direction the highest spatial frequency passed is reduced to

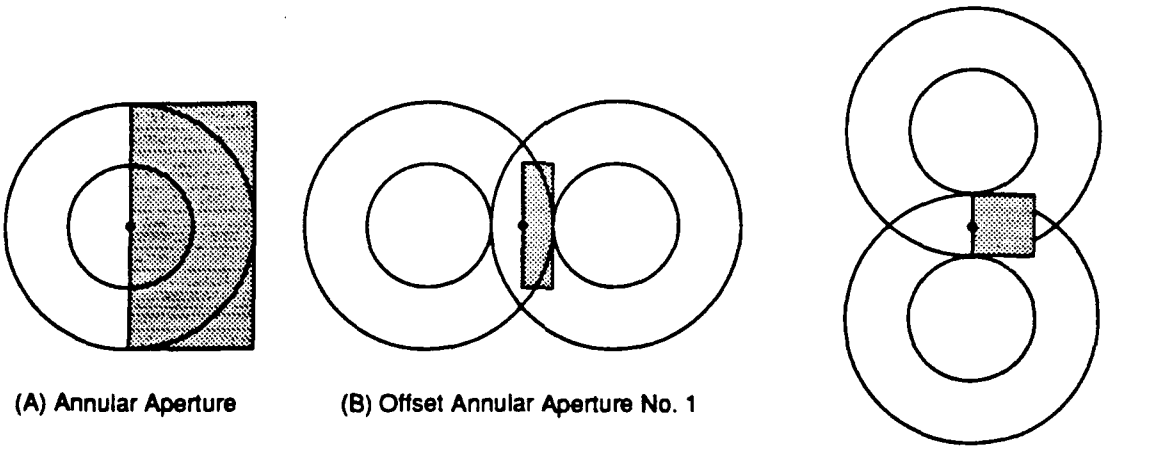


Figure 6-1. Alternative Pupil Shearing and Detection Geometries for Annular Apertures. The shaded rectangles are potential areas for the detector array to cover. (A) Conventional geometry; (B) alternative geometry with horizontal effort; (C) alternative geometry with vertical offset.

1/4 that of the usual geometry, and in the vertical direction the highest spatial frequency passed is $\sqrt{7/16}$ = 0.66 that of the usual geometry. Since the width of the overlap region is narrow near the highest spatial frequencies, the highest practical spatial frequency is about 1/2 that of the usual geometry. This dimension should be oriented along the dimension for which resolution is most important.

For the geometry of Figure 6-1(C), the highest spatial frequency passed is 0.66 that of the usual geometry in the horizontal direction and 1/4 in the vertical dimension compared with the usual geometry, and the detector array is closer to a square shape.

In these cases, for the same number of detector elements, the alternative geometries have twice the field-of-view in each dimension as the usual geometry, and the fraction of detector elements that are used is increased from 58.9% to 82.6%. Most importantly, the low and mid spatial frequencies, where $|\gamma|$ is large, are measured, enabling image reconstruction at much lower light levels.

Another potential operating mode would be to have a system which flips between the two geometries, which could be accomplished with, say, a movable mirror. Then alternately both the low spatial frequencies and the highest spatial frequencies could be measured.

7.0 IMAGE RECONSTRUCTION USING A DECONVOLUTION ALGORITHM

An alternative to the iterative transform phase retrieval algorithm (which was the workhorse algorithm for most of this effort) was developed. It is a version of the Ayers-Dainty blind deconvolution algorithm modified to solve the phase retrieval problem, using support and nonnegativity constraints.

In the blind deconvolution problem, one is given an image g(x) which is the convolution of two arrays, f(x) and h(x), neither of which is known, and both of which we wish to reconstruct from g(x). The Fourier transform, G(u), of g(x) is given by the product of the Fourier transforms of f(x) and h(x):

$$G(u) = F(u)\, H(u) \quad . \tag{7-1}$$

Phase retrieval is a special case of blind deconvolution for which g(x) is the autocorrelation of the object (given by the inverse Fourier transform of the squared Fourier modulus), f(x) is the unknown object, and h(x) is the twin image (hermitian conjugate) of f(x), and we are given $|F(u)|^2 = F(u)\, F^*(u)$. The Ayers/Dainty algorithm iteratively estimates F, f, H, and h by inverting Eq. (7-1) and using constraints, such as support and nonnegativity, on f and h.

Our analysis showed that the algorithm, modified to the phase retrieval problem, has properties similar to the error-reduction version of the iterative transform algorithm. It converges slowly but seems to handle noise well, perhaps due to a built-in Wiener filter that we use to invert Eq. (7-1).

A detailed description of the new phase retrieval algorithm and some results of computer simulations and reconstructions are given in Appendix E.

8.0 NUMERICAL INVESTIGATION OF PHASE RETRIEVAL UNIQUENESS

A question that always arises for image reconstruction by phase retrieval is whether the image obtained is unique. If it were likely that other images were also consistent with the data and constraints, then the method would not be reliable. A new methodology of quantifying the uniqueness of the solution was developed and exercised. The subspace of all ambiguous solutions was analytically derived for the case of small (up to 3 x 3 pixels) images. If an image is a distance from this subspace less than the measurement noise of the Fourier modulus data, then it is consistent with an ambiguous image. If the ambiguous counterpart to the ambiguous image is very different from the original object, then the solution is ambiguous in a practical sense. For 2 x 2 and 3 x 2 images, Monte Carlo experiments were conducted to determine the probability that a random image would lie within a certain distance of this subspace. It involved a reduced-gradient search along the subspace of ambiguous images to determine the ambiguous image closest to a given image. It was found that for small amounts of noise, the probability of having an ambiguous image is small. As the noise level increases, the probability of having a practical ambiguity increases.

The surface of ambiguous images for the 3 x 2 case is five-dimensional, embedded in a six-dimensional space. On the other hand, for the 3 x 3 case, the ambiguous images lie in a seven-dimensional surface embedded in a nine-dimensional space. Since the ambiguous images in the latter case have dimension two less than the space, it seems that they would be far less likely to occur. Therefore, for larger images of practical interest, the probability of ambiguity is probably less than what we computed for the 3 x 2 images.

We also explored the relationship between ambiguous solutions and local minima encountered by phase retrieval algorithms.

The most important aspect of this task was the development of a methodology for determining the probability of uniqueness in a practical

sense. If successfully extended to the larger images of interest, it could yield a practical estimate of probability of ambiguity, and of the reliability of phase retrieval.

A detailed description of this study of the uniqueness of phase retrieval is given in Appendix F.

9.0 ASSESSMENT OF COMPUTATIONAL REQUIREMENTS

The computational requirements for phase retrieval were analyzed. Versions of the algorithm were also sent to other researchers to implement on particular computer architectures, such as the Carnegie-Mellon Warp.

Each iteration of the iterative transform phase-retrieval algorithm involves two 2-D fast Fourier transforms (FFT's) and some additional operations in the two domains. These additional operations include addition, subtraction, multiplication, division, and square root. For some versions of the algorithm it is also necessary to compute sin, cos, arctangent (i.e. conversion between real-imaginary and modulus-phase), logical NOT, and clipping (≥ 0). All of these operations are done independently on 2-D arrays of numbers, so that they are well-suited to vector processor or parallel computing architectures. The 2-D FFT's are similarly well-suited to vector or parallel architectures, since the row (or column) 1-D FFT's can be done in parallel. If fully optimized, the largest computational burden will ordinarily be the FFT's. Since typically dozens to hundreds of iterations are required for convergence, depending on the difficulty of the particular reconstruction problem, the primary computational burden is dozens to hundreds of FFT's to compute a single image. For the SDI discrimination application, all this must be done in a short time, say 20 msec. Consider this example: if the FFT array size were $N \times N = 64 \times 64$, considering that each 2-D FFT requires about $2N^2 \log_2 N$ complex floating-point operations (CFLOP's), the computational rate required to perform 100 iterations (200 FFT's) in 20 msec would be about 500 MegaCFLOP/sec. Consequently, substantial parallelism in the computing architecture is currently necessary to perform these algorithms in the short times allowed. This could be done currently with a Cray Y/MP supercomputer. Efforts are underway to put this level of computing power in a small package (a size less than that of a five-pound coffee can).

Our own computing hardware experience has seen a substantial speed-up over time with machines of comparable cost. A single iteration for array size N x N = 128 x 128 took 1.00 seconds with a Floating Point Systems AP120B array processor in 1980, 0.60 seconds with a Mars Numerix 432 array processor in 1986, and 0.15 seconds with the Carnegie-Mellon Warp computer (tests performed by H.T. Kung's group) in 1988. The latter time was dominated by (a) a non-optimized, slow square root function and (b) a corner-turn required to be performed in the host computer rather than interior to the Warp. With an optimized square root function and a larger memory within the Warp, this time could be reduced by a factor of two.

In a realistic space-based scenario, special-purpose electronic processors would be used instead of the general-purpose processors described above. Typical speed-ups of special-purpose electronic processors over general-purpose processors is typically in the range of 100 times to 1,000 times. Projected general-purpose processors should be adequate for the job. Therefore, if special-purpose electronic processors were developed, then the computational requirements for phase retrieval would be easily achieved.

10.0 LABORATORY EXPERIMENTS

It was intended that images be reconstructed from MAAI data gathered in the laboratory. The data was to be collected by the University of Maryland (UMd) in their laboratories. ERIM prepared test targets of appropriate objects for use in the experiments and delivered them to UMd. The targets were those digitized images shown in Section 2.2. They were written onto fine-grained film using an Eikonix laser-beam recording system. Transparencies, as opposed to reflective objects, were used in order to maximize the intensity of the light that would enter the MAAI. Transparencies were produced at a variety of magnifications in order to match the size requirements of the experimental setup. Special care was taken to make the background density of the transparencies as dark as possible to avoid a background term. No MAAI data was gathered in the laboratory during this effort.

Phase retrieval/image reconstruction software that resided on a Heurikon-hosted Mercury Zip Array Processor at ERIM was delivered to UMd and extensive assistance was given to UMd by ERIM to get the software to work on the Micro-Vax-hosted Zip at UMd. Considerable effort was required to overcome operating system incompatibilities (Unix vs. VMS). This transition was made to enable UMd to perform image reconstruction both at UMd locally and at remote test range sites.

APPENDIX A
EXPRESSIONS FOR BIAS AND MEAN-SQUARED-ERROR

In this Appendix, approximate algebraic expressions are derived for the bias and mean-squared error (MSE) of several of the estimators for $|\gamma_{ij}|^2$ discussed in Section 3. These bias and MSE expressions can then be used to compute the normalized bias (NB) and normalized root mean-squared error given by Equations (3-30). To aid in the computation of expressions for the bias and MSE, the symbolic computation software *Maple* [A-1] was used. Section A.1 contains expressions for the bias and and MSE for four estimators: D1, Equation (3-17); D2, Equation (3-18); C1, Equation (3-26); and C2, Equation (3-28). Listings of the *Maple* sources used to generate the resulting expressions are given in Section A.2

A.1 ALGEBRAIC EXPRESSIONS

The following methodology was used in computing expressions for the bias and MSE associated with estimators D1 and C1. Note that estimator D1 consists of a sum of terms which involve the ratio of the photon difference $N^2_{ijk} - N^1_{ijk}$ to the photon sum $N^2_{ijk} + N^1_{ijk}$, as in (3-17). Similar ratios are required for Estimator C1. Consequently, direct expressions for bias and mean-squared error associated with D1 and C1 are difficult to compute. Instead, to compute the bias and MSE of these two estimators, we use asymptotic expansions for terms involving $(N^2_{ijk} + N^1_{ijk})^{-1}$. The resulting expressions for bias and MSE can then be expressed in terms of a power series in I_0^{-1}. Approximate expressions for bias and MSE are then calculated by truncating the respective series representations after the first few terms. In the expressions below, the first four terms (zeroth, first, second and third-order) are maintained and the resulting accuracy of the expressions for bias and MSE are therefore of order $O(I_0^{-4})$. *Maple* was used as an aid performing the required symbolic computations. In the following expressions, the subscripts denote the estimator, c is defined in (3-13), and the term Q is related to the number of frames

as:

$$Q = \frac{1}{K}\sum_{k=1}^{K}\sin(\arg\gamma_{ij} + \Psi_{ijk}) \quad \text{(A-1)}$$

ESTIMATOR D1:

$$\text{Bias}_{D1} \simeq \frac{I_0^{-1}c}{K} - \frac{\frac{1}{2}I_0^{-1}\gamma_{ij}{}^2}{c} + \frac{\frac{1}{2}I_0^{-2}}{K} - \frac{\frac{1}{4}I_0^{-2}\gamma_{ij}{}^2}{c^2} - \frac{13.375I_0^{-3}\gamma_{ij}{}^2}{c^3} + \frac{\frac{1}{2}I_0^{-3}}{cK} \quad \text{(A-2)}$$

$$\begin{aligned}\text{MSE}_{D1} \simeq\ & \frac{4I_0^{-1}c\gamma_{ij}{}^2}{K} - \frac{8I_0^{-1}\gamma_{ij}{}^4Q}{cK} \\ & - \frac{5I_0^{-2}\gamma_{ij}{}^2}{K} + \frac{6I_0^{-2}\gamma_{ij}{}^4Q}{Kc^2} + \frac{\frac{1}{4}I_0^{-2}\gamma_{ij}{}^4}{c^2} + \frac{3I_0^{-2}c^2}{K^2} \\ & + \frac{\frac{1}{4}I_0^{-3}\gamma_{ij}{}^4}{c^3} - \frac{516.5I_0^{-3}\gamma_{ij}{}^4Q}{c^3K} - \frac{6.5I_0^{-3}\gamma_{ij}{}^2}{cK} + \frac{3.5I_0^{-3}c}{K^2}\end{aligned} \quad \text{(A-3)}$$

ESTIMATOR C1:

$$\begin{aligned}\text{Bias}_{C1} \simeq\ & 2I_0^{-1}c - \frac{1.383728637I_0^{-1}\gamma_{ij}{}^2}{c} \\ & + I_0^{-2} - \frac{0.6918643184I_0^{-2}\gamma_{ij}{}^2}{c^2} \\ & - \frac{45.87288805I_0^{-3}\gamma_{ij}{}^2}{c^3} + \frac{I_0^{-3}}{c}\end{aligned} \quad \text{(A-4)}$$

$$\begin{aligned}\text{MSE}_{C1} \simeq\ & 20.51851925I_0^{-1}c\gamma_{ij}{}^2 - \frac{22.64794786I_0^{-1}\gamma_{ij}{}^4}{c} \\ & - 11.06982909I_0^{-2}\gamma_{ij}{}^2 + 8c^2I_0^{-2} + \frac{4.951018069I_0^{-2}\gamma_{ij}{}^4}{c^2} \\ & + 8.5I_0^{-3}c - \frac{15.50759459I_0^{-3}\gamma_{ij}{}^2}{c} - \frac{901.2612945I_0^{-3}\gamma_{ij}{}^4}{c^3}\end{aligned} \quad \text{(A-5)}$$

An alternative methodology was used in the computation of expressions for the bias and MSE associated with estimators D2 and C2. Expressions for the bias and

mean-squared error for estimators D2 and C2 are simplified by ignoring the variance and higher-order moments of denominator term involving $\hat{I}_0$. As can be seen from Equation (3-6), the standard deviation of $\hat{I}_0$ is inversely proportional to the square-root of the product of the number of pixels N^2, the number of frames K, and the integration time Δt. Furthermore, since $(N^1_{ijk} + N^2_{ijk})$ is a Poisson random variable, its standard deviation is the square-root of its expectation, i.e., $\sqrt{I_0}$. Combining these relationships, the normalized standard deviation of $\hat{I}_0$ is:

$$\text{Std. Dev.}\left(\frac{\hat{I}_0}{I_0}\right) = \frac{1}{N\sqrt{K I_0}}. \tag{A-6}$$

Therefore, for sufficiently large I_0, N and K, we can ignore fluctuations in $\hat{I}_0$ in the computation of the bias and mean-squared error of estimators D2 and C2. As a result, formulas for the first and second moments of estimators D2 and C2 are straightforward to compute. Again, *Maple* was used as an aid in the computation.

Estimator D2:

$$\text{Bias}_{D2} = \frac{I_0^{-1} c}{K} \tag{A-7}$$

$$\text{MSE}_{D2} \simeq \frac{4 I_0^{-1} {\gamma_{ij}}^2 c}{K} + \frac{3 I_0^{-2} c^2}{K^2} + \frac{2 I_0^{-2} {\gamma_{ij}}^2}{K} + \frac{\frac{1}{2} I_0^{-3} c}{K^2} \tag{A-8}$$

Estimator C2:

$$\text{Bias}_{C2} = 2 I_0^{-1} c \tag{A-9}$$

$$\text{MSE}_{C2} \simeq 20.51851925 {\gamma_{ij}}^2 c I_0^{-1}$$

$$+8I_0^{-2}c^2 + 5.129629813I_0^{-2}\gamma_{\cdot\cdot}^2$$
$$+\frac{1}{2}I_0^{-3}c \qquad \text{(A-10)}$$

References:

[A-1] Char, B. W., Geddes, K. O., Gaston, H. G, Monagan, M. B., and Watt, S. M., *Maple Reference Manual*, 5th ed., WATCOM Pub. Ltd., Waterloo, Ont, 1988.

A.2 MAPLE SOURCE CODE

Listings of the *Maple* input used to generate expressions for NB and NRMSE for each of the four estimators considered above are included in this section. The following file `visibility` contains procedures used in all of the computations. *Maple* listings related to each of the estimators D1, D2, C1 and C2 follow.

visibility:

```
#
# File: visibility
# Date: 18 Jul 88
# Author: J. D. Gorman, ERIM
#
# Intent: Computes an expansion for the fringe visibility measurement V
# in terms of two new random variables PSI and ETA, and raises it to
# the Nth power.
# Let NS(K) and ND(K) be the number of photons detected at the
# sinisterous and dexterous  arms of the amplitude interferometer
# respectively so that:
# E{ ND(K) } = I_0 (1 - Gm(K)) = LambdaD(K)
# E{ NS(K) } = I_0 (1 + Gm(K)) = LambdaS(K).
#
# Then we define the random variables:
#
# PSI(K) = {[ND(K)-LambdaD(K)] + [NS(K)-LambdaS(K)]}/ sqrt(2 I_0)
#
# ETA(K) = C*{[ND(K)-LambdaD(K)] - [NS(K)-LambdaS(K)]}
#                 / {Gm(K)*sqrt(2I_0)},
```

```
# where:
# C = {1 + 2*I_Bckgnd}/I_0.
#
# The result is that:
#
# [ND(K) - NS(K)]   [1 + ETA(K)]
# --------------- ≈ ------------.
# [ND(K) + NS(K)]   [1 + PSI(K)]
#
# This ratio is expanded as a series and then the terms of order 0 or
# greater are retained.
#

mean_ratio := proc(K)
result := expand( (Gm(K)^1) * simplify( expectation( visibility(1,K),K ) ) );
end;

mean_ratio_sq := proc(K)
result := expand( (Gm(K)^2) * simplify( expectation( visibility(2,K),K ) ) );
end;

mean_ratio_t := proc(K)
result := expand( (Gm(K)^3) * simplify( expectation( visibility(3,K),K ) ) );
end;

mean_ratio_q := proc(K)
result := expand( (Gm(K)^4) * simplify( expectation( visibility(4,K),K ) ) );
end;

visibility := proc(N, K)
local tmp, result;
tmp := subs( X=psi(K), convert( taylor(1/(1+X),X=0,10), polynom ) );
result := (1+eta(K))^N * tmp^N;
end;

ls := proc(K)
result := (c - Gm(K))/(i0inv);
end;

ld := proc(K)
result := (c + Gm(K))/(i0inv);
```

```
end;

#
#  The following procedures are used to calculate the expectation of
#  various moments of the fringe visibility.
#

etapsi := proc(m, n, k)
local i, j, jnki, jnkj, result, rsum, t1, t2, t3;
option remember;
if type(m,integer) and m > 0 and type(n,integer) and n > 0 then
rsum := 0;
for i from 0 by 1 to m do
for j from 0 by 1 to n do
jnki := i;
jnkj := j;
t1 := binomial(m,i) * binomial(n,j) * (-1)^i;
t2 := pcm(ls(k),i+j) * pcm(ld(k),m+n-i-j);
t3 := ((2.0*Gm(k)/i0inv)^m) * ((2.0*c/i0inv)^n);
rsum := rsum + ((t1 *t2)/t3);
od;
od;
elif type(m,integer) and m > 0 and type(n,integer) and n = 0 then
rsum := 0;
for i from 0 by 1 to m do
jnki := i;
t1 := binomial(m,i) * (-1)^i;
t2 := pcm(ls(k),i) * pcm(ld(k),m-i);
t3 := (2.0*Gm(k)/i0inv)^m;
rsum := rsum + ((t1 *t2)/t3);
od;
elif type(m,integer) and m = 0 and type(n,integer) and n > 0 then
rsum := 0;
for j from 0 by 1 to n do
jnkj := j;
t1 := binomial(n,j);
t2 := pcm(ls(k),j) * pcm(ld(k),n-j);
t3 := ((2.0*c/i0inv)^n);
rsum := rsum + ((t1 *t2)/t3);
od;
fi;
```

```
result := rsum;
end;


subetapsi := proc(X,K)
local i, result, t1, t2;
option remember;
t1 := X;
for i from 12 by -1 to 1 do
for j from 12 by -1 to 1 do
if i > 5 or j > 5 then
t2 := subs( eta(K)^i * psi(K)^j = 0, t1 );
else
t2 := subs( eta(K)^i * psi(K)^j = etapsi(i,j,K), t1 );
fi;
t1 := t2;
od
od
end;

subpsi := proc(X,K)
local i, result, t1, t2;
option remember;
t1 := X;
for i from 12 by -1 to 1 do
if i > 5 then
t2 := subs( psi(K)^i = 0, t1 )
else
t2 := subs( psi(K)^i = etapsi(0,i,K), t1 );
fi;
t1 := t2;
od
end;

subeta := proc(X,K)
local i, result, t1, t2;
option remember;
t1 := X;
for i from 12 by -1 to 1 do
if i > 5 then
t2 := subs( eta(K)^i = 0, t1 )
```

```
else
t2 := subs( eta(K)^i = etapsi(i,0,K), t1 )
fi;
t1 := t2;
od
end;

expectation := proc(X,K)
local result, t1, t2;
option remember;
t1 := expand(X);
t2 := subetapsi(t1,K);
t1 := subeta(t2,K);
t2 := subpsi(t1,K);
result := t2;
end;


#
#  This procedure calculates the Nth central moment of a Poisson
#  random variable having parameter X.
#

pcm := proc(X,N)
local result, Y, tmp;
if type(N,integer) then
if N = 0 then result := 1
elif N = 1 then result := 0
elif N = 2 then result := X
elif N = 3 then result := X
elif N = 4 then result := X + 3*X^2
elif N = 5 then result := X + 10*X^2
elif N = 6 then result := X + 25*X^2 + 15*X^3
elif N = 7 then result := X + 56*X^2 + 105*X^3
elif N = 8 then result := X + 119*X^2 + 409*X^3 + 105*X^4
else
tmp := Y*N*pcm(Y,(N-2)) + diff( pcm(Y,(N-1)), Y );
result := subs(Y=X, tmp);
fi;
fi;
result
```

```
end;
```

Estimator D1:

```
read( visibility );

#
# File: Ncurrie_mse
# Date: 19 Oct 88
# Author: J. D. Gorman
#

mean_ratio_sq_K := mean_ratio_sq(K):

#
#  Calculate BIAS of Discrete-Phase Normalized Estimator
#

sum_mean_ratio_sq := proc()
local tmp1, tmp2, result;
tmp1 := expand( mean_ratio_sq_K );
tmp2 := subs( Gm(K)^2 = GM^2*(nframes/2), tmp1 );
result := tmp2;
end;

bias := simplify( ((2/nframes)*sum_mean_ratio_sq()) - GM^2 );

#
#  Calculate Terms in MSE
#

sq_sum_mean_ratio_sq := proc()
local tmp1, result;
tmp1 := sum_mean_ratio_sq();
result := expand( tmp1^2 );
end;

sum_sq_mean_ratio_sq := proc()
local tmp1, tmp2, tmp3, tmp4, result;
tmp1 := mean_ratio_sq_K;
tmp2 := expand( tmp1^2 );
tmp3 := subs( Gm(K)^4 = GM^4*(qsum*nframes), tmp2 );
tmp4 := subs( Gm(K)^2 = GM^2*(nframes/2), tmp3 );
result := tmp4;
```

```
end;

sum_mean_ratio_q := proc()
local tmp1, tmp2, tmp3, result;
tmp1 := expand( mean_ratio_q(K) );
tmp2 := subs( Gm(K)^4 = GM^4*(qsum*nframes), tmp1 );
tmp3 := subs( Gm(K)^2 = GM^2*(nframes/2), tmp2 );
result := tmp3;
end;

expected_quad_term := (4/nframes^2) * ( sum_mean_ratio_q()
- sum_sq_mean_ratio_sq() + sq_sum_mean_ratio_sq() );

mse := simplify( expected_quad_term - GM^4 - (2*bias*GM^2) );

#
#  Simplify BIAS and MSE
#

bias := simplify( expand( bias ) );
mse := simplify( expand( mse ) );

bias_c0 := simplify( coeff( expand( bias ), i0inv, 0 ) );
bias_c1 := simplify( coeff( expand( bias ), i0inv, 1 ) );
bias_c2 := simplify( coeff( expand( bias ), i0inv, 2 ) );
bias_c3 := simplify( coeff( expand( bias ), i0inv, 3 ) );

mse_c0 := simplify( coeff( mse, i0inv, 0 ) );
mse_c1 := simplify( coeff( mse, i0inv, 1 ) );
mse_c2 := simplify( coeff( mse, i0inv, 2 ) );
mse_c3 := simplify( coeff( mse, i0inv, 3 ) );

bias_c0 := expand( bias_c0 );
bias_c1 := expand( bias_c1 );
bias_c2 := expand( bias_c2 );
bias_c3 := expand( bias_c3 );

mse_c0 := expand( mse_c0 );
mse_c1 := expand( mse_c1 );
mse_c2 := expand( mse_c2 );
mse_c3 := expand( mse_c3 );
```

```
bias_expr := bias_c0 + bias_c1*i0inv + bias_c2*(i0inv^2) + bias_c3*(i0inv^3);
mse_expr := mse_c0 + mse_c1*i0inv + mse_c2*(i0inv^2) + mse_c3*(i0inv^3);
snr_expr := mse_expr - bias_expr^2;

bias_expr := expand( bias_expr );
mse_expr := expand( mse_expr );
snr_expr := expand( snr_expr );

latex( bias_expr );
latex( mse_expr );
latex( snr_expr );

quit;
```

Estimator D2:

```
read( visibility ):

#
# File: ac_mse
# Date: 19 Oct 88
# Author: John D. Gorman
#

ND := proc(k)
result := ld(k) + dd(k);
end;

NS := proc(k)
result := ls(k) + ds(k);
end;

mean_diff := proc()
result := expectddds( expand( (0.5*i0inv)^1 * (ND(K) - NS(K))^1 ), K );
end;

mean_diff_sq := proc()
result := expectddds( expand( (0.5*i0inv)^2 * (ND(K) - NS(K))^2 ), K );
end;

mean_diff_t := proc()
result := expectddds( expand( (0.5*i0inv)^3 * (ND(K) - NS(K))^3 ), K );
end;

mean_diff_q := proc()
result := expectddds( expand( (0.5*i0inv)^4 * (ND(K) - NS(K))^4 ), K );
end;

expectddds := proc(X,K)
local i, result, t1, t2, t3;
option remember;
t1 := X;
for i from 4 by -1 to 1 do
t2 := subs( ( dd(K) )^i = pcm( ld(K), i ), t1 );
t3 := subs( ( ds(K) )^i = pcm( ls(K), i ), t2 );
```

```
t1 := t3;
od
end;

mean_diff_sq_K := expand( mean_diff_sq(K) );

#
#  Calculate BIAS of Discrete-Phase Normalized Estimator
#

sum_mean_diff_sq := proc()
local tmp1, tmp2, tmp3, result;
tmp1 := mean_diff_sq_K;
tmp2 := expand( tmp1 );
tmp3 := subs( Gm(K)^2 = GM^2*(nframes/2), tmp2 );
result := tmp3;
end;

bias := simplify( ((2/nframes)*sum_mean_diff_sq()) - GM^2 );

#
#  Calculate Terms in MSE
#

sq_sum_mean_diff_sq := proc()
local tmp1, result;
tmp1 := sum_mean_diff_sq();
result := expand( tmp1^2 );
end;

sum_sq_mean_diff_sq := proc()
local tmp1, tmp2, tmp3, tmp4, tmp5, result;
tmp1 := mean_diff_sq_K;
tmp2 := expand( tmp1^2 );
tmp3 := subs( Gm(K)^4 = GM^4*(qsum*nframes), tmp2 );
tmp4 := subs( Gm(K)^2 = GM^2*(nframes/2), tmp3 );
result := tmp4;
end;

sum_mean_diff_q := proc()
local tmp1, tmp2, tmp3, tmp4, result;
```

```
tmp1 := mean_diff_q(K);
tmp2 := expand( tmp1 );
tmp3 := subs( Gm(K)^4 = GM^4*(qsum*nframes), tmp2 );
tmp4 := subs( Gm(K)^2 = GM^2*(nframes/2), tmp3 );
result := tmp4;
end;

expected_quad_term := (4/nframes^2) * ( sq_sum_mean_diff_sq() - sum_sq_mean_diff.

mse := simplify( expected_quad_term - GM^4 - (2*bias*GM^2) );

#
#  Simplify BIAS and MSE
#

bias := simplify( expand( bias ) );
mse := simplify( expand( mse ) );

bias_c0 := simplify( coeff( expand( bias ), i0inv, 0 ) );
bias_c1 := simplify( coeff( expand( bias ), i0inv, 1 ) );
bias_c2 := simplify( coeff( expand( bias ), i0inv, 2 ) );
bias_c3 := simplify( coeff( expand( bias ), i0inv, 3 ) );

mse_c0 := simplify( coeff( mse, i0inv, 0 ) );
mse_c1 := simplify( coeff( mse, i0inv, 1 ) );
mse_c2 := simplify( coeff( mse, i0inv, 2 ) );
mse_c3 := simplify( coeff( mse, i0inv, 3 ) );

bias_c0 := expand( bias_c0 );
bias_c1 := expand( bias_c1 );
bias_c2 := expand( bias_c2 );
bias_c3 := expand( bias_c3 );

mse_c0 := expand( mse_c0 );
mse_c1 := expand( mse_c1 );
mse_c2 := expand( mse_c2 );
mse_c3 := expand( mse_c3 );

bias_expr := bias_c0 + bias_c1*i0inv + bias_c2*(i0inv^2) + bias_c3*(i0inv^3);
mse_expr := mse_c0 + mse_c1*i0inv + mse_c2*(i0inv^2) + mse_c3*(i0inv^3);
snr_expr := mse_expr - bias_expr^2;
```

```
bias_expr := expand( bias_expr );
mse_expr := expand( mse_expr );
snr_expr := expand( snr_expr );

latex( bias_expr );
latex( mse_expr );
latex( snr_expr );

quit;
```

Estimator C1:

```
#
# File: Nac_mse
# Date: 22 Aug 88
# Author: John D. Gorman
#

read( visibility ):

G := expand( (A - C)^2 + (B - D)^2 );
Gsq := expand( G^2 );

tildeGA := expand( (tildeA - tildeC)^2 + (tildeB - tildeD)^2 );
tildeGAsq := expand( tildeGA^2 );

tildeGB := subs( tildeA^2 = tildeAsq, tildeGA );
tildeGBq := subs( tildeA^4 = tildeAq, tildeGAsq );
tildeGBt := subs( tildeA^3 = tildeAt, tildeGBq );
tildeGBsq := subs( tildeA^2 = tildeAsq, tildeGBt );

tildeGC := subs( tildeB^2 = tildeBsq, tildeGB );
tildeGCq := subs( tildeB^4 = tildeBq, tildeGBsq );
tildeGCt := subs( tildeB^3 = tildeBt, tildeGCq );
tildeGCsq := subs( tildeB^2 = tildeBsq, tildeGCt );

tildeGD := subs( tildeC^2 = tildeCsq, tildeGC );
tildeGDq := subs( tildeC^4 = tildeCq, tildeGCsq );
tildeGDt := subs( tildeC^3 = tildeCt, tildeGDq );
tildeGDsq := subs( tildeC^2 = tildeCsq, tildeGDt );

tildeG := expand( subs( tildeD^2 = tildeDsq, tildeGD ) );
tildeGq := subs( tildeD^4 = tildeDq, tildeGDsq );
tildeGt := subs( tildeD^3 = tildeDt, tildeGq );
tildeGsq := expand( subs( tildeD^2 = tildeDsq, tildeGt ) );

A := Gm(1);
B := Gm(2);
C := Gm(3);
D := Gm(4);

tildeA := mean_ratio(1):
```

```
tildeAsq := mean_ratio_sq(1):
tildeAt := mean_ratio_t(1):
tildeAq := mean_ratio_q(1):

tildeB := mean_ratio(2):
tildeBsq := mean_ratio_sq(2):
tildeBt := mean_ratio_t(2):
tildeBq := mean_ratio_q(2):

tildeC := mean_ratio(3):
tildeCsq := mean_ratio_sq(3):
tildeCt := mean_ratio_t(3):
tildeCq := mean_ratio_q(3):

tildeD := mean_ratio(4):
tildeDsq := mean_ratio_sq(4):
tildeDt := mean_ratio_t(4):
tildeDq := mean_ratio_q(4):

tildeGexp := expand( tildeG );
tildeGsqexp := expand( tildeGsq );

bias := tildeGexp - G;

sqterm := tildeGsqexp - Gsq;
oterm := expand( 2 * G * bias );
mse := sqterm - oterm;

#
#  Simplify BIAS and MSE
#

bias := simplify( expand( bias ) );
mse := simplify( expand( mse ) );

bias_c0 := simplify( coeff( expand( bias ), i0inv, 0 ) );
bias_c1 := simplify( coeff( expand( bias ), i0inv, 1 ) );
bias_c2 := simplify( coeff( expand( bias ), i0inv, 2 ) );
bias_c3 := simplify( coeff( expand( bias ), i0inv, 3 ) );

mse_c0 := simplify( coeff( mse, i0inv, 0 ) );
```

```
mse_c1 := simplify( coeff( mse, i0inv, 1 ) );
mse_c2 := simplify( coeff( mse, i0inv, 2 ) );
mse_c3 := simplify( coeff( mse, i0inv, 3 ) );

bias_c0 := expand( bias_c0 );
bias_c1 := expand( bias_c1 );
bias_c2 := expand( bias_c2 );
bias_c3 := expand( bias_c3 );

mse_c0 := expand( mse_c0 );
mse_c1 := expand( mse_c1 );
mse_c2 := expand( mse_c2 );
mse_c3 := expand( mse_c3 );

bias_expr := bias_c0 + bias_c1*i0inv + bias_c2*(i0inv^2) + bias_c3*(i0inv^3);
mse_expr := mse_c0 + mse_c1*i0inv + mse_c2*(i0inv^2) + mse_c3*(i0inv^3);
snr_expr := mse_expr - bias_expr^2;

bias_expr := expand( bias_expr );
mse_expr := expand( mse_expr );
snr_expr := expand( snr_expr );

latex( bias_expr );
latex( mse_expr );
latex( snr_expr );

quit,
```

Estimator C2:

```
#
# File: ac_mse
# Date: 22 Aug 88
# Author: John D. Gorman
#

read( visibility ):

ND := proc(k)
result := ld(k) + dd(k);
end;

NS := proc(k)
result := ls(k) + ds(k);
end;

G := expand( (A - C)^2 + (B - D)^2 );
Gsq := expand( G^2 );

tildeG := expand( (tildeA - tildeC)^2 + (tildeB - tildeD)^2 );
tildeGsq := expand( tildeG^2 );

A := Gm(1);
B := Gm(2);
C := Gm(3);
D := Gm(4);

tildeA := 0.5 * i0inv * (ND(1) - NS(1));
tildeB := 0.5 * i0inv * (ND(2) - NS(2));
tildeC := 0.5 * i0inv * (ND(3) - NS(3));
tildeD := 0.5 * i0inv * (ND(4) - NS(4));

expectdddd := proc(X)
local i, j, result, t1, t2, t3;
option remember;
t1 := X;
for i from 4 by -1 to 1 do
for j from 1 by 1 to 4 do
t2 := subs( ( dd(j) )^i = pcm( ld(j), i ), t1 );
```

```
t3 := subs( ( ds(j) )^i = pcm( ls(j), i ), t2 );
t1 := t3;
od
od
end;

tildeGexp := expectddds( expand( tildeG ) );
tildeGsqexp := expectddds( expand( tildeGsq ) );

bias := tildeGexp - G;

sqterm := tildeGsqexp - Gsq;
oterm := expand( 2 * G * bias );
mse := sqterm - oterm;

#
#  Simplify BIAS and MSE
#

bias := simplify( expand( bias ) );
mse := simplify( expand( mse ) );

bias_c0 := simplify( coeff( expand( bias ), i0inv, 0 ) );
bias_c1 := simplify( coeff( expand( bias ), i0inv, 1 ) );
bias_c2 := simplify( coeff( expand( bias ), i0inv, 2 ) );
bias_c3 := simplify( coeff( expand( bias ), i0inv, 3 ) );

mse_c0 := simplify( coeff( mse, i0inv, 0 ) );
mse_c1 := simplify( coeff( mse, i0inv, 1 ) );
mse_c2 := simplify( coeff( mse, i0inv, 2 ) );
mse_c3 := simplify( coeff( mse, i0inv, 3 ) );

bias_c0 := expand( bias_c0 );
bias_c1 := expand( bias_c1 );
bias_c2 := expand( bias_c2 );
bias_c3 := expand( bias_c3 );

mse_c0 := expand( mse_c0 );
mse_c1 := expand( mse_c1 );
mse_c2 := expand( mse_c2 );
mse_c3 := expand( mse_c3 );
```

```
bias_expr := bias_c0 + bias_c1*i0inv + bias_c2*(i0inv^2) + bias_c3*(i0inv^3);
mse_expr := mse_c0 + mse_c1*i0inv + mse_c2*(i0inv^2) + mse_c3*(i0inv^3);
snr_expr := mse_expr - bias_expr^2;

bias_expr := expand( bias_expr );
mse_expr := expand( mse_expr );
snr_expr := expand( snr_expr );

latex( bias_expr );
latex( mse_expr );
latex( snr_expr );

quit;
```

APPENDIX B

Lower Bounds For Parametric Estimation with Constraints

John D. Gorman
Alfred O. Hero

Reprinted from
IEEE TRANSACTIONS ON INFORMATION THEORY
Vol. 36, No. 6, November 1990

Lower Bounds For Parametric Estimation with Constraints

JOHN D. GORMAN, STUDENT MEMBER, IEEE, AND ALFRED O. HERO, MEMBER IEEE

Abstract—A Chapman–Robbins form of the Barankin bound is used to derive a multiparameter Cramér–Rao (CR) type lower bound on estimator error covariance when the parameter $\theta \in \mathbb{R}^n$ is constrained to lie in a subset of the parameter space. A simple form for the constrained CR bound is obtained when the constraint set Θ_c can be expressed as a smooth functional inequality constraint, $\Theta_c = \{\theta : \mathcal{J}_\theta \leq 0\}$. We show that the constrained CR bound is identical to the unconstrained CR bound at the regular points of Θ_c, i.e. where no equality constraints are active. On the other hand, at those points $\theta \in \Theta_c$ where pure equality constraints are active the full-rank Fisher information matrix in the unconstrained CR bound must be replaced by a rank-reduced Fisher information matrix obtained as a projection of the full-rank Fisher matrix onto the tangent hyperplane of the constraint set at θ. A necessary and sufficient condition involving the forms of the constraint and the likelihood function is given for the bound to be achievable, and examples for which the bound is achieved are presented. In addition to providing a useful generalization of the CR bound, our results permit analysis of the gain in achievable mse performance due to the imposition of particular constraints on the parameter space without the need for a global reparameterization. For the purpose of illustration, we apply the constrained bound to problems involving linear constraints and quadratic constraints. Specific examples considered include: linear constraints for Gaussian linear models, object support constraints in image reconstruction, signal subspace constraints in sensor array processing, and average power constraints in spectral estimation and signal extraction.

Index Terms—Constrained estimation, Cramér–Rao bounds, multiple parameter estimation, spectrum estimation.

I. Introduction

THE MULTIPLE PARAMETER Cramér–Rao (CR) lower bound is widely used to investigate the fundamental limits on estimator performance in multidimensional parameter estimation problems, and in single parameter estimation problems involving unknown nuisance parameters. The CR bound on estimator error covariance is computed as the inverse of the Fisher information matrix premultiplied and postmultiplied by the gradient of the mean vector of the estimator. Although elementary derivations, for instance [27, Section 2.4], may not explicitly make the assumption, the CR bound is typically derived under the assumption that the parameter space is an open subset of $\mathbb{R}^n$ [13, Section 1.7]. Frequently, however, the parameter is constrained to lie in a proper non-open subset of the original parameter space. Some examples are: bandwidth, support, and positivity constraints in phase retrieval [5], [9] and tomographic reconstruction [24], [29]; kernel-sieve constraints in probability-density estimation [25]; array geometry constraints in estimation of coupled times-of-arrival across multiple-sensor arrays [28]; and auto-correlation lag constraints in maximum-entropy spectral analysis and image reconstruction [23]. Constraints restrict the allowable parameter variations and hence the local structure of the log-likelihood function over the constrained parameter space may be changed. Specifically, the average curvature of the log-likelihood function, and in particular the Fisher information matrix, may be affected, thereby invalidating the unconstrained CR bound.

We present a multiparameter CR type bound for parametric estimators when the vector parameter θ is constrained to lie in a subset Θ_C of $\mathbb{R}^n$. We refer to this bound as a *constrained* CR bound. The constrained CR bound is derived directly from a version of the Barankin bound: the multiple parameter Chapman–Robbins bound. The tightest such Barankin bound is nonincreasing as Θ_c decreases. Thus, in general, a bound reduction occurs as a result of incorporating constraints. When θ is a nonisolated point in a locally convex region of Θ_C, and the log-likelihood function is smooth, the constrained CR bound depends on Θ_C only through the linear span of a set of basis vectors for the region. When the constraints on the parameter take the form of smooth functional inequality constraints $\mathcal{J}_\theta \leq 0$ more explicit results are obtained. Specifically, let the inequality constraint be decomposed into a finite vector of equality constraints $G_\theta = 0$ and a finite vector of pure inequality constraints $H_\theta \leq 0$ (defined in Section II-C). Then the constrained CR bound is obtained by implementing the classical unconstrained CR bound with a different "constrained" Fisher matrix. The structure of the constrained Fisher matrix depends on whether or not θ is a regular point of Θ_C, where a regular point is a point where no equality constraints are active. As examples, points on the interior and points on the boundary of open regions in Θ_C are

Manuscript received April 26, 1989; revised November 27, 1989. This work was supported in part by the Office of Naval Research under contract N00014-86-C-0587 and in part by the National Cancer Institute of the National Institutes of Health, DHHS, under PHS Grant R01-CA46622-01. This work was presented in part at the Fourth Annual ASSP Workshop on Spectrum Estimation and Modeling, August 1988.

J. D. Gorman is with the Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, MI 48109 and also with Environmental Research Institute of Michigan, Box 8618, Ann Arbor, MI 48107-8618.

A. O. Hero is with the Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, MI 48109.

IEEE Log Number 9038001.

0018-9448/90/1100-1285$01.00

regular points. It is shown that if θ is a regular point then the constrained Fisher matrix is identical to the unconstrained Fisher matrix for that point. Conversely, if θ is not a regular point, the constrained Fisher matrix is the product of the unconstrained Fisher matrix and a θ-dependent, rank-deficient, idempotent matrix whose columns span a hyperplane that is tangent to the constraint set at θ.

The constrained CR bound presented here has the following attributes.

- For range constraints, orthant constraints, positivity constraints, and any other constraint sets Θ_c with no isolated boundaries, the constrained CR bound is identical to the unconstrained CR bound restricted to Θ_c. Hence the incorporation of these types of constraints provides no CR bound reduction.
- For constraints which restrict θ to a lower-dimensional manifold of parameter space, e.g., through active equality constraints of the form $G_\theta = 0$, the unconstrained CR bound is invalid and a reduced-rank Fisher matrix must be used.
- While an equivalent lower-dimensional unconstrained parameter estimation problem can sometimes be specified via a reparameterization of parameter space, such a global reparameterization is not necessary for the computation of the constrained CR bound. Rather, the constrained CR bound only depends on the local properties of the constraint set through its tangent hyperplanes. Since the tangent hyperplanes can typically be computed much more easily than can a global reparameterization of parameter space, the amount of bound reduction due to particular constraints is more easily analyzed.
- Conditions under which the constrained CR bound is achieved are similar to those required for achievement of the unconstrained CR bound. Examples are provided for which the constrained CR bound is achievable.

The following geometrical interpretation is helpful in interpreting the effect of constraints on the CR bound. The Fisher information matrix J_θ being the expected value of the Hessian matrix of the (n-dimensional) log-likelihood surface at θ, can be related to the average curvature of the log-likelihood surface at θ along n different directions in $\mathbb{R}^n$. Thus the unconstrained CR bound is a function of the variation of the likelihood surface over an n-dimensional neighborhood of θ. When the parameter constraint $G_u = 0$, $u \in \mathbb{R}^n$, is introduced, local parameter variations will generally be restricted to lie in a lower dimensional neighborhood. This neighborhood is contained in the linear vector space which is tangent to the constraint set $\{u: G_u = 0\}$ at the point $u = \theta$. As the parameter varies over the lower-dimensional neighborhood, only certain "constrained" trajectories are traversed on the likelihood surface. Thus the average curvature of the surface appears different for the constrained parameter, as compared to the unconstrained parameter for which all local trajectories are allowed. This results in a change in the associated Fisher information matrix and a different CR bound. This constrained CR bound depends on the constraint set only through its tangent space at the point θ.

It is interesting to note that tangent space approximations to subsets of parameter space arise in general asymptotic statistical theory [15], [19] and specific applications have appeared in the statistical literature. For example tangent spaces arise in: the study [7] of the asymptotic distribution of the likelihood ratio for testing composite hypotheses involving smooth boundaries; the study [18] of the asymptotic distribution of a specific estimator arising in a composite detection problem with inequality constraints on the unknown parameter; the study [4] of asymptotic efficiency of estimators in partially parametric models; the study [1] of the asymptotic distribution of maximum likelihood estimators subject to equality constraints. While the study of finite sample CR bounds and the study of asymptotic properties of estimators have points in common, it is important to distinguish between the results of this paper and the aforementioned references. First, our result is a general finite sample CR lower bound on estimator covariance for fully parametric models. Second, the bound is of a simple and explicit form which is useful for studying the impact of particular parameter constraints on estimation error covariance. Third, while the CR bound holds for any estimator whose mean is smooth, the CR bound is not applicable to cases where the estimator has a nondifferentiable mean, such as the estimator considered in [18]. Furthermore, since the bound is a finite sample bound on covariance, methods of large sample theory are not needed for our derivation permitting a more elementary, and therefore more accessible, presentation.

To illustrate the utility of the constrained CR bound, we investigate the effect of constraints on the achievable estimator error for several representative problems in signal processing. First we consider the problem of estimation of parameters subject to linear constraints in the general linear Gaussian model. For this problem the tangent hyperplanes of the constraint set are functionally independent of the parameter θ, and hence the constrained CR lower bound can be achieved by projecting the unconstrained minimum variance unbiased (MVU) estimator onto the tangent hyperplane. The amount of bound reduction depends on the rank of the projection of the covariance matrix of the unconstrained MVU onto the linear constraint subspace.

Second, we consider the problem of image reconstruction subject to support constraints on the image. The constrained CR bound is equal to the pseudo-inverse of a constrained Fisher matrix, obtained by zeroing out the rows and columns of the unconstrained Fisher information matrix which are associated with estimator errors outside of the region of support. It is significant that this is not generally the same as zeroing out rows and columns

of the unconstrained CR bound, unless the image pixels are statistically independent. This establishes that, if an efficient estimator of the unconstrained image exists, zeroing the unconstrained efficient estimator outside of the support region does not, in general, provide an efficient constrained estimator.

Third, power spectral density (PSD) estimation subject to average power constraints over disjoint frequency intervals, called frequency bands, is considered. For the case where the unconstrained Fisher information matrix is diagonal, corresponding to large observation time, it is shown that the constrained Fisher matrix is block diagonal. This means that average power constraints effectively couple the PSD estimation errors over a particular frequency band, but do not couple errors across different frequency bands. Within a particular frequency band where average power constraints are active, our results indicate that bound reduction is greatest over frequency bands where there are highly resolved spectral peaks, while there is virtually no reduction over bands where the true spectrum is smooth. This suggests that average power constraints make peaks easier to estimate but have little impact on the estimation of the rest of the spectrum.

Fourth, the estimation of the eigenvalues of a structured covariance matrix subject to signal subspace constraints is considered. We put this problem in the context of estimating the eigenvalues and eigenvectors of the array covariance matrix when it is known *a priori* that p of the eigenvalues, the "signal dependent eigenvalues," are larger than the remaining eigenvalues, the "noise eigenvalues," and that these latter eigenvalues are identical. When the unconstrained Fisher matrix is block diagonal, the constrained CR bound can be achieved by averaging the noise eigenvalues of an efficient unconstrained estimator, if one exists.

Finally, we consider the problem of estimation of a deterministic time varying signal, and its Fourier transform, subject to average power constraints applied to its spectrum (squared Fourier magnitudes). Unlike the PSD estimation problem previously mentioned, here the constraints on the parameters (the signal) are nonlinear. Nonetheless, it is shown that if the unconstrained Fisher information is an identity matrix, e.g., corresponding to observation of the signal in additive-white-Gaussian noise, the structure of the constrained Fisher matrix is identical to the structure found in the PSD estimation problem, with the signal spectrum taking the place of the PSD.

An outline of the paper is as follows. Section II is divided into several subsections. In Section II-A a Barankin lower bound on the estimator covariance is given for general constrained parameters. In Section II-B the constrained CR bound is derived from this Barankin bound for locally convex regions of the constrained parameter space Θ_C. In Section II-C the constrained CR bound of Section II-B is extended to the case of smooth nonlinear functional inequality constraints. In Section III, examples of the implementation of the constrained CR bound are presented.

II. Lower Bounds on the Error Covariance

Throughout the paper the notation $\boldsymbol{\theta}$ and $[\theta_i]_{i=1,\cdots,n}$ will denote a column vector, $[\theta_1, \cdots, \theta_n]^T$, of unknown parameters contained in the unconstrained parameter space $\Theta = \mathbb{R}^n$. For a particular value of the vector $\boldsymbol{\theta}$ we specify a probability distribution $P_{\boldsymbol{\theta}}$ governing the observations X, taking values x in a sample space Ω. The collection of probability spaces $\mathcal{E} = \{(\Omega, \mathcal{F}, P_{\boldsymbol{\theta}})\}_{\boldsymbol{\theta} \in \Theta}$ defines a $\boldsymbol{\theta}$-indexed set of possible models for X, and is called a *statistical experiment* over Θ. If it is known that $\boldsymbol{\theta}$ is restricted to a subset of Θ, called the constrained parameter space Θ_C, the relevant statistical experiment becomes the reduced set of models $\mathcal{E}_C = \{(\Omega, \mathcal{F}, P_{\boldsymbol{\theta}})\}_{\boldsymbol{\theta} \in \Theta_C}$. In this context, the constrained parameter estimation problem can be stated as follows: given a statistical experiment $\mathcal{E}_C$, a random variable X is observed which has distribution $P_{\boldsymbol{\theta}}$; the objective is to specify an estimator $\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}(X) \in \Theta$ for the parameter vector $\boldsymbol{\theta}$. Define the vector mean $\boldsymbol{m}_{\boldsymbol{\theta}} \stackrel{\text{def}}{=} E_{\boldsymbol{\theta}}\{\hat{\boldsymbol{\theta}}\}$ of $\hat{\boldsymbol{\theta}}$, where $E_{\boldsymbol{\theta}}$ denotes expectation with respect to the distribution $P_{\boldsymbol{\theta}}$. The objective of this paper is to investigate the impact of parameter constraints on bounds for the minimum estimation error, where error is measured by the covariance matrix

$$\Sigma_{\boldsymbol{\theta}} \stackrel{\text{def}}{=} E_{\boldsymbol{\theta}}\left\{(\hat{\boldsymbol{\theta}} - \boldsymbol{m}_{\boldsymbol{\theta}})(\hat{\boldsymbol{\theta}} - \boldsymbol{m}_{\boldsymbol{\theta}})^T\right\}. \tag{1}$$

We say that a matrix B is a lower bound on a matrix A if $A \geq B$ in the sense that $A - B$ is nonnegative definite.

A. A Multiple Parameter Barankin Bound

We first present a Chapman–Robbins version of the multiple parameter Barankin lower bound on the covariance matrix $\Sigma_{\boldsymbol{\theta}}$ for the case where $\boldsymbol{\theta} \in \Theta_C$. Unlike the CR bound, the Barankin bound requires no regularity conditions on the distribution $P_{\boldsymbol{\theta}}$. To achieve a unified treatment of the cases of continuous and discrete random variables X, we let $P_{\boldsymbol{\theta}}$ have a density function $f_{\boldsymbol{\theta}} = f_{\boldsymbol{\theta}}(x)$ with respect to some reference measure μ. $P_{\boldsymbol{\theta}}(A) = \int_A f_{\boldsymbol{\theta}}\, d\mu$, where $P_{\boldsymbol{\theta}}(A)$ is the probability that $X \in A$, $A \in \mathcal{F}$. For a continuous sample space Ω the previous integral can be interpreted as the standard (Lebesgue) integral over A, while for discrete Ω, μ is the counting measure and the integral can be interpreted as a sum over elements $x \in A$.

For arbitrary vectors $\boldsymbol{\nu}_1, \cdots, \boldsymbol{\nu}_k \in \mathbb{R}^n$ and scalars $\Delta_1, \cdots, \Delta_k \in \mathbb{R}$, define the scalar and vector finite differences, $\delta_i f_{\boldsymbol{\theta}}$ and $\delta_i \boldsymbol{m}_{\boldsymbol{\theta}}$, of the density function and of the mean vector for $\hat{\boldsymbol{\theta}}$, respectively, which are produced by a change in the underlying parameter from the point $\boldsymbol{\theta}$ to the point $\boldsymbol{\theta} + \Delta_i \boldsymbol{\nu}_i$:

$$\delta_i f_{\boldsymbol{\theta}} \stackrel{\text{def}}{=} \frac{f_{\boldsymbol{\theta} + \Delta_i \boldsymbol{\nu}_i} - f_{\boldsymbol{\theta}}}{\Delta_i}, \tag{2}$$

$$\delta_i \boldsymbol{m}_{\boldsymbol{\theta}} \stackrel{\text{def}}{=} \frac{\boldsymbol{m}_{\boldsymbol{\theta} + \Delta_i \boldsymbol{\nu}_i} - \boldsymbol{m}_{\boldsymbol{\theta}}}{\Delta_i}. \tag{3}$$

These finite differences are the variations in $f_{\boldsymbol{\theta}}$ and $\boldsymbol{m}_{\boldsymbol{\theta}}$

along the directions of the vectors $\boldsymbol{v}_1, \cdots, \boldsymbol{v}_k$; a set of vectors which are henceforth referred to as *direction vectors*. Define the row vector of k finite differences,

$$\delta f_{\boldsymbol{\theta}} \stackrel{\text{def}}{=} [\delta_1 f_{\boldsymbol{\theta}}, \cdots, \delta_k f_{\boldsymbol{\theta}}], \tag{4}$$

and the $n \times k$ matrix of finite differences

$$\delta \boldsymbol{m}_{\boldsymbol{\theta}} \stackrel{\text{def}}{=} [\delta_1 \boldsymbol{m}_{\boldsymbol{\theta}}, \cdots, \delta_k \boldsymbol{m}_{\boldsymbol{\theta}}]. \tag{5}$$

With these definitions we have the following multiple parameter Chapman–Robbins version of the Barankin bound [6], [17] when $\boldsymbol{\theta}$ is constrained to lie in the set Θ_C.

Proposition 1: Let the $k+1$ vectors $\boldsymbol{\theta}, \boldsymbol{\theta} + \Delta_1 \boldsymbol{v}_1, \cdots, \boldsymbol{\theta} + \Delta_k \boldsymbol{v}_k$ be arbitrary points contained in the constrained parameter set $\Theta_C \subset \mathbb{R}^n$. Then for any estimator $\hat{\boldsymbol{\theta}}$ having mean $\boldsymbol{m}_{\boldsymbol{\theta}}$, the estimator error covariance matrix $\Sigma_{\boldsymbol{\theta}}$ satisfies the matrix inequality

$$\Sigma_{\boldsymbol{\theta}} \geq B_k \tag{6}$$

where

$$B_k = [\delta \boldsymbol{m}_{\boldsymbol{\theta}}] \cdot \left\{ E_{\boldsymbol{\theta}} \left[\frac{\delta f_{\boldsymbol{\theta}}}{f_{\boldsymbol{\theta}}} \right]^T \left[\frac{\delta f_{\boldsymbol{\theta}}}{f_{\boldsymbol{\theta}}} \right] \right\}^{\cdot} \cdot [\delta \boldsymbol{m}_{\boldsymbol{\theta}}]^T, \tag{7}$$

and the plus sign denotes pseudo-inverse. Equality holds in (6) *if and only if there exists a nonrandom* $n \times k$ *matrix* Γ such that the estimator $\hat{\boldsymbol{\theta}}$ satisfies

$$\hat{\boldsymbol{\theta}} - \boldsymbol{m}_{\boldsymbol{\theta}} = \Gamma \left[\frac{\delta f_{\boldsymbol{\theta}}}{f_{\boldsymbol{\theta}}} \right]^T \quad \text{(w.p.1)}. \tag{8}$$

In Proposition 1, the pseudo-inverse of a matrix A is defined as the unique matrix $A^{\cdot}$ that satisfies the Moore–Penrose conditions [2, Ch. 3], [21, Section 1.b5]:

$$\begin{aligned} &1)\ AA^{\cdot} \text{ and } A^{\cdot}A \text{ are symmetric,} \\ &2)\ AA^{\cdot}A = A, \\ &3)\ A^{\cdot}AA^{\cdot} = A^{\cdot}. \end{aligned} \tag{9}$$

The conditions 1)–3) are a statement of the fact that $AA^{\cdot}$ *and* $A^{\cdot}A$ *are projection* operators onto the range of A and $A^{\cdot}$, respectively. Pseudo-inverses always exist, are continuous under certain conditions [26], and if A is invertible $A^{\cdot} = A^{-1}$.

Before proving Proposition 1, we make the following observations. Since only a pseudo-inverse is required for the bound B_k of Proposition 1, the covariance matrix, $E_{\boldsymbol{\theta}}[\delta f_{\boldsymbol{\theta}} / f_{\boldsymbol{\theta}}]^T [\delta f_{\boldsymbol{\theta}} / f_{\boldsymbol{\theta}}]$, of the finite difference vector does not have to be invertible. This general form is necessary for the present application since parameter constraints can reduce the rank of the covariance matrix. In view of the definition (4) of the finite difference vector $\delta f_{\boldsymbol{\theta}}$ the bound (6) is a measure of the variation of the probability density $f_{\boldsymbol{\theta}}$ relative to the set of "test" points $\boldsymbol{\theta} + \Delta_1 \boldsymbol{v}_1, \cdots, \boldsymbol{\theta} + \Delta_k \boldsymbol{v}_k$, which are arbitrarily specified in the constrained parameter space Θ_C. On the other hand, since $\Theta_C \subset \Theta$, it is obvious that

$$\max_{\Theta} B_k \geq \max_{\Theta_C} B_k,$$

where each maximization is performed over the set of admissible test points in the parameter space. Hence constraining the parameter space can only reduce the (greatest) lower bound of the form (6). Thus it is clear that some bound reduction can occur due to incorporation of parameter constraints. Due to the difficulty in finding the best test vectors for (6), however, the amount of bound reduction is difficult to quantify in general. In the next section we will derive a constrained CR bound as a limiting form of the bound (6) for which the impact of constraints will be much easier to evaluate.

The proof of Proposition 1 depends on the following generalized version of the Cauchy–Schwarz inequality.

Lemma 1: Let $\boldsymbol{U} \in \mathbb{R}^n$ and $\boldsymbol{V} \in \mathbb{R}^k$ be random column vectors. Then

$$E_{\boldsymbol{\theta}}\{\boldsymbol{U}\boldsymbol{U}^T\} \geq E_{\boldsymbol{\theta}}\{\boldsymbol{U}\boldsymbol{V}^T\} \left[E_{\boldsymbol{\theta}}\{\boldsymbol{V}\boldsymbol{V}^T\} \right]^{\cdot} E_{\boldsymbol{\theta}}\{\boldsymbol{V}\boldsymbol{U}^T\}, \tag{10}$$

where the plus sign denotes pseudo-inverse. Moreover, equality holds if and only if there is an $n \times k$ nonrandom matrix Γ such that $\boldsymbol{U} = \Gamma \boldsymbol{V}$ w.p.1.

Note that if the $k \times k$ matrix $E_{\boldsymbol{\theta}}\{\boldsymbol{V}\boldsymbol{V}^T\}$ is nonsingular, the matrix inequality (10) is the standard Cauchy–Schwarz inequality for random vectors.

Proof of Lemma 1: Define the $\mathbb{R}^{n+k}$ vector $\boldsymbol{Z} = [\boldsymbol{U}^T \boldsymbol{V}^T]^T$. Then $E_{\boldsymbol{\theta}}\{\boldsymbol{Z}\boldsymbol{Z}^T\} \geq 0$ implies the matrix inequality

$$E_{\boldsymbol{\theta}}\{\boldsymbol{Z}\boldsymbol{Z}^T\} = E_{\boldsymbol{\theta}} \left\{ \begin{bmatrix} \boldsymbol{U}\boldsymbol{U}^T & \boldsymbol{U}\boldsymbol{V}^T \\ \boldsymbol{V}\boldsymbol{U}^T & \boldsymbol{V}\boldsymbol{V}^T \end{bmatrix} \right\} \geq 0.$$

Let D be the $n \times (n+k)$ partitioned matrix

$$D = \left[I \mid -E_{\boldsymbol{\theta}}\{\boldsymbol{U}\boldsymbol{V}^T\} \left[E_{\boldsymbol{\theta}}\{\boldsymbol{V}\boldsymbol{V}^T\} \right]^{\cdot} \right],$$

where I is the $n \times n$ identity. Since $E_{\boldsymbol{\theta}}\{\boldsymbol{Z}\boldsymbol{Z}^T\}$ is symmetric and nonnegative-definite, it has a nonnegative square root: $E_{\boldsymbol{\theta}}\{\boldsymbol{Z}\boldsymbol{Z}^T\} = E_{\boldsymbol{\theta}}^{1/2}\{\boldsymbol{Z}\boldsymbol{Z}^T\} E_{\boldsymbol{\theta}}^{T/2}\{\boldsymbol{Z}\boldsymbol{Z}^T\}$. Thus, $DE_{\boldsymbol{\theta}}\{\boldsymbol{Z}\boldsymbol{Z}^T\} D^T = [DE_{\boldsymbol{\theta}}^{1/2}\{\boldsymbol{Z}\boldsymbol{Z}^T\}][DE_{\boldsymbol{\theta}}^{1/2}\{\boldsymbol{Z}\boldsymbol{Z}^T\}]^T \geq 0$, and use of property 3) of (9) results in

$$E_{\boldsymbol{\theta}}\{\boldsymbol{U}\boldsymbol{U}^T\} - E_{\boldsymbol{\theta}}\{\boldsymbol{U}\boldsymbol{V}^T\} \left[E_{\boldsymbol{\theta}}\{\boldsymbol{V}\boldsymbol{V}^T\} \right]^{\cdot} E_{\boldsymbol{\theta}}\{\boldsymbol{V}\boldsymbol{U}^T\} \geq 0.$$

This equation can be reexpressed as $E_{\boldsymbol{\theta}}\{(\boldsymbol{U} - \Gamma \boldsymbol{V})(\boldsymbol{U} - \Gamma \boldsymbol{V})^T\} \geq 0$, where $\Gamma \stackrel{\text{def}}{=} E_{\boldsymbol{\theta}}\{\boldsymbol{U}\boldsymbol{V}^T\}[E_{\boldsymbol{\theta}}\{\boldsymbol{V}\boldsymbol{V}^T\}]^{\cdot}$. Equality holds if and only if the eigenvalues, λ_i, of the matrix $E_{\boldsymbol{\theta}}\{(\boldsymbol{U} - \Gamma \boldsymbol{V})(\boldsymbol{U} - \Gamma \boldsymbol{V})^T\}$ are zero. Furthermore, the nonnegative definiteness of this matrix implies that $\lambda_1 = \cdots = \lambda_n = 0$ if and only if $0 = \Sigma \lambda_i = \text{tr}[E_{\boldsymbol{\theta}}\{(\boldsymbol{U} - \Gamma \boldsymbol{V})(\boldsymbol{U} - \Gamma \boldsymbol{V})^T\}] = E_{\boldsymbol{\theta}}\{(\boldsymbol{U} - \Gamma \boldsymbol{V})^T (\boldsymbol{U} - \Gamma \boldsymbol{V})\}$. Hence, equality holds in (10) if and only if $\boldsymbol{U} = \Gamma \boldsymbol{V}$ w.p.1. □

Using the previous Lemma, Proposition 1 is proven next.

Proof of Proposition 1: Define the n-vector $\boldsymbol{U}$ and the k-vector $\boldsymbol{V}$

$$\boldsymbol{U} \stackrel{\text{def}}{=} \hat{\boldsymbol{\theta}} - \boldsymbol{m}_{\boldsymbol{\theta}},$$

$$\boldsymbol{V} \stackrel{\text{def}}{=} \left[\frac{\delta f_{\boldsymbol{\theta}}}{f_{\boldsymbol{\theta}}} \right]^T,$$

where $\boldsymbol{m}_{\boldsymbol{\theta}}$ is the mean vector of $\hat{\boldsymbol{\theta}}$ and $\delta f_{\boldsymbol{\theta}}$ is the vector of

finite differences defined in (4). With these definitions, application of Lemma 1 gives a lower bound involving the pseudo-inverse of the $k \times k$ matrix $E_\theta\{VV^T\}$ and the $k \times n$ and $n \times k$ matrices $E_\theta\{VU^T\}$ and $E_\theta\{UV^T\}$, respectively. If it can be shown that $E_\theta\{UV^T\} = \delta m_\theta$ Proposition 1 would be established. Consider the jth column of $E_\theta\{UV^T\}$ and recall the definition (4) of δf_θ.

$$\begin{aligned}\left[E_\theta\{UV^T\}\right]_{\cdot,j} &= E_\theta\left\{\left[\hat{\theta} - m_\theta\right]\frac{\delta_j f_\theta}{f_\theta}\right\} \\ &= E_\theta\left\{\left[\hat{\theta} - m_\theta\right]\frac{f_{\theta - \Delta_j v_j} - f_\theta}{\Delta_j f_\theta}\right\} \\ &= \frac{E_{\theta - \Delta_j v_j}\{\hat{\theta} - m_\theta\} - E_\theta\{\hat{\theta} - m_\theta\}}{\Delta_j} \\ &= \frac{m_{\theta - \Delta_j v_j} - m_\theta}{\Delta_j} \\ &= \delta_j m_\theta. \end{aligned}$$ □

B. The Constrained CR bound

We first obtain a constrained CR bound for locally convex Θ_C directly from the bound (6). We then show that the same bound holds for points $\theta \in \Theta_C$ at which Θ_C can be approximated by a union of locally convex sets. These results are then used in Section II-C to construct CR bounds when Θ_C is specified by continuously differentiable functional constraints.

Let θ and the k linearly independent test vectors $\theta + \Delta_1 v_1, \cdots, \theta + \Delta_k v_k$ be contained in the reduced parameter space Θ_C for all sufficiently small Δ_i, $i = 1, \cdots, k$. Such test vectors can always be found for points θ that are in locally convex regions of Θ_C with dimension at least k. Assuming the exchange of limiting and expectation operations is valid, the limit of the bound B_b, (6) of Proposition 1, as $\Delta_i \to 0$, $i = 1, \cdots, k$, gives a bound which depends only on the directional derivatives, $\lim_{\Delta_i \to 0} \delta_i f_\theta$ and $\lim_{\Delta_i \to 0} \delta_i m_\theta$, of f_θ and the mean vector, m_θ, along the directions of the vectors v_i, $i = 1, \cdots, k$, at the point θ. Specifically, by the chain rule we would have: $\lim_{\Delta_1, \cdots, \Delta_k \to 0} \delta f_\theta = \nabla f_\theta K$ and $\lim_{\Delta_1, \cdots, \Delta_k \to 0} \delta m_\theta = \nabla m_\theta K$, where $K = [v_1, \cdots, v_k]$ is the $n \times k$ matrix of direction vectors; ∇f_θ is the $1 \times n$ (row-vector) gradient of f_θ; and ∇m_θ is the $n \times n$ matrix whose rows are the gradient vectors associated with each scalar component of m_θ. If we could substitute the above limiting expressions into the right-hand side of (6) we would obtain

$$\Sigma_\theta \geq [\nabla m_\theta] K \left[K^T J_\theta K\right]^{-} K^T [\nabla m_\theta]^T, \qquad (11)$$

where

$$\begin{aligned} J_\theta &= E_\theta\left\{\left[\frac{\nabla f_\theta}{f_\theta}\right]^T\left[\frac{\nabla f_\theta}{f_\theta}\right]\right\} \\ &= E_\theta\{[\nabla \ln f_\theta]^T[\nabla \ln f_\theta]\} \end{aligned} \qquad (12)$$

is the $n \times n$ Fisher information matrix. Under appropriate regularity conditions [13, Lemma 8.1], [27, Section 2.4], the Fisher matrix is equivalent to

$$J_\theta = -E_\theta\{\nabla^2 \ln f_\theta\}. \qquad (13)$$

where $\nabla^2 \ln f_\theta$ is the Hessian matrix of partial derivatives of $\ln f_\theta$ with respect to elements of θ. This motivates the following lemma.

Lemma 2: Let the vector θ be in the constrained parameter space $\Theta_C \subset \mathbb{R}^n$, and let $\{v_i\}_{i=1}^k$ be k linearly independent vectors such that $\theta - \Delta_i v_i \in \Theta_C$ for all sufficiently small $\Delta_i > 0$, $i = 1, \cdots, k$. Then for any estimator $\hat{\theta}$ having mean m_θ, the estimator error covariance matrix Σ_θ satisfies the matrix inequality

$$\Sigma_\theta \geq B_c \overset{\text{def}}{=} \limsup_{\Delta_1, \cdots, \Delta_k \to 0} B_b, \qquad (14)$$

where B_b is the bound (6) of Proposition 1. If in addition the following four regularity conditions hold:

- $\hat{\theta}$ has finite variance: $\operatorname{var}\{\hat{\theta}_i\} < \infty$; (15)
- f_θ has continuous partial derivatives; (16)
- $E_\theta\left\{\left|\frac{\partial \ln f_\theta}{\partial \theta_i}\frac{\partial \ln f_\theta}{\partial \theta_j}\right|\right\} < \infty$; (17)
- the matrix $E_\theta\{[\nabla \ln f_\theta]^T[\nabla \ln f_\theta]\}$ is positive definite: (18)

then

$$B_c = [\nabla m_\theta] A \left[A^T J_\theta A\right]^{-} A^T [\nabla m_\theta]^T, \qquad (19)$$

where J_θ is the positive definite $n \times n$ Fisher matrix (12), and A is any $n \times n$ matrix whose column space equals $\operatorname{span}\{v_1, \cdots, v_k\}$. Under these regularity conditions, equality is achieved in the lower bound (14) if and only if there exists a non-random $n \times n$ matrix Γ such that:

$$\hat{\theta} - m_\theta = \Gamma A^T [\nabla \ln f_\theta]^T \quad \text{(w.p.1)}. \qquad (20)$$

If such an estimator $\hat{\theta}$ exists, this estimator is called an efficient constrained estimator.

Proof of Lemma 2: By assumption, $\theta + \Delta_1 v_1, \cdots, \theta + \Delta_k v_k$ are contained in Θ_C for all Δ_i sufficiently small, $i = 1, \cdots, k$, and the bound (14) follows directly from the Barankin bound of Proposition 1.

The regularity conditions (15)–(17) ensure that the Fisher matrix J_θ (12) exists and has bounded elements [13, Section 1.7], and condition (18) says that J_θ is positive definite.

We first derive the limits as $\Delta_1, \cdots, \Delta_k \to 0$ of the matrices $E_\theta\left\{\left[\frac{\delta f_\theta}{f_\theta}\right]^T\left[\frac{\delta f_\theta}{f_\theta}\right]\right\}$ and δm_θ under the stated regularity conditions of Lemma 2. Define $\Delta \overset{\text{def}}{=} \max_i |\Delta_i|$. Let K be the $n \times k$ matrix with columns $v_1, \cdots, v_k$. By condition (16) and the chain rule

$$\begin{aligned}\lim_{\Delta_1, \cdots, \Delta_k \to 0} \frac{\delta f_\theta}{f_\theta} &= \frac{1}{f_\theta}\nabla f_\theta K \\ &= \nabla \ln f_\theta K. \end{aligned}$$

From this, and the stated continuity of $\nabla \ln f_\theta$, condition (16), the ijth element of $\left[\frac{\delta f_\theta}{f_\theta}\right]^T\left[\frac{\delta f_\theta}{f_\theta}\right]$ is dominated by $\sum_{l,m=1}^{n} K_{li}\left[\frac{\partial \ln f_\theta}{\partial \theta_l}\frac{\partial \ln f_\theta}{\partial \theta_m}\right]K_{mj} + O(\Delta)$, which has finite expectation by condition (17). Hence, by dominated convergence [3, Theorem 16.4], we have the finite limit

$$\begin{aligned}\lim_{\Delta_1,\cdots,\Delta_k \to 0} & E_\theta\left[\frac{\delta f_\theta}{f_\theta}\right]^T\left[\frac{\delta f_\theta}{f_\theta}\right] \\ &= K^T E_\theta\{[\nabla \ln f_\theta]^T[\nabla \ln f_\theta]\}K \\ &= K^T J_\theta K. \end{aligned} \tag{21}$$

Next consider the $n \times k$ matrix

$$\begin{aligned}\delta \boldsymbol{m}_\theta &= \left[\frac{\boldsymbol{m}_{\theta+\Delta_j \boldsymbol{\nu}_j} - \boldsymbol{m}_\theta}{\Delta_j}\right]_{j=1,\cdots,k} \\ &= \left[\frac{E_{\theta+\Delta_j \boldsymbol{\nu}_j}\{\hat{\boldsymbol{\theta}}\} - E_\theta\{\hat{\boldsymbol{\theta}}\}}{\Delta_j}\right]_{j=1,\cdots,k} \\ &= \left[E_\theta\left\{\hat{\boldsymbol{\theta}}\frac{f_{\theta+\Delta_j \boldsymbol{\nu}_j} - f_\theta}{\Delta_j}\frac{1}{f_\theta}\right\}\right]_{j=1,\cdots,k} \\ &= \left[E_\theta\left\{\hat{\boldsymbol{\theta}}\frac{\delta_j f_\theta}{f_\theta}\right\}\right]_{j=1,\cdots,k} = E_\theta\left\{\hat{\boldsymbol{\theta}}\frac{\delta f_\theta}{f_\theta}\right\}, \\ &= E_\theta\left\{(\hat{\boldsymbol{\theta}} - \boldsymbol{m}_\theta)\frac{\delta f_\theta}{f_\theta}\right\}, \end{aligned}$$

where the last equality results from the identity $E_\theta\left[\frac{\delta f_\theta}{f_\theta}\right] = 0$. Now from condition (16) the elements of the $n \times k$ matrix $(\hat{\boldsymbol{\theta}} - \boldsymbol{m}_\theta)\delta f_\theta / f_\theta$ are equal to the elements of $(\hat{\boldsymbol{\theta}} - \boldsymbol{m}_\theta)\nabla \ln f_\theta K$ to order $O(\Delta)$. The Schwarz inequality and the regularity conditions (15) and (17) can be used to establish that the elements of the latter matrix have finite absolute expectation

$$\begin{aligned}|E_\theta\{(\hat{\boldsymbol{\theta}} - \boldsymbol{m}_\theta)\nabla \ln f_\theta\}_{i,j}| &= \left|E_\theta\left\{(\hat{\theta}_i - [\boldsymbol{m}_\theta]_i)\frac{\partial \ln f_\theta}{\partial \theta_j}\right\}\right| \\ &\le \operatorname{var}^{1/2}\{\hat{\theta}_i\}E_\theta^{1/2}\left\{\left|\frac{\partial \ln f_\theta}{\partial \theta_j}\right|^2\right\} \\ &< \infty. \end{aligned}$$

Hence, by dominated convergence, the limit

$$\lim_{\Delta_1,\cdots,\Delta_k \to 0} \delta \boldsymbol{m}_\theta$$

exists and is equal to the finite matrix

$$\begin{aligned}\lim_{\Delta_1,\cdots,\Delta_k \to 0} \delta \boldsymbol{m}_\theta &= E_\theta\left\{\hat{\boldsymbol{\theta}}\frac{\nabla f_\theta}{f_\theta}\right\}K \\ &= \nabla E_\theta\{\hat{\boldsymbol{\theta}}\}K \\ &= \nabla \boldsymbol{m}_\theta K. \end{aligned} \tag{22}$$

Since the columns, $\{\boldsymbol{\nu}_j\}_{j=1}^k$, of K are linearly independent, by condition (18) $K^T J_\theta K$ is a full rank invertible matrix and $[K^T J_\theta K]^- = [K^T J_\theta K]^{-1}$. Since the matrix $K^T J_\theta K$ is symmetric and positive definite the eigenvalues of the perturbed matrix $K^T J_\theta K + E$ are positive for a sufficiently small matrix perturbation E [12, Corollary 6.3.4]. This implies that the inverse of $K^T J_\theta K$ is continuous in perturbations of its elements

$$\begin{aligned}\left(E_\theta\left[\frac{\delta f_\theta}{f_\theta}\right]^T\left[\frac{\delta f_\theta}{f_\theta}\right]\right)^- &= (K^T J_\theta K + \underline{\underline{O}}(\Delta))^- \\ &= [K^T J_\theta K]^- + \underline{\underline{o}}(1), \end{aligned} \tag{23}$$

where $\underline{\underline{O}}(\Delta)$ and $\underline{\underline{o}}(1)$ are matrices whose elements are of order $O(\Delta)$ and of order $o(1)$, respectively. In view of (21) we therefore have

$$\begin{aligned}\limsup B_\theta = & \lim_{\Delta_1,\cdots,\Delta_k \to 0}[\delta \boldsymbol{m}_\theta] \\ & \cdot \lim_{\Delta_1,\cdots,\Delta_k \to 0}\left(E_\theta\left[\frac{\delta f_\theta}{f_\theta}\right]^T\left[\frac{\delta f_\theta}{f_\theta}\right]\right)^- \\ & \cdot \lim_{\Delta_1,\cdots,\Delta_k \to 0}[\delta \boldsymbol{m}_\theta]^T \\ = & [\nabla \boldsymbol{m}_\theta]K[K^T J_\theta K]^- K^T[\nabla \boldsymbol{m}_\theta]^T. \end{aligned} \tag{24}$$

It remains to show that the bound (24) depends only on the range space of $K = [\boldsymbol{\nu}_1, \cdots, \boldsymbol{\nu}_k]$. Let A be an $n \times n$ matrix whose column span is identical to the span of $\boldsymbol{\nu}_1, \cdots, \boldsymbol{\nu}_k$. Since the column spaces of A and K are identical, there exists an invertible $n \times n$ matrix T such that

$$[K \quad O_1]T = A,$$

where O_1 is an $n \times (n-k)$ matrix of zeros. Let O_2 and O_3 be $(n-k) \times (n-k)$ and $k \times (n-k)$ matrices of zeros, respectively. Then,

$$\begin{aligned} A[A^T J_\theta A]^- A^T &= [K \quad O_1]T\left[T^T\begin{bmatrix}K^T \\ O_1^T\end{bmatrix}J_\theta[K \quad O_1]T\right]^- T^T\begin{bmatrix}K^T \\ O_1^T\end{bmatrix} \\ &= [K \quad O_1]\left[\begin{bmatrix}K^T \\ O_1^T\end{bmatrix}J_\theta[K \quad O_1]\right]^-\begin{bmatrix}K^T \\ O_1^T\end{bmatrix} \\ &= [K \quad O_1]\begin{bmatrix}K^T J_\theta K & O_3 \\ O_3^T & O_2\end{bmatrix}^-\begin{bmatrix}K^T \\ O_1^T\end{bmatrix} \\ &= [K \quad O_1]\begin{bmatrix}[K^T J_\theta K]^- & O_3 \\ O_3^T & O_2\end{bmatrix}\begin{bmatrix}K^T \\ O_1^T\end{bmatrix} \\ &= K[K^T J_\theta K]^- K^T, \end{aligned}$$

where the second equality follows from (65) of Lemma 5 in the Appendix.

The condition for equality in the bound (14), under the regularity conditions (15)–(17), can be obtained by mak-

ing the identifications $U = (\hat{\theta} - \theta)$, $V = K^T[\nabla \ln f_\theta]^T$ in Lemma 1, verifying that the right side of the resultant bound (10) is identical to the right side of the bound (14) and invoking the necessary and sufficient condition for equality in (10): $U = TV$ for some $k \times n$ matrix T. This gives:

$$\hat{\theta} - \boldsymbol{m}_\theta = TK^T[\nabla \ln f_\theta]^T \quad \text{(w.p.1)}.$$

Since A has the identical column span as K, the above is equivalent to condition (20). □

The constrained CR bound (19) of Lemma 2 is in a general form that is applicable to nonisolated points θ in locally convex regions of the parameter space Θ_C. It is significant that, unlike the Barankin bound of Proposition 1, the constrained CR bound (19) only depends on the test points through the span of the set $\{\boldsymbol{v}_1, \cdots, \boldsymbol{v}_k\}$. In particular, when Θ_C is only p-dimensional in the neighborhood of θ, and $p < n$, all p-dimensional sets of test points are equivalent in the sense that the limit (19) of the Barankin bound is the same.

The construction of Lemma 2 requires that Θ_C be locally convex or star-shaped in the neighborhood of θ. Lemma 2 can be extended to include nonisolated points in regions of Θ_C that have the property that local neighborhoods can be approximated to order $o(\Delta)$ by locally convex neighborhoods. The result is the following lemma.

Lemma 3: Let the vector θ be in the constrained parameter space $\Theta_C \subset \mathbb{R}^n$, and let $\{\boldsymbol{v}_i\}_{i=1}^k$ be k linearly independent vectors such that $\theta + \Delta_i \boldsymbol{v}_i + o(\Delta_i) \in \Theta_C$ for all Δ_i sufficiently small, $i = 1, \cdots, k$, where $o(\Delta_i)$ is a $\mathbb{R}^n$ vector whose length is of order $o(\Delta_i)$. Then the conclusions of Lemma 2 remain valid when the vectors $\theta + \Delta_i \boldsymbol{v}_i$ are replaced by $\theta + \Delta_i \boldsymbol{v} + o(\Delta_i)$, $i = 1, \cdots, k$.

Proof of Lemma 3: Similarly to (2), let $\delta' f_\theta$ denote the k-length vector of scalar differences $\delta' f_\theta = [\delta_1' f_\theta, \cdots, \delta_k' f_\theta]$ where

$$\delta_i' f_\theta \overset{\text{def}}{=} \frac{f_{\theta + \Delta_i \boldsymbol{v} + o(\Delta_i)} - f_\theta}{\Delta_i}. \tag{25}$$

Define $\delta' \boldsymbol{m}_\theta$ similarly. Let B_k' denote the Barankin bound of Proposition 1 formed with the k test points $\{\theta + \Delta_1 \boldsymbol{v}_1 + o(\Delta_1), \cdots, \theta + \Delta_k \boldsymbol{v}_k + o(\Delta_k)\}$. We need to establish that the limits $\limsup_{\Delta_1, \cdots, \Delta_k \to 0} B_k'$ and $\limsup_{\Delta_1, \cdots, \Delta_k \to 0} B_k$ (14) are identical.

By assumption (16) f_θ is continuous and therefore: $f_{\theta + \Delta_i \boldsymbol{v} + o(\Delta_i)} = f_{\theta + \Delta_i \boldsymbol{v}} + o(\Delta_i)$. In view of (25) this implies

$$\begin{aligned}\frac{\delta' f_\theta}{f_\theta} &= \frac{1}{f_\theta}\left[\frac{f_{\theta + \Delta_i \boldsymbol{v}} - f_\theta}{\Delta_i} + \frac{o(\Delta_i)}{\Delta_i}\right]_{i=1, \cdots, k} \\ &= \frac{\delta f_\theta}{f_\theta} + \frac{1}{f_\theta}\left[\frac{o(\Delta_i)}{\Delta_i}\right]_{i=1, \cdots, k} \\ &= \nabla \ln f_\theta K + o(1).\end{aligned}$$

Using the definition of the Fisher matrix and the continuity of the inverse of the full rank matrix $K^T J_\theta K$,

$$\begin{aligned}\left(E_\theta\left[\frac{\delta' f_\theta}{f_\theta}\right]^T\left[\frac{\delta' f_\theta}{f_\theta}\right]\right)^{-} &= \left(K^T J_\theta K + \underline{\underline{o}}(1)\right)^{-} \\ &= \left[K^T J_\theta K\right]^{-} + \underline{\underline{o}}(1). \end{aligned} \tag{26}$$

where $\underline{\underline{o}}(1)$ is a matrix that has $o(1)$ entries that go to zero as the Δ_i's go to zero. In a similar manner it can be shown that $\delta \boldsymbol{m}_\theta = \nabla \boldsymbol{m}_\theta K + \underline{\underline{o}}(1)$, which, when taken with (26), implies $B_k' = B_k + \underline{\underline{o}}(1)$. This establishes the lemma. □

C. Functional Constraints

Often the constrained parameter space Θ_C can be defined in terms of an implicit functional inequality constraint of the form

$$\mathcal{G}_\theta \le 0, \tag{27}$$

where $\mathcal{G} = [\mathcal{G}^1, \cdots, \mathcal{G}^q]^T$ is a vector function on $\mathbb{R}^n$, $\mathcal{G}: \mathbb{R}^n \to \mathbb{R}^q$, and the inequality is to be interpreted element by element. We will assume that the inequality constraints are consistent, i.e., there exists at least one $\theta \in \mathbb{R}^n$ that satisfies (27), and that $\mathcal{G}$ is continuously differentiable in the sense that the $q \times n$ gradient matrix

$$\nabla \mathcal{G}_\theta = \begin{bmatrix} \nabla \mathcal{G}_\theta^1 \\ \vdots \\ \nabla \mathcal{G}_\theta^q \end{bmatrix} = \begin{bmatrix} \frac{\partial \mathcal{G}_\theta^1}{\partial \theta_1} & \cdots & \frac{\partial \mathcal{G}_\theta^1}{\partial \theta_n} \\ \vdots & & \vdots \\ \frac{\partial \mathcal{G}_\theta^q}{\partial \theta_1} & \cdots & \frac{\partial \mathcal{G}_\theta^q}{\partial \theta_n} \end{bmatrix}, \tag{28}$$

exists and has continuous elements.

With the parameterization (27) of Θ_C, the boundary of Θ_C is defined as the set of points where at least one component, $\mathcal{G}_\theta^i$, of the vector function $\mathcal{G}_\theta$ is equal to zero. The interior of Θ_C is defined as the set $\{\theta: \mathcal{G}_\theta < 0\}$, where the strict inequality means $\mathcal{G}^i < 0$, for each $i = 1, \cdots, q$.

Note that equality constraints can be imbedded in (27) by letting $\mathcal{G}_\theta^i = -\mathcal{G}_\theta^j$ for some i, j, $i \neq j$. It is customary to extract the equality constraints from the inequality constraints (27), denoting what remains as *pure inequality constraints*. This yields the equivalent description of Θ_C

$$\boldsymbol{G}_\theta = 0, \tag{29}$$

$$\boldsymbol{H}_\theta \le 0, \tag{30}$$

where $\boldsymbol{G} = [G^1, \cdots, G^k]^T$ and $\boldsymbol{H} = [H^1, \cdots, H^l]^T$ are vector functions of θ, $\boldsymbol{G}: \mathbb{R}^n \to \mathbb{R}^k$, $\boldsymbol{H}: \mathbb{R}^n \to \mathbb{R}^l$. We will say that the equality constraint (29) is *active* if it restricts θ to a lower dimensional subset of $\mathbb{R}^n$. Otherwise the equality constraint is said to be *inactive*.

The decomposition (29) and (30) is accomplished by partitioning the constraint set Θ_C into a set of *regular points* and *nonregular points*.

Definition [16, Section 9.4]: The point $\theta_0 \in \mathbb{R}^n$ is called a regular point of the inequality $\mathcal{G}_\theta \le 0$ (a regular point of the constraint set Θ_C) if: $\mathcal{G}_{\theta_0} \le 0$ and if there exists a $\nu \in \mathbb{R}^n$ such that $\mathcal{G}_{\theta_0} + \nabla \mathcal{G}_{\theta_0} \nu < 0$.

There can be no active equality constraints at a regular point θ_0. Specifically, it can be shown that θ_0 is a regular point of Θ_C if and only if $\mathcal{G}_{\theta_0 - \Delta\nu} < 0$ for some $\nu \in \mathbb{R}^n$ and all sufficiently small $\Delta > 0$ (see proof of Lemma 4). This implies that there exits a sequence of interior points (e.g., $\{\theta_0 - \frac{1}{n}\nu\}_n$) that converge to θ_0. Hence regular points are points that are in the closure of the interior of Θ_C. In particular, all interior points of Θ_C are regular points and points on the boundary of pure inequality constraints $H_\theta \le 0$ are regular points. See Figs. 1 and 2 for graphical illustrations.

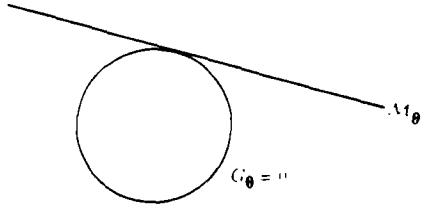


Fig. 1. Equality constraint $G_\theta \stackrel{\text{def}}{=} (\theta_1 - \theta_1^0)^2 + (\theta_2 - \theta_2^0)^2 - a^2 = 0$. Here θ can only vary along boundary of disk. Set of admissible directions, $\{\nu\}$, in which parameter can move must lie on tangent hyperplane $\mathcal{M}_\theta$. Since Θ_C has no interior points, there are no regular points of constraint set.

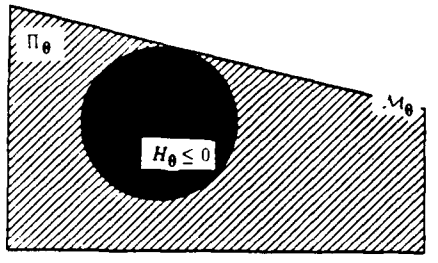


Fig. 2. Inequality-constraint $H_\theta \le 0$, where $H_\theta \stackrel{\text{def}}{=} (\theta_1 - \theta_1^0)^2 + (\theta_2 - \theta_2^0)^2 - a^2$. Here θ can move into interior of disk. Set of admissible directions is contained in half-space Π_θ that is supported by tangent hyperplane $\mathcal{M}_\theta$. Since any point $\theta \in \Theta_C$ can be represented as a limit of interior points, all points in Θ_C are regular points.

The following Lemma shows that if θ is a regular point of Θ_C the constrained CR bound is identical to the unconstrained CR bound.

Lemma 4: Assume that the conditions (15)–(18) of Lemma 2 hold. Let the parameter space Θ_C be defined by the general inequality constraint $\mathcal{G}_\theta \le 0$ where the vector function $\mathcal{G} = [\mathcal{G}^1, \cdots, \mathcal{G}^q]^T$ is differentiable. Let θ be a regular point of Θ_C. Then for any estimator $\hat{\theta}$ having mean m_θ, the estimator error covariance matrix Σ_θ satisfies the classical unconstrained CR matrix inequality

$$\Sigma_\theta \ge B_u, \tag{31}$$

where

$$B_u \stackrel{\text{def}}{=} [\nabla m_\theta] J_\theta^{-1} [\nabla m_\theta]^T, \tag{32}$$

and J_θ is the Fisher matrix (12). Equality holds in (31) if and only if there exists an $n \times n$ matrix Γ such that

$$\hat{\theta} - m_\theta = \Gamma[\Delta \ln f_\theta]^T. \tag{33}$$

If such an estimator $\hat{\theta}$ exists, it is called an efficient unconstrained estimator.

Proof of Lemma 4: Since θ is a regular point, there exists a $\nu \in \mathbb{R}^n$ such that for all Δ, $0 < \Delta < 1$, we have: $(1 - \Delta)\mathcal{G}_\theta < 0$ and $\Delta[\mathcal{G}_\theta + \nabla \mathcal{G}_\theta \nu] < 0$. Hence $(1 - \Delta)\mathcal{G}_\theta + \Delta[\mathcal{G}_\theta + \nabla \mathcal{G}_\theta \nu] = \mathcal{G}_\theta + \nabla \mathcal{G}_\theta \nu \Delta < 0$. Since for fixed ν

$$\mathcal{G}_{\theta - \Delta\nu} - \mathcal{G}_\theta - \nabla \mathcal{G}_\theta \nu \Delta = o(\Delta),$$

it follows that for all sufficiently small Δ, $\mathcal{G}_{\theta - \Delta\nu} < 0$. In a similar manner, it can be verified that there exists an $\epsilon > 0$ such that for all $\xi \in \mathbb{R}^n$ with length $\|\xi\| \le 1$

$$\mathcal{G}_{\theta - \Delta(\nu - \epsilon\xi)} < 0, \qquad \text{for all sufficiently small } \Delta > 0. \tag{34}$$

that is, $\theta + \Delta\nu$ is an interior point of Θ_C. Choose n linearly independent unit length vectors $\xi_1, \cdots, \xi_n$, and define $\nu_i = \nu + \epsilon\xi_i$, $i = 1, \cdots, n$. Then, using (34) it is seen that $\{\theta + \Delta\nu_i\}_{i=1}^n$ is a set of n linearly independent vectors contained Θ_C for all sufficiently small $\Delta > 0$. Application of Lemma 2 thus gives the lower bound on the covariance matrix

$$B_c = [\nabla m_\theta] A [A^T J_\theta A]^{-1} A^T [\nabla m_\theta]^T,$$

where A is any $n \times n$ matrix with identical column space as $[\nu_1, \cdots, \nu_n]$. But the column space of this latter matrix is identical to $\mathbb{R}^n$, by linear independence of the ν's, so taking $A = I$ in the previous equation for B_c we obtain

$$B_c = B_u = [\nabla m_\theta] J_\theta^{-1} [\nabla m_\theta]^T. \qquad \square$$

The bound (31) of Lemma 4 is identical to the classical multiparameter unconstrained CR bound [21], [27]. Since no equality constraints can be active at the regular points of Θ_C, the Lemma establishes that pure inequality constraints on θ do not affect the CR bound on the error covariance of estimators having a given mean gradient ∇m_θ. A number of parameter estimation problems have parameter constraint sets for which all of the points are regular. Examples include: orthant constraints, e.g., positivity of each of the elements θ_i in the parameter vector θ; range constraints, e.g., magnitude of θ_i less than 1; length constraints, e.g., $\sum_{i=1}^n \theta_i^2 \le 1$. For these types of constraints the classical unconstrained CR bound applies to all points in Θ_C.

On the other hand, many estimation problems are formulated with parameter constraint sets for which some or all of the points are not regular. In particular, as previously mentioned, for the case of active equality constraints (29), if Θ_C is a k-dimensional surface, $k < n$, then Θ_C contains no regular points. Examples of these problems are provided in Section III of this paper. For this case, the classical CR bound is invalid and bound reduction occurs due to the constraints.

We now consider the construction of a CR bound under continuously differentiable equality constraints. As-

sume the equality constraint $G_\theta = 0$ (29) is active at θ. Define the $k \times n$ gradient matrix, ∇G_θ, of the function G. Also define the hyperplane, $\mathcal{M}_\theta$, tangent to the constraint set Θ_c at the point θ:

$$\mathcal{M}_\theta = \{ y \in \mathcal{R}^n : \nabla G_\theta y = 0 \}. \tag{35}$$

If G is a linear function, e.g., $G_\theta = F\theta$ for some $n \times k$ matrix F, $\mathcal{M}_\theta \equiv \Theta_c$. Otherwise, when G is a continuously differentiable function, any set of points in Θ_c that are in the local Δ-neighborhood of the point $\theta \in \Theta_c$ are approximated to $o(\Delta)$ by a set of points in the tangent hyperplane $\mathcal{M}_\theta$. Using Lemma 3 this implies that the constrained CR bound $B_c(\theta)$ depends on the equality-constraint function G only through its associated tangent hyperplane at the point θ.

The constrained CR bound for smooth inequality constraints is given in the following theorem.

Theorem 1: Let the regularity conditions (15)–(18) of Lemma 2 be satisfied. Let the parameter space $\Theta_C \subset \mathcal{R}^n$ be defined by the consistent set of equality and pure inequality constraints: $G_\theta = 0$, $H_\theta \le 0$, where the vector functions $G = [G^1, \cdots, G^k]^T$ and $H = [H^1, \cdots, H^l]^T$ are continuously differentiable. Assume that the $k \times n$ gradient matrix ∇G_θ has rank p, $p \le k$. Then for any estimator $\hat{\theta}$ having mean m_θ, the estimator error covariance matrix Σ_θ satisfies the matrix inequality

$$\Sigma_\theta \ge B_c, \tag{36}$$

where

$$B_c = [\nabla m_\theta] Q_\theta J_\theta^{-1} [\nabla m_\theta]^T, \tag{37}$$

and Q_θ is the $n \times n$, idempotent, rank $n - p$ matrix

$$Q_\theta = \begin{cases} I, & \text{if } \theta \text{ is a regular point of } \Theta_C \\ I - J_\theta^{-1} [\nabla G_\theta]^T \{ [\nabla G_\theta] J_\theta^{-1} [\nabla G_\theta]^T \}^+ [\nabla G_\theta], & \text{otherwise.} \end{cases} \tag{38}$$

Furthermore, equality holds in (36) if and only if there exists an $n \times n$ matrix Γ such that

$$\hat{\theta} - m_\theta = \Gamma Q_\theta^T [\nabla \ln f_\theta]^T \quad \text{(w.p.1)}. \tag{39}$$

If such an estimator $\hat{\theta}$ exists, it is called an efficient constrained estimator.

Proof of Theorem 1: For the case that θ is a regular point, in view of Lemma 4, there is nothing left to prove. Conversely, suppose that θ is not a regular point. We will show that any sequence of test points in Θ_C that converges to θ approximates an equivalent sequence in $\mathcal{M}_\theta$. Then, for $0 < k \le n - p$, we define k sequences of test points in Θ_C whose associated approximating sequences in $\mathcal{M}_\theta$ converge to θ along linearly independent line paths $\theta + \Delta_1 \nu_1, \cdots, \theta + \Delta_k \nu_k, \Delta_1, \cdots, \Delta_k \to 0$, $\nu_i \in \mathcal{M}_\theta$. Finally, with B_b the Barankin bound (7), we show that $\limsup B_b$ is equal to the expression (37) for B_c, where the "limsup" is taken over all such sequences of test points.

Let $\xi = \xi(\Delta)$ be a vector such that $\|\xi(\Delta)\| \le \Delta \to 0$ and assume that $\theta + \xi$ is a vector in Θ_C that converges to $\theta \in \Theta_C$. By the assumed continuous differentiability of G_θ, $\nabla G_\theta \xi = o(\Delta)$:

$$\begin{aligned} 0 &= G_{\theta - \xi} - G_\theta \\ &= \nabla G_\theta \xi + o(\|\xi\|) \\ &= \nabla G_\theta \xi + o(\Delta), \end{aligned} \tag{40}$$

where $o(\Delta)$ is a vector of length $o(\Delta)$. Now define $P_{\mathcal{M}_\theta} = I - \nabla G_\theta^T [\nabla G_\theta \nabla G_\theta^T]^+ \nabla G_\theta$. $P_{\mathcal{M}_\theta}$ is an orthogonal-projection operator onto the null space of ∇G_θ, i.e., onto $\mathcal{M}_\theta$ [21, Section 1c.4]. This induces an orthogonal decomposition of $\xi = \xi(\Delta)$ relative to $\mathcal{M}_\theta$: $\xi = P_{\mathcal{M}_\theta} \xi + [I - P_{\mathcal{M}_\theta}] \xi$. From (40), $[I - P_{\mathcal{M}_\theta}] \xi = \nabla G_\theta^T [\nabla G_\theta \nabla G_\theta^T]^+ \nabla G_\theta \xi = o(\Delta)$ so that

$$\xi = P_{\mathcal{M}_\theta} \xi + o(\Delta). \tag{41}$$

Hence to order Δ, ξ is equal to the vector $P_{\mathcal{M}_\theta} \xi$ that is contained in $\mathcal{M}_\theta$.

Now let $\{\theta + \xi_i(\Delta_i)\}_{i=1}^k$ be k sequences in Θ_C indexed by $\Delta_1, \cdots, \Delta_k$ such that $P_{\mathcal{M}_\theta} \xi_i(\Delta_i) = \Delta_i \nu_i$, $i = 1, \cdots, k$, where $\nu_1, \cdots, \nu_k$ are fixed linearly independent vectors and $0 < k \le n - p$. Since G_θ is continuously differentiable and $\mathcal{M}_\theta$ has dimension $n - p$, such sequences exist [8, Prop. 26.1]. Hence, in view of (41), for fixed $\Delta_1, \cdots, \Delta_k$ the k test points $\theta + \xi_1(\Delta_1), \cdots, \theta + \xi_k(\Delta_k)$ are equal to $\theta + \Delta_1 \nu_1 + o(\Delta_1), \cdots, \theta + \Delta_k \nu_k + o(\Delta_k)$. Define $B_b(\theta + \xi_1(\Delta_1), \cdots, \theta + \xi_k(\Delta_k))$ the Barankin bound of Proposition 1 evaluated at these test points and define $B_c(\nu_1, \cdots, \nu_k)$ the CR bound of Lemma 2 evaluated with the direction vectors $\nu_1, \cdots, \nu_k$. Lemma 3 implies

$$\begin{aligned} &B_b(\theta + \xi_1(\Delta_1), \cdots, \theta + \xi_k(\Delta_k)) \\ &\quad = B_c(\nu_1, \cdots, \nu_k) + \underline{\underline{o}}(1) \\ &\quad = [\nabla m_\theta] A [A^T J_\theta A]^+ A^T [\nabla m_\theta]^T + \underline{\underline{o}}(1), \end{aligned} \tag{42}$$

where $\underline{\underline{o}}(1)$ is a matrix of $o(1)$ elements that go to zero as the Δ_i's go to zero, and A is an $n \times n$ matrix with column space equal to the span of $\nu_1, \cdots, \nu_k$.

Next we show that if $\nu_1, \cdots, \nu_k$ and $\nu_1', \cdots, \nu_k'$ are sets of vectors in $\mathcal{M}_\theta$ such that $\text{span}\{\nu_1, \cdots, \nu_k\} \supset \text{span}\{\nu_1', \cdots, \nu_k'\}$ then $A[A^T J_\theta A]^+ A^T \ge B[B^T J_\theta B]^+ B^T$, where A and B are $n \times n$ matrices which have identical column spaces as $\text{span}\{\nu_1, \cdots, \nu_k\}$ and $\text{span}\{\nu_1', \cdots, \nu_k'\}$, respectively. Since by definition $\nu_i \in \mathcal{M}_\theta$, $i = 1, \cdots, k$, this will establish that the matrix $[\nabla m_\theta] A [A^T J_\theta A]^+ A^T [\nabla m_\theta]^T$ on the right of (42) is maximized when the column space of A is equal to $\mathcal{M}_\theta$. With $J_\theta^{-1/2}$ the positive square root matrix corresponding to J_θ^{-1}, the previous relation between the two spans holds if and only if $\text{span}\{J_\theta^{-1/2} \nu_1, \cdots, J_\theta^{-1/2} \nu_k\} \supset \text{span}\{J_\theta^{-1/2} \nu_1', \cdots, J_\theta^{-1/2} \nu_k'\}$. Hence it is sufficient to show that $A[A^T A]^+ A^T \ge B[B^T B]^+ B^T$ when the column space of A contains the column space of B. Now $A[A^T A]^+ A^T$ and $I - B[B^T B]^+ B^T$ are idempotent,

symmetric, orthogonal-projection matrices onto the column space of A and the null space of B [21, Section 1c.4], respectively. Therefore, since the column space of A contains the column space of B: $A[A^TA]^-A^TB = B$ and $B^TA[A^TA]^-A^T = B^T$. Since idempotent matrices are nonnegative definite, it follows that $A[A^TA]^-A^T - B[B^TB]^-B^T = A[A^TA]^-A^T\{I - B[B^TB]^-B^T\} = A[A^TA]^-A^T\{I - B[B^TB]^-B^T\}A[A^TA]^-A^T$, which is nonnegative-definite. Therefore we have from (42)

$$\limsup B_n = [\nabla \boldsymbol{m}_\theta]A\left[A^TJ_\theta A\right]^- A^T[\nabla \boldsymbol{m}_\theta]^T. \quad (43)$$

where A is a matrix whose column span equals $\mathscr{H}_\theta$.

Finally we show that the column span of Q_θ (38) is equal to $\mathscr{H}_\theta$ and that, setting $A = Q_\theta$ in (43), we obtain (37). Since ∇G_θ has rank p, there exists a row-echelon representation

$$\nabla G_\theta = T\begin{bmatrix} B \\ O_1 \end{bmatrix}$$

where T is a nonsingular $k \times k$ matrix, B is a $p \times n$ full-row-rank matrix, and O_1 is a $(k-p)\times n$ matrix of zeros. Let O_2, O_3 and O_4 denote matrices of zeros having dimensions $(k-p)\times(k-p)$, $(k-p)\times p$ and $k \times n$, respectively. Use of (38) and (65) of Lemma 5 in the Appendix results in

$$\begin{aligned}
\nabla G_\theta Q_\theta &= T\begin{bmatrix} B \\ O_1 \end{bmatrix}\left[I - J_\theta^{-1}\begin{bmatrix} B^T & O_1^T \end{bmatrix}\right. \\
&\quad \left. \cdot T^T\left(T\begin{bmatrix} B \\ O_1 \end{bmatrix}J_\theta^{-1}\begin{bmatrix} B^T & O_1^T \end{bmatrix}T^T\right)^- T\begin{bmatrix} B \\ O_1 \end{bmatrix}\right] \\
&= T\left[\begin{bmatrix} B \\ O_1 \end{bmatrix} - \begin{bmatrix} B \\ O_1 \end{bmatrix}J_\theta^{-1}\begin{bmatrix} B^T & O_1^T \end{bmatrix}\right. \\
&\quad \left. \cdot\left(\begin{bmatrix} B \\ O_1 \end{bmatrix}J_\theta^{-1}\begin{bmatrix} B^T & O_1^T \end{bmatrix}\right)^-\begin{bmatrix} B \\ O_1 \end{bmatrix}\right] \\
&= T\left[\begin{bmatrix} B \\ O_1 \end{bmatrix} - \begin{bmatrix} BJ_\theta^{-1}B^T & O_3^T \\ O_3 & O_2 \end{bmatrix}\right. \\
&\quad \left. \cdot\begin{bmatrix} BJ_\theta^{-1}B^T & O_3^T \\ O_3 & O_2 \end{bmatrix}^-\begin{bmatrix} B \\ O_1 \end{bmatrix}\right] \\
&= T\left[\begin{bmatrix} B \\ O_1 \end{bmatrix} - \begin{bmatrix} I & O_3^T \\ O_3 & O_2 \end{bmatrix}\begin{bmatrix} B \\ O_1 \end{bmatrix}\right] \\
&= O_4,
\end{aligned}$$

where the invertibility of the full rank $p \times p$ matrix $BJ_\theta^{-1}B^T$ has been used on the third line of this equation. This establishes that the columns of Q_θ are contained in the hyperplane $\mathscr{H}_\theta$. A straightforward calculation shows that $Q_\theta Q_\theta = Q_\theta$ and $Q_\theta^TQ_\theta^T = Q_\theta^T$, i.e., both Q_θ and Q_θ^T are idempotent. Hence the rank of Q_θ is equal to its trace

$$\begin{aligned}
&\operatorname{rank}\{Q_\theta\} \\
&\quad = \operatorname{tr}\{Q_\theta\} \\
&\quad = \operatorname{tr}\left\{I - J_\theta^{-1}[\nabla G_\theta]^T\left\{[\nabla G_\theta]J_\theta^{-1}[\nabla G_\theta]^T\right\}^-[\nabla G_\theta]\right\} \\
&\quad = n - \operatorname{tr}\left\{[\nabla G_\theta]J_\theta^{-1}[\nabla G_\theta]^T\left\{[\nabla G_\theta]J_\theta^{-1}[\nabla G_\theta]^T\right\}^-\right\} \\
&\quad = n - p.
\end{aligned}$$

and Q_θ has $n-p$ linearly independent columns. Since these columns are contained in $\mathscr{H}_\theta$ and since $n - \operatorname{rank}\{\nabla G_\theta\} = n - p$ is the dimension of $\mathscr{H}_\theta$, this establishes that the column space of Q_θ is identical to $\mathscr{H}_\theta$. Hence, using $A = Q_\theta$ in Lemma 2, we obtain the bound

$$B_n = [\nabla \boldsymbol{m}_\theta]Q_\theta\left[Q_\theta^TJ_\theta Q_\theta\right]^- Q_\theta^T[\nabla \boldsymbol{m}_\theta]^T$$

Now it is evident from symmetry that $Q_\theta J_\theta^- = J_\theta^- Q_\theta^T$. Define $J_{t_\theta}^- \overset{\text{def}}{=} Q_\theta^TJ_\theta Q_\theta$. One can verify that the matrix $Q_\theta J_\theta^{-1} = J_\theta^- Q_\theta^T$ satisfies the Penrose conditions (9) for the pseudo-inverse, $J_{t_\theta}^-$, of J_{t_θ}. Using these results and the fact that Q_θ and Q_θ^T are idempotent results in

$$\begin{aligned}
Q_\theta\left[Q_\theta^TJ_\theta Q_\theta\right]^- Q_\theta^T &= Q_\theta J_{t_\theta}^- Q_\theta^T \\
&= Q_\theta\left[Q_\theta J_\theta^-\right]Q_\theta^T \\
&= Q_\theta J_\theta^- Q_\theta^T \\
&= Q_\theta Q_\theta J_\theta^- \\
&= Q_\theta J_\theta^-.
\end{aligned}$$

Hence (37) is established. □

In reference to Theorem 1 we make the following remarks.

Remark 1: If the set of constraints $G_\theta = 0$ is defined so that the rows of ∇G_θ are linearly independent, the $k \times k$ matrix $[\nabla G_\theta]J_\theta^{-1}[\nabla G_\theta]^T$ will be of full rank and Q_θ (38) will only involve the more familiar inverse matrix $([\nabla G_\theta]J_\theta^{-1}[\nabla G_\theta]^T)^{-1}$. Although a reformulation eliminating redundant constraints can always be accomplished, frequently the most natural description of a constraint involves a rank-deficient ∇G_θ, e.g., see Example 4 of Section III. In this case the general result of Theorem 1 is applicable.

Remark 2: Comparison between the bound of Lemma 4 and the bound of Theorem 1 indicates that the presence of constraints on the parameter space has the effect of reducing the rank of the Fisher information matrix. In particular if the k equality constraints $G_\theta = 0$ reduce the dimension of the parameter space from n to $n-p$ then the rank n inverse Fisher information J_θ^{-1} becomes the rank $n-p$ inverse *constrained Fisher information* $Q_\theta J_\theta^{-1}$. Hence active equality constraints have the effect of reducing the rank of the Fisher information matrix. In the proof of Theorem 1 it was shown that the column span of Q_θ is the tangent hyperplane $\mathscr{H}_\theta$, and that $Q_\theta J_\theta^{-1} = Q_\theta[Q_\theta^TJ_\theta Q_\theta]^-Q_\theta^T$. Furthermore, by Lemma 2,

$$Q_\theta\left[Q_\theta^TJ_\theta Q_\theta\right]^- Q_\theta^T = A\left[A^TJ_\theta A\right]^- A^T$$

if A has the same column span as Q_θ. Using these facts we have

$$Q_\theta J_\theta^{-1} = P_{\mathscr{H}_\theta}\left[P_{\mathscr{H}_\theta}^TJ_\theta P_{\mathscr{H}_\theta}\right]^- P_{\mathscr{H}_\theta}^T,$$

where $P_{\mathscr{H}_\theta} = I - [\nabla G_\theta]^T([\nabla G_\theta][\nabla G_\theta]^T)^-[\nabla G_\theta]$ is the $n \times n$ orthogonal-projection matrix that projects vectors in Ξ^n onto $\mathscr{H}_\theta$. Hence the inverse constrained Fisher matrix $Q_\theta J_\theta^{-1}$ is obtained from a projection of the rows and columns of the unconstrained Fisher matrix J_θ onto the tangent hyperplanes of the constraint set.

Remark 3: The matrix B_c (37) in Theorem 1 can be represented as the quantity

$$B_c = E\{[\nabla m_\theta P_{\mathscr{M}_\theta}]\}\{E\{[\nabla \ln f_\theta P_{\mathscr{M}_\theta}]^T[\nabla \ln f_\theta P_{\mathscr{M}_\theta}]\}^{\#}\}$$
$$\cdot E\{[\nabla m_\theta P_{\mathscr{M}_\theta}]^T\}^{\#},$$

where $P_{\mathscr{M}_\theta}$ is the projection operator defined in Remark 2. The vectors $\nabla m_\theta P_{\mathscr{M}_\theta}$ and $\nabla \ln f_\theta P_{\mathscr{M}_\theta}$ are the projections of the unconstrained gradients of the mean and log-likelihood (score) functions onto the constraint tangent hyperplane $\mathscr{M}_\theta$, that is, these vectors correspond to constrained gradient vectors. In [10] these constrained gradient vectors were used along with Lemma 1 to give an alternative derivation of the inequality $\Sigma_\theta \geq B_c$.

Remark 4: Theorem 1 indicates that a certain bound reduction is induced by adding constraints on θ. In particular, it is easy to show that the constrained CR bound B_c of Theorem 1 is always less than the unconstrained CR bound B_u in the sense that $B_u - B_c$ is nonnegative definite. This follows from: 1) the idempotence of $I - Q_\theta$; 2) the symmetry of J_θ^{-1} and $Q_\theta J_\theta^{-1}$, which imply that $(I - Q_\theta)J_\theta^{-1} = J_\theta^{-1}(I - Q_\theta)^T$; and 3) the nonnegative definiteness of J_θ^{-1}. In particular, for unbiased estimators $\nabla m_\theta = I$ and

$$\begin{aligned} B_c &= Q_\theta J_\theta^{-1} \\ &= J_\theta^{-1} - (I - Q_\theta)J_\theta^{-1} \\ &= J_\theta^{-1} - (I - Q_\theta)(I - Q_\theta)J_\theta^{-1} \\ &= J_\theta^{-1} - (I - Q_\theta)J_\theta^{-1}(I - Q_\theta)^T \\ &\leq J_\theta^{-1} = B_u. \end{aligned} \tag{44}$$

An important implication of (44) is that the incorporation of constraints can only reduce the CR bound on the component error variances.

Remark 5: In many examples of interest Q_θ is nondiagonal, accounting for the functional relationships between individual components of θ introduced by the constraint. Thus even if J_θ is diagonal, suggesting uncorrelated unconstrained estimator errors, the rank-reduced inverse Fisher information $Q_\theta J_\theta^{-1}$ in Theorem 1 can have off-diagonal terms, suggesting correlated constrained estimator errors.

Remark 6: A result of Lemma 4 and Theorem 1 is that pure inequality constaints $H_\theta \leq 0$ do not affect the CR bound on error covariance of estimators with a given mean gradient ∇m_θ. This is true even when θ is on the boundary of this set. An interpretation of this fact is obtained by recalling that the Fisher information matrix J_θ (12) is a function of the gradient of the likelihood surface at θ. For a smooth surface, the gradient of the surface at θ is completely determined by the set of directional derivatives along directions contained in a convex cone with vertex at θ, e.g., the half-space indicated in Fig. 2. In the case of one-dimensional differentiable functions, this simply reflects the equivalence of right and left derivatives. Therefore, the restriction of allowable local variations of a parameter at the boundary of $H_\theta \leq 0$ does not affect the CR bound.

Remark 7: While Theorem 1 is stated as a lower bound on the estimator error covariance matrix, it can be used to specify a bound on the mean-square error (mse) matrix, $E_\theta\{(\hat{\theta} - \theta)(\hat{\theta} - \theta)^T\}$. Specifically, since the mse matrix is equal to $\Sigma_\theta + (m_\theta - \theta)(m_\theta - \theta)^T$, application of the theorem gives a constrained CR bound on mse:

$$E_\theta\{(\hat{\theta} - \theta)(\hat{\theta} - \theta)^T\} \geq B_c + (m_\theta - \theta)(m_\theta - \theta)^T,$$

where B_c is given by (37).

Remark 8: Remarks 6 and 7 notwithstanding, when Θ_c corresponds to a pure inequality constraint Theorem 1 does not imply that improvement in mse is impossible. Indeed the minimum-distance projection of an unconstrained estimator $\hat{\theta}_u$ onto Θ_c can yield an estimator with lower mse than that of $\hat{\theta}_u$. Such an estimator arises in the example studied in [18]. However, if the estimators differ the projected estimator may have a different mean from that of $\hat{\theta}_u$ which generally is not differentiable, whereas Theorem 1 applies to classes of estimators with identical differentiable means m_θ.

Remark 9: In the course of proof of Theorem 1 it was established that the lower bound B_c (36) is the tightest bound of the form (14) in the sense that $B_c = \limsup_{\Delta_1,\cdots,\Delta_k \to 0} B_h(\theta + \xi_1(\Delta_1), \cdots, \theta + \xi_k(\Delta_k))$ where $\{\theta + \xi_i(\Delta_i)\}_{i=1}^k$ are k arbitrary sequences converging to θ along paths whose projections onto the tangent plane $\mathscr{M}_\theta$ are k linearly independent line segments, $0 < k \leq n - p$. For linear constraints and exponential families of f_θ more can be proven: B_c is the "lim sup" of the Barankin bound B_h (7) with respect to arbitrary sequences of test points converging to θ, i.e., B_c is the tightest local Barankin bound.

III. Applications

In this section we illustrate the application of the constrained CR bound (37) by specializing to the cases of linear and quadratic constraints.

Example 1) Linearly Constrained Gauss-Markov Problem: Let F be an $m \times n$ matrix of rank n, $n \leq m$, and suppose that one observes the vector $X \in \mathbb{R}^m$,

$$X = F\theta + \eta,$$

where $\theta \in \mathbb{R}^n$, $\eta \in \mathbb{R}^m$ and η is a zero-mean Gaussian vector with nonsingular $m \times m$ covariance matrix $K = E_\theta\{\eta\eta^T\}$. Since the model is linear and Gaussian, the Fisher information matrix is simply calculated as $J \overset{\text{def}}{=} J_\theta = F^T K^{-1} F$, which is independent of θ. Furthermore, by the Gauss-Markov theorem [21, Ch. 4], the minimum variance unbiased (MVU) estimator $\hat{\theta}_u$ is a linear function of X,

$$\hat{\theta}_u = J^{-1} F^T K^{-1} X.$$

The error covariance of $\hat{\theta}_u$ is

$$\Sigma_\theta^u = J^{-1}.$$

Thus $\hat{\theta}_u$ achieves the unconstrained CR bound, (31) of

Lemma 2, for unbiased estimators. (Recall that for unbiased estimators, $\nabla \boldsymbol{m}_\theta = I$.)

Consider, however, the problem of estimating $\boldsymbol{\theta}$ subject to the k linear equality constraints $\boldsymbol{G}_\theta = A\boldsymbol{\theta} = 0$, where A is a $k \times n$ matrix, $k \le n$. Using the fact that $\nabla \boldsymbol{G}_\theta = A$, Theorem 1 gives the constrained CR bound $B_c = [\nabla \boldsymbol{m}_\theta] Q J^{-1} [\nabla \boldsymbol{m}_\theta]^T$, where

$$Q \overset{\text{def}}{=} Q_\theta = I - J^{-1}A^T\left[AJ^{-1}A^T\right]^{-} A.$$

Since the matrix Q is independent of $\boldsymbol{\theta}$, one can define the estimator

$$\begin{aligned}\hat{\boldsymbol{\theta}} &= QJ^{-1}F^TK^{-1}X \\ &= Q\hat{\boldsymbol{\theta}}_u. \end{aligned} \tag{45}$$

Due to the constraint $A\boldsymbol{\theta} = 0$, $\hat{\boldsymbol{\theta}}$ is unbiased

$$E_\theta\{\hat{\boldsymbol{\theta}}\} = \left(I - J^{-1}A^T\left[AJ^{-1}A^T\right]^{-} A\right)\boldsymbol{\theta} = \boldsymbol{\theta}.$$

The error covariance of $\hat{\boldsymbol{\theta}}$ can be calculated directly from (45) using the idempotence of Q:

$$\begin{aligned}\Sigma_\theta &= QJ^{-1}Q^T \\ &= QQJ^{-1} \\ &= QJ^{-1} \\ &= B_c,\end{aligned}$$

where B_c is the constrained CR bound, (37) of Theorem 1, for unbiased estimation. This establishes that: 1) the estimator $\hat{\boldsymbol{\theta}}$ of (45) is the MVU constrained estimator, and 2) the constrained CR bound of Theorem 1 is achievable for the Gaussian linear model with linear constraints.

Example 2) Image Reconstruction with a Support Constraint: Support constraints are frequently used in image reconstruction problems such as those arising in tomographic imaging [24], [29] and phase retrieval [5], [9]. Suppose that the parameter vector of interest consists of a sampled two-dimensional image that is represented by a complex-valued vector with elements $\theta_{(k_1,k_2)}$, $k_1, k_2 = 0, 1, \cdots, M-1$. We will represent the parameter vector $\boldsymbol{\theta}$ as the $\mathbb{R}^{2M \times M}$ vector

$$\boldsymbol{\theta} = \left[\theta^R_{(0,0)}, \theta^I_{(0,0)}, \theta^R_{(0,1)}, \theta^I_{(0,1)}, \cdots, \theta^R_{(M-1,M-1)}, \theta^I_{(M-1,M-1)}\right]^T,$$

where the superscripts R and I denote respectively the real and imaginary parts of $\theta_{(k_1,k_2)}$.

If the support of the object is known, it can be used as a constraint in the estimation of $\boldsymbol{\theta}$. Let S be the support of $\boldsymbol{\theta}$,

$$S = \left\{(k_1,k_2): \theta_{(k_1,k_2)} \neq 0;\ k_1, k_2 = 0, 1, \cdots, M-1\right\}.$$

Let $\mathbf{1}_S$ denote the $2M^2 \times 2M^2$ diagonal matrix with $[\mathbf{1}_S]_{ii} = 1$ if the ith element of $\boldsymbol{\theta}$ lies inside the support set S and $[\mathbf{1}_S]_{ii} = 0$ otherwise, i.e., $\mathbf{1}_S$ is the matrix indicator function of S. The support constraint then has the form $\boldsymbol{G}_\theta = [I - \mathbf{1}_S]\boldsymbol{\theta} = 0$. From Theorem 1 we have the constrained CR bound $B_c = [\nabla \boldsymbol{m}_\theta] Q_\theta J_\theta^{-} [\nabla \boldsymbol{m}_\theta]^T$. Using $\nabla \boldsymbol{G}_\theta = [I - \mathbf{1}_S]$ it is easy to verify:

$$\begin{aligned} Q_\theta J_\theta^{-} &= J_\theta^{-} - J_\theta^{-}[I - \mathbf{1}_S]^T \\ &\quad \cdot \left\{[I - \mathbf{1}_S]J_\theta^{-}[I - \mathbf{1}_S]^T\right\}^{-}[I - \mathbf{1}_S]J_\theta^{-} \\ &= T\mathcal{J}^{-1} - \mathcal{J}^{-1}[I - \mathbf{1}_S]^T \\ &\quad \cdot \left\{[I - \mathbf{1}_S]\mathcal{J}^{-1}[I - \mathbf{1}_S]^T\right\}^{-}[I - \mathbf{1}_S]\mathcal{J}^{-1}T^T, \end{aligned} \tag{46}$$

where $\mathcal{J} = T^T J_\theta T$ and T is an orthogonal matrix such that

$$\mathbf{1}_S = T\begin{bmatrix} I & O_1 \\ O_1^T & O_2 \end{bmatrix} T^T. \tag{47}$$

where O_1 and O_2 are zero matrices. In other words, T is a transformation that rearranges the image pixels so that the support is in the upper left hand corner of the image. Now let $\mathcal{J}$ and $\mathcal{J}^{-1}$ have the partitions

$$\mathcal{J} = \begin{bmatrix} A & B \\ B^T & C \end{bmatrix} \tag{48}$$

$$\mathcal{J}^{-1} = \begin{bmatrix} K & L \\ L^T & M \end{bmatrix}, \tag{49}$$

where A and K are matrices of the same dimension as the identity matrix I on the right-hand side of (47). With this notation $[I - \mathbf{1}_S]\mathcal{J}^{-1}[I - \mathbf{1}_S]^T$ is the partitioned matrix $\begin{bmatrix} O_0 & O \\ O^T & M \end{bmatrix}$, where O_0 is a zero matrix of the appropriate dimensions. Therefore the pseudo-inverse on the right-hand side of (46) is simply $\begin{bmatrix} O_0 & O \\ O^T & M^{-} \end{bmatrix}$. Performing the rest of the matrix algebra indicated on the right-hand side of (46) we obtain

$$Q_\theta J_\theta^{-} = T\begin{bmatrix} K - LM^{-1}L^T & O_1 \\ O_1^T & O_2 \end{bmatrix} T^T.$$

Using identities for the inverse of a partitioned matrix [11, Theorem 8.2.1] and the definitions of A, B, C and K, L, M, (48) and (49), the matrix $K - LM^{-1}L^T$ can be identified as the inverse of the block matrix A. Hence,

$$\begin{aligned} Q_\theta J_\theta^{-} &= T\begin{bmatrix} A^{-1} & O_1 \\ O_1^T & O_2 \end{bmatrix} T^T \\ &= T\begin{bmatrix} A & O_1 \\ O_1^T & O_2 \end{bmatrix}^{-} T^T \\ &= T\left\{\begin{bmatrix} I & O_1 \\ O_1^T & O_2 \end{bmatrix}\begin{bmatrix} A & B \\ B^T & C \end{bmatrix}\begin{bmatrix} I & O_1 \\ O_1^T & O_2 \end{bmatrix}\right\}^{-} T^T \\ &= (\mathbf{1}_S J_\theta \mathbf{1}_S)^{-}, \end{aligned} \tag{50}$$

where the last equality follows by the orthogonality of T, the application of (47), (48), and the identification $T\mathcal{J}T^T = J_\theta$. For the case of unbiased estimation $\nabla \boldsymbol{m}_\theta = I$ and (50) is the constrained CR bound. Comparing the constrained CR bound (50) to the unconstrained CR bound

J_θ^{-1} it is evident that the incorporation of support constraints has the effect of zeroing out those rows and columns of the Fisher information matrix corresponding to image pixels θ for which it is known *a priori* that the pixel values are zero.

It is useful to compare the covariance of the estimator errors within the support region for the unconstrained cases. Using the same transformation T (47) as before, we can assume without loss of generality that the support is in the upper left corner of image, i.e., the support matrix indicator function is $\mathbf{1}_S = \begin{bmatrix} I & O_1 \\ O_1^T & O_2 \end{bmatrix}$. In this case the unconstrained bound within the support region is $(A - BC^{-1}B^T)^{-1}$, which is the upper left block element of the inverse matrix $J_\theta^{-1} = \mathcal{F}^{-1}$ (48), while the constrained CR bound for these pixels is A^{-1}. If the Fisher matrix is block diagonal then B is a matrix of zeros in (48), indicating that the errors of an unbiased efficient estimator of pixels inside and outside of the support region are uncorrelated; in this case the constrained CR bound is identical to the unconstrained CR bound. If the Fisher matrix is not block diagonal, however, there may be substantial reduction in the constrained CR bound over the support region. It is also significant that, unless J_θ is block diagonal, setting the pixels of an efficient (CR bound achieving) unconstrained estimator to zero outside the image support region does not produce an estimator that achieves the constrained CR bound. This is in contrast to the results obtained in [5] for diagonal J_θ.

Example 3) Spectrum Estimation with Power Constraints: When there is prior information on the power of a random process over some regions of frequency, it is reasonable to expect that the achieveable error covariance of spectral estimators will be affected. This example quantifies the effect of such prior information on the constrained CR bound.

Let $\{X_t\}_{t=1}^N$ be a segment of a real wide sense stationary random process with power spectral density (PSD) $\{\mathcal{P}(f)\}_{f \in [-1/2, 1/2]}$. The objective is to estimate the PSD, $\theta_i \overset{\text{def}}{=} \mathcal{P}(f_i)$, at n distinct frequencies $f_1, \cdots, f_n$. Let the average power of $\{X_t\}$ be known over P nonoverlapping frequency bands

$$\sum_{S_p} \theta_k = E_p, \qquad p = 1, \cdots, P, \tag{51}$$

where S_p is the index set of the pth frequency band, and E_p is the known average power of $\{X_t\}$ over this frequency band. The equations (51) correspond to P linear constraints on the unknown PSD, known as the *P-point constraint* in robust Wiener filtering theory [20]. The concatenation of the P equalities (51) gives the P equations

$$G_\theta = \begin{bmatrix} \chi_1^T \\ \vdots \\ \chi_P^T \end{bmatrix} \begin{bmatrix} \theta_1 \\ \vdots \\ \theta_n \end{bmatrix} = \begin{bmatrix} E_1 \\ \vdots \\ E_P \end{bmatrix} \tag{52}$$

where χ_p is an $n \times 1$ column vector with ith element equal to 1 if $i \in S_p$ and 0 otherwise, i.e., χ_p is the vector indicator function of S_p. The gradient matrix ∇G_θ is given by $\nabla G_\theta = [\chi_1, \cdots, \chi_P]^T$, resulting in

$$Q_\theta J_\theta^{-1} = J_\theta^{-1} - J_\theta^{-1}[\chi_1 \cdots \chi_P] \cdot \begin{bmatrix} \chi_1^T J_\theta^{-1} \chi_1 & \cdots & \chi_1^T J_\theta^{-1} \chi_P \\ \vdots & \ddots & \vdots \\ \chi_P^T J_\theta^{-1} \chi_1 & \cdots & \chi_P^T J_\theta^{-1} \chi_P \end{bmatrix}^{-} \begin{bmatrix} \chi_1^T \\ \vdots \\ \chi_P^T \end{bmatrix} J_\theta^{-1}. \tag{53}$$

The structure of $Q_\theta J_\theta^{-1}$ is considerably simplified when J_θ is the diagonal matrix:

$$J_\theta = \text{diag}_i\{\theta_i^{-2}\},$$

which is appropriate for the case of Gaussian observations $\{X_t\}_{t=1}^N$ and large N. Since the frequency bands $\{S_p\}$ are nonoverlapping the pseudo-inverse on the right-hand side of (53) becomes the pseudo-inverse of a diagonal matrix and

$$Q_\theta J_\theta^{-1} = J_\theta^{-1} - \sum_{i=1}^{P} \frac{J_\theta^{-1} \chi_i \chi_i^T J_\theta^{-1}}{\chi_i^T J_\theta^{-1} \chi_i}. \tag{54}$$

Let $e_l = [0, \cdots, 0, 1, 0, \cdots, 0]^T$ denote the lth standard basis vector in $\mathbb{R}^n$. Let l be an index in the constraint set S_p. Then for an unbiased estimator, $\hat{\theta}$, the constrained CR bound on the variance of the lth component, $\hat{\theta}_l$, is obtained from (54)

$$\begin{aligned} [B_c]_{ll} &= e_l^T B_c e_l \\ &= e_l^T \left(J_\theta^{-1} - \sum_{i=1}^{P} \frac{J_\theta^{-1} \chi_i \chi_i^T J_\theta^{-1}}{\chi_i^T J_\theta^{-1} \chi_i} \right) e_l \\ &= [J_\theta^{-1}]_{ll} \left(1 - \frac{[J_\theta]_{ll}^{-1}}{\sum_{\{i \in S_p\}} [J_\theta^{-1}]_{ii}} \right). \end{aligned} \tag{55}$$

Using the unconstrained CR bound $[B_u]_{ll} = \theta_l^2 = \mathcal{P}^2(f_l)$, we obtain the relative reduction in the CR bound due to the constraint

$$\frac{[B_c]_{ll}}{[B_u]_{ll}} = 1 - \frac{1}{1 + \sum_{\{i \in S_p, i \neq l\}} \frac{\mathcal{P}^2(f_i)}{\mathcal{P}^2(f_l)}}. \tag{56}$$

Since the term on the right hand side of (56) is between 0 and 1, the average power constraint induces a CR bound reduction on the component PSD estimation errors. The bound reduction factor (56) is independent of the other constraint sets S_k, $k = 1, \cdots, P$, $k \neq p$, and therefore average power constraints over S_p do not affect PSD estimator errors at frequencies outside of S_p. The amount of bound reduction depends on two factors: 1) the relative magnitude of the spectral component of interest, $\mathcal{P}^2(f_l)$, compared to the magnitude of the other frequency components within the frequency band S_p; and 2) the length, $|S_p|$ = number of indices, of S_p. In particular, little or no reduction in the variance bound occurs for the case where $\mathcal{P}^2(f_i)/\mathcal{P}^2(f_l)$ is small for all $i \in S_p, i \neq l$. However, when $\mathcal{P}^2(f_l)$ is large compared to the other $\mathcal{P}^2(f_i)$,

$f_i \in S_p$, a substantial reduction in the bound occurs. This implies that the most bound reduction will be achieved over those constraint regions S where the PSD has a high dynamic range, i.e., large peaks. The particular dynamic range required for a significant bound reduction is proportional to $|S_p|$. As a rule of thumb, for a reduction in the CR bound at frequency f_i by a factor α or more, the ratio of $\mathscr{P}(f_i)$ to the root mean-squared value of the remaining spectral components in S_p,

$$\frac{\mathscr{P}(f_i)}{\bar{\mathscr{P}}} \triangleq \frac{\mathscr{P}(f_i)}{\sqrt{\dfrac{1}{|S_p|-1}\displaystyle\sum_{f_l \in S_p, l \neq i} \mathscr{P}^2(f_l)}},$$

must satisfy

$$\frac{\mathscr{P}(f_i)}{\bar{\mathscr{P}}} \geq \sqrt{\frac{1-\alpha}{\alpha}\left[|S_p|-1\right]}.$$

Example 4) Signal Subspace Constraints: Signal subspace constraints are used in sensor array processing estimation problems to take account of a particular structure of the array covariance matrix [14]. Specifically, assume that p zero-mean Gaussian signals arrive at different angles of incidence on an m-sensor array having a zero-mean, spatially incoherent array noise of power σ^2. Further, assume that $p < m$. Then the covariance matrix of the set of sensor outputs has the singular value decomposition

$$R = \sum_{i=1}^{p} \lambda_i' v_i v_i^H + \sigma^2 I = \sum_{i=1}^{m} \lambda_i v_i v_i^H,$$

where $\{v_i\}_{i=1}^m$ are the eigenvectors of R and $\{\lambda_i\}_{i=1}^m$ are the eigenvalues:

$$\lambda_i = \begin{cases} \lambda_i' + \sigma^2, & i = 1, \cdots, p \\ \sigma^2, & i = p+1, \cdots, m \end{cases}$$

and $\{\lambda_i'\}_{i=1}^p$ denote the signal-dependent eigenvalues of R. The span of $v_1, \cdots, v_p$ is called the *signal subspace*.

Consider the problem of estimating the eigenvalues of R when p is known but all of the other parameters are unknown. This partial knowledge induces the following constraints on the λ_i:

$$\begin{aligned} &a) \quad \lambda_j > 0, && j = 1, \cdots, m \\ &b) \quad \lambda_j \geq \frac{1}{m-p} \sum_{i=p+1}^{m} \lambda_i, && j = 1, \cdots, m \\ &c) \quad \lambda_j - \frac{1}{m-p} \sum_{i=p+1}^{m} \lambda_i = 0, && j = p+1, \cdots, m \end{aligned} \quad (57)$$

where constraint a) arises from the assumed positive-definiteness of R, constraint b) takes account of the positivity of the signal eigenvalues $\{\lambda_i'\}_{i=1}^p$, and constraint c) reflects the equality of the $m-p$ noise eigenvalues.

Let each unknown eigenvector $v_i \in \mathbb{R}^m$ be parameterized by its $m-1$ direction cosines, $\rho_i = [\rho_{i,1}, \cdots, \rho_{i,m-1}]^T$, $i = 1, \cdots, m$. The combination of the m unknown eigenvalues and the $m(m-1)$ unknown direction cosines yields the $n = m^2$ element parameter vector $\theta = [\lambda_m, \cdots, \lambda_1, \rho_m^T, \cdots, \rho_1^T]^T$. The constraint c) can be then be expressed as the $(m-p) \times n$ matrix constraint

$$G_\theta = \left[I_{m-p} - \frac{1}{m-p} \mathbf{1}\mathbf{1}^T O_1 \right] \theta,$$

where I_k denotes a $k \times k$ identity matrix, O_1 is a $(m-p) \times (n-m-p)$ matrix of zero entries, and $\mathbf{1}$ is a $(m-p)$-vector of ones.

Observe that the rows of ∇G_θ are not linearly independent due to the fact that there is one redundant constraint in c) of (57). Observe also that the equality constraint c) creates a dimension $n-m+p+1$ linear subspace in the unconstrained parameter space $\mathbb{R}^n$. Hence, despite the presence of inequality constaints a) and b), the constrained parameter space Θ_c contains no regular points, and, by Theorem 1, the constraints a), b) do not impact the form of the constrained CR bound.

As in Example 2, partition J_θ according to

$$J_\theta = \begin{bmatrix} A & B \\ B^T & C \end{bmatrix},$$

where A is $(m-p) \times (m-p)$, B is $(m-p) \times (n-m-p)$, and C is $(n-m+p) \times (n-m+p)$. Then the $n \times n$ inverse constrained Fisher matrix, $Q_\theta J_\theta^{-1}$, of Theorem 1 is given by

$$Q_\theta J_\theta^{-1} = J_\theta^{-1} - J_\theta^{-1} \begin{bmatrix} Z^- & O_1 \\ O_1^T & O_2 \end{bmatrix} J_\theta^{-1},$$

where O_1 and O_2 are zero matrices of dimensions $(m-p) \times (n-m+p)$ and $(n-m+p) \times (n-m+p)$, respectively, and Z is the $(m-p) \times (m-p)$ matrix

$$\begin{aligned} Z &= \nabla G_\theta J_\theta^{-1} [\nabla G_\theta]^T \\ &= \left[I_{m-p} - \frac{1}{m-p} \mathbf{1}\mathbf{1}^T \right] [A - BC^{-1}B^T] \\ &\quad \cdot \left[I_{m-p} - \frac{1}{m-p} \mathbf{1}\mathbf{1}^T \right]. \end{aligned} \quad (58)$$

As a simple example, consider the case where the Fisher information matrix is block diagonal with: $B = O_1$ and $A = \sigma_\lambda^{-2} I_{m-p}$. Then $Z = \sigma_\lambda^{-2}[I_{m-p} - \frac{1}{m-p}\mathbf{1}\mathbf{1}^T]$. Using condition 3) of (9) it is easy to show that $Z^- = \sigma_\lambda^2[I_{m-p} - \frac{1}{m-p}\mathbf{1}\mathbf{1}^T]$. This results in

$$Q_\theta J_\theta^{-1} = \begin{bmatrix} \sigma_\lambda^2 \dfrac{1}{m-p} \mathbf{1}\mathbf{1}^T & O_1 \\ O_1^T & C^{-1} \end{bmatrix}. \quad (59)$$

Suppose there exists an efficient unbiased estimator $\hat{\theta}_u$ for the eigenvalues and eigenvectors which satisfies constraints a) and b), and assume that the Fisher information is block diagonal as previously specified. The right-hand side of (59) is then the covariance matrix of the estimator obtained by replacing each of the $m-p$ noise eigenvalue estimates in $\hat{\theta}_u$ by their average $\frac{1}{m-p}\sum_{i=p+1}^m \hat{\lambda}_i$. Hence, if an efficient unconstrained estimator of the eigenvalues

can be found that has positive elements, the estimator obtained by averaging over the $m-p$ smallest eigenvalues of the efficient estimator achieves the constrained CR bound.

Example 5) Signal Estimation with Power Constraints: Consider the problem of estimating the discrete-time signal waveform, $\theta_1, \cdots, \theta_n$, subject to constraints on the squared-modulus of the DFT of $\boldsymbol{\theta}$. Here, the sum of the squared moduli over each of P nonoverlapping frequency intervals is constrained to be equal to known constants E_p, $p = 1, \cdots, P$. While similar to the case studied in Example 3, this problem involves nonlinear quadratic constraints on the parameters, and time rather than frequency domain estimation is performed.

Let $W = [W_1, \cdots, W_n]$ denote the $n \times n$ unitary matrix of orthonormal discrete Fourier transform columns: $W_l = 1/\sqrt{n}[1, e^{-j2\pi l/n}, \cdots, e^{-j2\pi(n-1)l/n}]^T$, $l = 1, \cdots, n$. Now suppose that for $P \leq n$ the constraint takes the form

$$\sum_{i \in S_p} |[W'\boldsymbol{\theta}]_i|^2 = E_p, \qquad p = 1, 2, \cdots, P. \tag{60}$$

Here, S_p denotes the index set of the pth interval and $[W'\boldsymbol{\theta}]_i$ is ith component of the n-point DFT of $\boldsymbol{\theta}$. When $P = n$, (60) specifies the modulus Fourier transform of $\boldsymbol{\theta}$. As in Example 2, we let $\mathbf{1}_p$ denote the $n \times n$ diagonal matrix with $[\mathbf{1}_p]_{ii} = 1$ if $i \in S_p$ and $[\mathbf{1}_p]_{ii} = 0$ otherwise. Then the constraint (60) can be written as the set of P equations

$$\boldsymbol{G}_{\boldsymbol{\theta}} = \begin{bmatrix} \boldsymbol{\theta}^T W^H \mathbf{1}_1 W \boldsymbol{\theta} \\ \vdots \\ \boldsymbol{\theta}^T W^H \mathbf{1}_P W \boldsymbol{\theta} \end{bmatrix} - \begin{bmatrix} E_1 \\ \vdots \\ E_P \end{bmatrix} = \begin{bmatrix} 0 \\ \vdots \\ 0 \end{bmatrix},$$

where the superscript H denotes hermitian transpose. The gradient $\nabla \boldsymbol{G}_{\boldsymbol{\theta}}$ is the $P \times n$ matrix

$$\nabla \boldsymbol{G}_{\boldsymbol{\theta}} = \begin{bmatrix} 2\boldsymbol{\theta}^T W^H \mathbf{1}_1 W \\ \vdots \\ 2\boldsymbol{\theta}^T W^H \mathbf{1}_P W \end{bmatrix}. \tag{61}$$

We now specialize to the linear observation model:

$$X_i = \theta_i + \eta_i, \qquad i = 1, \cdots, n$$

where η_i is a zero-mean Gaussian white noise with variance σ^2. Recalling Example 1, $J_{\boldsymbol{\theta}}$ can be seen to be the scaled identity matrix $\sigma^{-2}I$. Let O denote the $n \times n$ zero matrix. Using (61) and the fact that the intervals S_p are nonoverlapping $\mathbf{1}_i \mathbf{1}_j = O$, $i \neq j$, the inverse constrained Fisher matrix of Theorem 1 is the $n \times n$ matrix

$$Q_{\boldsymbol{\theta}} J_{\boldsymbol{\theta}}^{-1} = \sigma^2 W^H \left[I - \sum_{i=1}^{P} \frac{\mathbf{1}_i W \boldsymbol{\theta}\boldsymbol{\theta}^T W^H \mathbf{1}_i}{\boldsymbol{\theta}^T W^H \mathbf{1}_i W \boldsymbol{\theta}} \right] W. \tag{62}$$

Since W is the (linear) DFT operator, the matrix $\sigma^2[I - \Sigma(\cdot)]$ on the right-hand side of (62) is the inverse constrained Fisher information matrix for estimation of the DFT $W\boldsymbol{\theta}$. As in Example 3, let the index l be constrained in S_p. Then the ratio between the constrained and unconstrained CR bounds on the variance, $\text{var}\{[W\hat{\boldsymbol{\theta}}]_l\} = E_{\boldsymbol{\theta}}\{[W\hat{\boldsymbol{\theta}}]_l - E_{\boldsymbol{\theta}}[W\hat{\boldsymbol{\theta}}]_l|^2\}$, is obtained by evaluating the quadratic forms $e^T W^H Q_{\boldsymbol{\theta}} J_{\boldsymbol{\theta}}^{-1} W e$ and $e^T W^H J_{\boldsymbol{\theta}}^{-1} W e$

$$\frac{[B]_{ll}}{[B_u]_{ll}} = 1 - \frac{1}{1 - \sum_{i \in S_p, i \neq l} \frac{|[W\boldsymbol{\theta}]_i|^2}{|[W\boldsymbol{\theta}]_l|^2}}. \tag{63}$$

This is of identical form to the expression obtained for constrained PSD estimation, (55) of Example 3, when the power spectral density, $\mathcal{P}(f)$, is identified with the magnitude spectrum $|[W\boldsymbol{\theta}]_i|$, $i = 1, \cdots, n$. For unbiased estimators, a bound on the total mean-squared error in estimating the time domain signal θ_i's can be determined from (63) by using the unitary property of the DFT matrix W (Parseval's Theorem):

$$\begin{aligned} \sum_{i=1}^{n} E_{\boldsymbol{\theta}}|\theta_i - \hat{\theta}_i|^2 &= \text{tr}\{\Sigma_{\boldsymbol{\theta}}\} \\ &\geq \text{tr}\{Q_{\boldsymbol{\theta}} J_{\boldsymbol{\theta}}^{-1}\} \\ &= \sigma^2 \text{tr}\left\{ W^H \left[I - \sum_{i=1}^{P} \frac{\mathbf{1}_i W \boldsymbol{\theta}\boldsymbol{\theta}^T W^H \mathbf{1}_i}{\boldsymbol{\theta}^T W^H \mathbf{1}_i W \boldsymbol{\theta}} \right] W \right\} \\ &= \sigma^2 \text{tr}\left\{ I - \sum_{i=1}^{P} \frac{\mathbf{1}_i W \boldsymbol{\theta}\boldsymbol{\theta}^T W^H \mathbf{1}_i}{\boldsymbol{\theta}^T W^H \mathbf{1}_i W \boldsymbol{\theta}} \right\} \\ &= \sigma^2 [n - P]. \end{aligned}$$

Therefore, on the average, the constraints produce a factor of $1 - P/n$ reduction in the CR bound on the variances of unbiased estimators of the θ_i's.

IV. Conclusion

A constrained Cramér–Rao (CR) lower bound on the error covariance of estimators of multidimensional parameters has been obtained. The constrained CR bound was derived from a limiting form of a multiparameter Barankin-type bound. For constraint sets defined by a general smooth functional inequality constraint of the form $\mathcal{J}_{\boldsymbol{\theta}} \leq 0$, the constrained CR bound is equivalent to the unconstrained CR bound evaluated with a "constrained" Fisher information matrix. This constrained Fisher matrix was shown to be identical to the classical unconstrained Fisher matrix at all regular points of the constraint set, e.g., at interior points. However at nonregular points, such as points governed by equality constraints, the constrained Fisher matrix is a rank-deficient matrix. This constrained Fisher matrix is equivalent to a matrix of orthogonal projections of the rows and columns of the unconstrained Fisher matrix onto the tangent hyperplanes of the constraint set. The simple form of the constrained CR bound allows the effect of particular equality and inequality constraints to be easily studied through comparisons between the constrained and unconstrained CR bounds. It was shown that the incorporation of functional constraints necessarily decreases the CR bound for unbiased estimators. Not surprisingly, the constrained bound was shown to be achievable for the lin-

early-constrained Gauss–Markov problem. To illustrate the application of the constrained CR bound, several applications in the area of signal processing were considered. These included support constraints in image reconstruction, signal subspace constraints in array processing, and average power constraints in spectral estimation and in signal estimation.

In their present form, the results obtained in this paper only directly apply to a finite dimensional parameter space and a non-stochastic constraint. A generalization of these results to infinite dimensional parameter spaces would be useful for the study of constraints in filtering, prediction, and smoothing problems. Theorem 1 could perhaps be applied to complete separable infinite-dimensional parameter spaces, e.g., a separable Hilbert space, by taking the formal limit of the elements of the matrix bound (37) as the dimension of the indicated matrices goes to infinity. Stochastic constraints are of interest when the constraint depends on the particular realization of the statistical experiment, and they provide a model for partially-known constraints. A main difficulty in obtaining a generalization of the constrained CR bound to differentiable stochastic constraints is that the column space of the constraint equality gradient matrix, $\nabla_\theta \mathcal{G}_\theta$, is in general a random set and therefore Lemma 2 cannot be applied. On the other hand, a tractible analysis may be possible for simple stochastic constraints such as constraints obtained from random perturbations of the constraint function $\mathcal{G}_\theta$.

Acknowledgment

The authors would like to thank Dr. John Jayne of M.I.T. Lincoln Laboratory for pointing out an error in an early draft of this manuscript and for helpful discussions concerning this paper.

Appendix

Lemma 5: Let Q be an arbitrary $n \times m$ matrix and T be any $m \times m$ invertible matrix. Then

$$QT[T^TQ^TQT]^+T^TQ^T = Q[Q^TQ]^+Q^T, \tag{64}$$

where the plus sign denotes (Moore–Penrose) pseudo-inverse. As a consequence, if R is an arbitrary $m \times n$ matrix, J is an $m \times m$ positive definite matrix, and T is an invertible $n \times n$ matrix, then:

$$RT[T^TR^TJRT]^+T^TR^T = R[R^TJR]^+R^T. \tag{65}$$

Proof of Lemma 5: Let the left and right sides of the identity (64) be denoted as the $n \times n$ matrices P_1 and P_2, respectively. It is easily verified that P_1 and P_2 are symmetric and idempotent. Therefore P_1 and P_2 are orthogonal projections onto respective subsets, $\mathcal{M}_1$ and $\mathcal{M}_2$ say, of $\mathbb{R}^m$ [22, Section 105]. Furthermore, using properties 1)–3) of (9), it is easily verified that $P_2P_1P_2 = P_2$ and $P_1P_2P_1 = P_1$. Equivalently, since P_1P_2 and P_2P_1 are projections onto respective subsets of $\mathcal{M}_1 \cap \mathcal{M}_2$: $P_2P_1 = P_1$ and $P_1P_2 = P_2$. However, $P_1P_1 = P_1$ implies $P_2 \geq P_1$ [22, Prop. d of Section 104], and hence $P_1 = P_2$.

To show (65), first observe that, due to positive definiteness, there exists an invertible matrix $J^{1/2}$ such that $J^{1/2}J^{1/2}$. Define $Q = J^{1/2}R$. Then (65) reads

$$J^{-1/2}QT[T^TQ^TQT]^+T^TQ^TJ^{-1/2} = J^{-1/2}Q[Q^TQ]^+Q^TJ^{-1/2},$$

which follows directly from (64). This finishes the proof of Lemma 5. □

References

[1] J. Aitchison and S. D. Silvey, "Maximum-likelihood estimation of parameters subject to restraints," *Ann. Math. Stat.*, vol. 29, pp. 813–828, 1958.
[2] A. Albert, *Regression and the Moore–Penrose Pseudo-Inverse*. New York: Academic Press, 1972.
[3] P. Billingsley, *Probability and Measure*, 2nd ed. New York: Wiley, 1986.
[4] J. M. Begun, W. J. Hall, W.-M. Huang, and J. A. Wellner, "Information and asymptotic efficiency in parametric–nonparametric models," *Ann. Statist.*, vol. 11, no. 2, pp. 432–452, 1983.
[5] J. N. Cederquist and C. C. Wackerman, "Phase-retrieval error: A lower bound," *J. Opt. Soc. Amer. A*, vol. 4, pp. 1788–1792, Sept. 1987.
[6] D. G. Chapman and H. Robbins, "Minimum variance estimation without regularity assumptions," *Ann. Math. Statist.*, vol. 22, pp. 581–586, 1951.
[7] H. R. Chernoff, "On the distribution of the likelihood ratio," *Ann. Math. Statist.*, vol. 25, pp. 573–578, 1954.
[8] K. Deimling, *Nonlinear Functional Analysis*. New York: Springer-Verlag, 1985.
[9] J. R. Fienup, "Reconstruction and synthesis applications of an iterative algorithm," in *Transformations in Optical Signal Processing*, W. T. Rhodes, J. R. Fienup, and B. E. A. Saleh, Eds., *Proc. SPIE*, vol. 373, pp. 147–160, 1981.
[10] J. D. Gorman and A. O. Hero, "Lower bounds on parametric estimators with constraints," in *Proc. 4th ASSP Workshop on Spectrum Estimation and Modeling*, Minneapolis, MN, August 3–5, 1988, pp. 223–228.
[11] F. A. Graybill, *Matrices with Applications in Statistics*. Belmont, CA: Wadsworth, 1969.
[12] R. A. Horn and C. G. Johnson, *Matrix Analysis*. Cambridge, MA: Cambridge Univ. Press, 1985.
[13] I. A. Ibragimov and R. Z. Has'minskii, *Statistical Estimation — Asymptotic Theory*. New York: Springer-Verlag, 1981.
[14] R. Kumaresan and D. W. Tufts, "Estimating the angles of arrival of multiple plane waves," *IEEE Trans. Aerosp. Electron. Syst.*, vol. AES-19, pp. 134–139, Jan. 1983.
[15] L. LeCam, *Asymptotic Methods in Statistical Decision Theory*. New York: Springer-Verlag, 1984.
[16] D. G. Luenberger, *Optimization by Vector Space Methods*. New York: Wiley, 1969.
[17] R. J. McAulay and E. M. Hofstetter, "Barankin bounds on parameter estimation," *IEEE Trans. Inform. Theory*, vol. IT-17, no. 6, pp. 669–675, Nov. 1971.
[18] P. A. P. Moran, "Maximum-likelihood estimation in non-standard conditions," *Proc. Camb. Phil. Soc.*, vol. 70, 1971, pp. 441–450.
[19] J. Pfanzagl and W. Wefelmeyer, *Contributions to a General Asymptotic Statistical Theory*, *Lecture Notes in Statistics*, vol. 13. New York: Springer-Verlag, 1982.
[20] K. S. Vastola and V. Poor, "Robust Wiener–Kolmogorov theory," *IEEE Trans. Inform. Theory*, vol. IT-30, no. 2, pp. 316–327, Mar. 1984.
[21] C. R. Rao, *Linear Statistical Inference and its Applications*, 2nd ed. New York: Wiley, 1973.
[22] F. Riesz and B. Sz.-Nagy, *Functional Analysis*, 2nd ed. New York: Frederick Ungar Pub., New York, 1978.
[23] S. E. Roucos and D. G. Childers, "A two-dimensional maximum entropy spectral estimator," *IEEE Trans. on Inform. Theory*, vol. IT-26, no. 5, pp. 554–560, Sept. 1980.

[24] M. I. Sezan and H. Stark, "Applications of convex projection theory to image recovery in tomography and related areas," in *Image Recovery: Theory and Application*. H. Stark Ed. Orlando, FL: Academic Press, 1987.

[25] D. L. Snyder, M. I. Miller, T. J. Schulz, and J. D. O'Sullivan, "Constrained probability-density estimation from noisy data," in *Proc. 22nd Ann. Conf. Inform. Sci. Syst.* Princeton Univ., Mar. 1988.

[26] G. W. Stewart, "On the continuity of the generalized inverse," *SIAM J. Appl. Math.*, vol. 17, no. 1, pp. 33–45, Jan. 1969.

[27] H. L. Van Trees, *Detection, Estimation and Modulation Theory, Part I*. New York: Wiley, 1968.

[28] P. A. Yansouni and R. J. Inkol, "The use of linear constraints to reduce the variance of time of arrival difference estimates for source location," *IEEE Trans. Acoust. Speech and Signal Processing*, vol. ASSP-32, pp. 907–912, Aug. 1984.

[29] D. C. Youla, "Mathematical theory of image restoration by the method of convex projections," in *Image Recovery: Theory and Application*. H. Stark Ed. Orlando, FL: Academic Press, 1987.

APPENDIX C
REFLECTION BY AN ILLUMINATED CYLINDER

Figure C-1 depicts a cross-section of a cylinder illuminated from an angle Ψ_i below the horizon and viewed from an angle Ψ_o below the horizon. For simplicity we assume that the sun illuminates it at broadside and we are viewing it from broadside. Let θ_o be the clockwise angle between the viewing angle and the surface normal at a given point on the surface, and let θ_i be the counterclockwise angle between the illumination angle and the surface normal at a given point. Then

$$\theta_i = \pi - \Psi_i - \Psi_o - \theta_o \quad . \tag{C-1}$$

Let the diameter of the cylinder be d_o. The distance from a given point on the surface and the center of the cylinder, projected along the perpendicular to the viewer's line-of-sight is

$$x_p = (d_o/2) \sin \theta_o \quad . \tag{C-2}$$

The illuminated part of the cylinder seen by the viewer goes from $\theta_o = \pi/2$, at the edge of the cylinder as seen by the viewer, where

$$\theta_i(\theta_o = \pi/2) = \pi/2 - \Psi_i - \Psi_o \tag{C-3}$$

and

$$x_p(\theta_o = \pi/2) = d_o/2 \tag{C-4}$$

to the edge of the shadow (at $\theta_i = \pi/2$), where

$$\theta_o(\theta_i = \pi/2) = \pi/2 - \Psi_i - \Psi_o \tag{C-5}$$

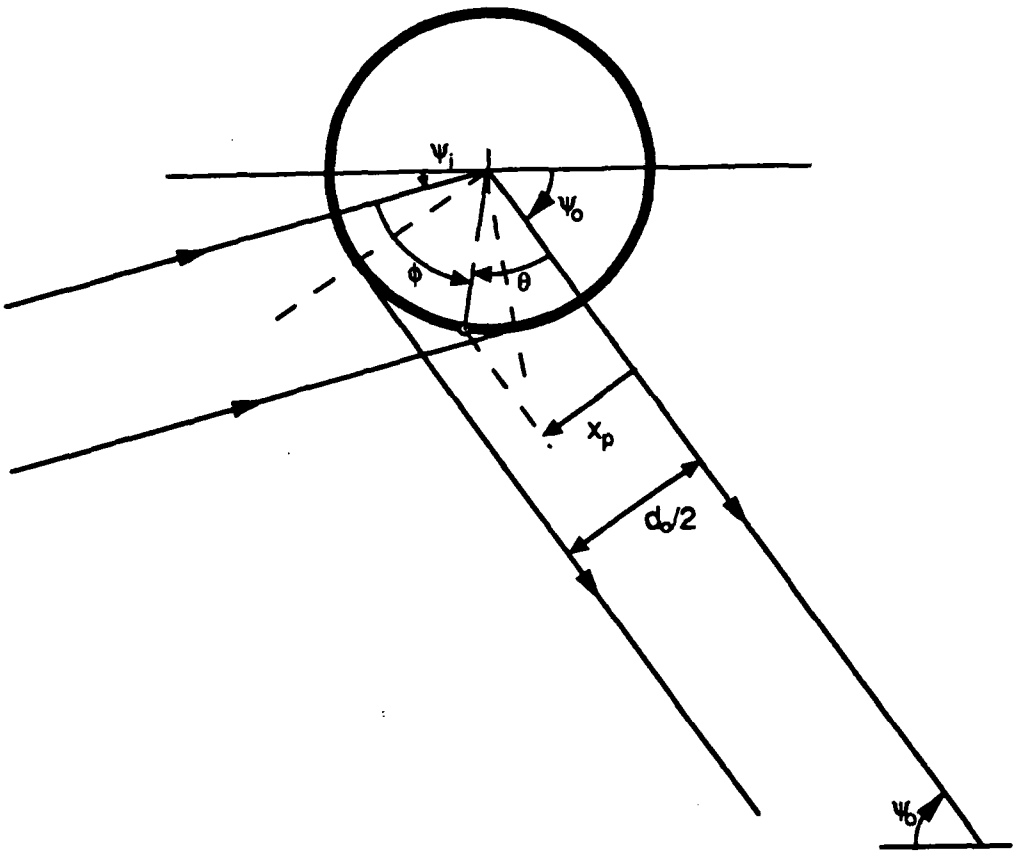


Figure C-1. Illumination and Viewing Geometry of a Cylinder (Axis Normal to the Plane of the Page).

and

$$x_p(\theta_i = \pi/2) = (d_o/2)\cos(\Psi_i + \Psi_o) \quad . \tag{C-6}$$

(Note that a negative value of x_p would indicate a point counterclockwise from the viewing angle.) The illuminated width of the cylinder from the viewing perspective is

$$d_p = x_p(\theta_o = \pi/2) - x_p(\theta_i = \pi/2) = (d_o/2)\,[1 - \cos(\Psi_i + \Psi_o)] \quad . \tag{C-7}$$

The angles over which light is scattered toward the viewing angle are

$$\pi/2 - \Psi_i - \Psi_o \leq \theta_i \leq \pi/2 \tag{C-8a}$$

and

$$\pi/2 \geq \theta_o \geq \pi/2 - \Psi_i - \Psi_o \quad . \tag{C-8b}$$

Consider a reflecting area

$$\Delta A = \Delta y(d_o/2)\, d\theta_o \tag{C-9}$$

where Δy is the length of the area along the axis perpendicular to the plane of Figure C-1. For a Lambertian surface of reflectivity r_o, the energy density scattered into a solid angle $d\Omega_o$ is

$$\begin{aligned} \Delta E_o &= (r_o/\pi)\, E_i \cos\theta_i \cos\theta_o\, \Delta A\, d\Omega_o \\ &= (r_o\, d_o/2\pi)\, \Delta y\, E_i \cos\theta_i \cos\theta_o\, d\theta_o\, d\Omega_o \end{aligned} \tag{C-10}$$

[where Eq. (C-8) is valid], where E_i is the incident energy density.

The apparent spatial brightness distribution of the object depends on the projection of this area onto the plane perpendicular to the line-of-sight, where the projected area is

$$\Delta A_p = \Delta A \cos \theta_o \quad . \tag{C-11}$$

This comes from the fact that from Eq. (C-2)

$$dx_p = (d_o/2) \cos \theta_o \, d\theta_o \quad . \tag{C-12}$$

Consequently the projected energy is

$$\Delta E_{op} = (r_o/\pi) \, \Delta y \, E_i \cos \theta_i \, dx_p \, d\Omega_o \quad . \tag{C-13}$$

From Eqs. (C-1) and (C-2),

$$\cos \theta_i = \cos(\pi - \Psi_i - \Psi_o - \theta_o)$$

$$= \sin(\Psi_i + \Psi_o - \pi/2 + \theta_o) \tag{C-14a}$$

$$= -\cos(\Psi_i + \Psi_o) \cos \theta_o + \sin(\Psi_i + \Psi_o) \sin \theta_o \tag{C-14b}$$

$$= -\cos(\Psi_i + \Psi_o) \sqrt{1 - (2x_p/d_o)^2} + \sin(\Psi_i + \Psi_o) \, (2x_p/d_o) . \tag{C-14c}$$

Eqs. (C-13) and (C-14c) give the apparent brightness as a function of the viewed coordinate, x_p. Figure C-2 shows Eq. (C-13) plotted as a function of x_p for $(\Psi_i + \Psi_o)$ = 20° to 180° in 20° increments. [Note that the apparently continuous curves at x_p = 1 are pairs of curves that approach x_p = 1 with the same values and slopes, one of the pair of curves for $(\Psi_i + \Psi_o)$ and the other for $(180° - \Psi_i - \Psi_o)$.]

Now consider the total energy density arriving at a detector. This can be obtained by a integrating Eq. (C-13) over x_p or by integrating Eq. (C-10), using Eq. (C-14b), over $d\theta_o$. The latter is given by

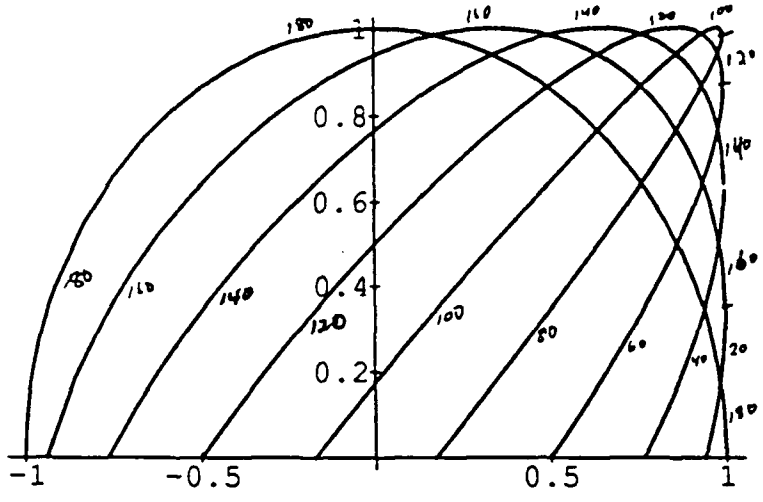


Figure C-2. Relative Brightness (Intensity) Across the Projected Image of the Cylinder, for $(\Psi_i + \Psi_o)$ = 20° to 180° in 20° Increments.

$$\int_0^L \Delta y \int_{\pi/2-\Psi_i-\Psi_o}^{\pi/2} (d_o/2)\, d\theta_o\, \Delta E_o$$

$$= L(r_o\, d_o/2\pi)\, E_i\, d\Omega_o \int_{\pi/2-\Psi_i-\Psi_o}^{\pi/2} [-\cos(\Psi_i + \Psi_o) \cos\theta_o + \sin(\Psi_i + \Psi_o) \sin\theta_o] \cos\theta_o\, d\theta_o$$

$$= L(r_o d_o/4\pi)\, E_i [\sin(\Psi_i + \Psi_o) - (\Psi_i + \Psi_o) \cos(\Psi_i + \Psi_o)]\, d\Omega_o$$

$$\equiv L(r_o d_o/2\pi)\, E_i\, V(\Psi_i + \Psi_o)\, d\Omega_o \qquad \text{(C-15)}$$

where $d\Omega_o$ is the angular subtense of a detector as viewed from the target. The function

$$V(\Psi_i + \Psi_o) = (1/2)\, [\sin(\Psi_i + \Psi_o) - (\Psi_i + \Psi_o) \cos(\Psi_i + \Psi_o)] \qquad \text{(C-16)}$$

is shown in Figure C-3, plotted as a function of $(\Psi_i + \Psi_o)$ (in degrees).

Example

Suppose that $\Psi_i = 10°$ and $\Psi_o = 55°$ so that $(\Psi_i + \Psi_o) = 65°$. Then the illuminated region can be seen for $25° \leq \theta_o \leq 90°$, for which $90° \geq \theta_i \geq 25°$. The relative perceived reflectivity, given by Eqs. (13) and (14) is proportional to $\cos\theta_i$, which varies from 0 to $\cos 25° = 0.906$, following a curve slightly above the 60° curve shown in Figure C-2. The perceived width of the cylinder is $d_p = (d_o/2)\,(1 - \cos 65°) = 0.577\,(d_o/2)$; so for a 0.8m diameter cylinder, the perceived width would be 0.231m. From Eq. (C-16), $V(65°) = 0.213$ (as compared with the maximum possible value of $\pi/2$).

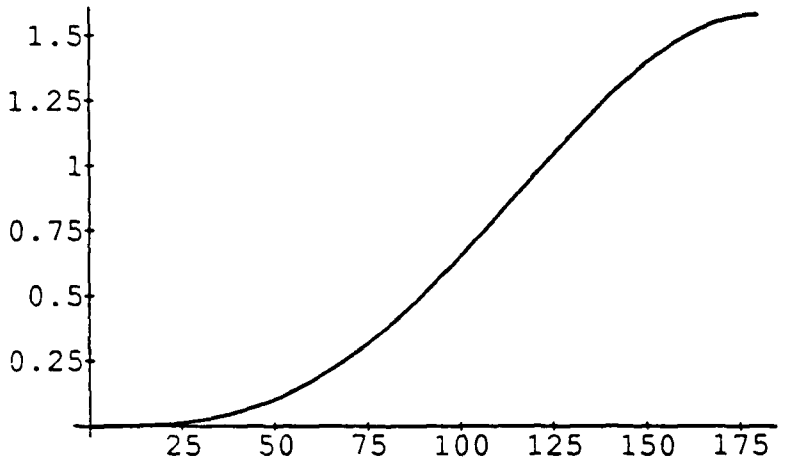


Figure C-3. Relative Energy Density Arriving at an Aperture-Plane Detector as a Function of $(\Psi_i + \Psi_o)$ (in Degrees).

APPENDIX D

IMAGE RECONSTRUCTION FOR AN ABERRATED AMPLITUDE INTERFEROMETER WITH A PARTIALLY-FILLED APERTURE

J. R. Fienup and J. D. Gorman

Optical Science Laboratory
Environmental Research Institute of Michigan
P.O. Box 8618, Ann Arbor, Michigan 48107-8618, USA

Presented at the NOAO/ESO Conference on
"High Resolution Imaging by Interferometry,"
Garching bei München, West Germany, 15-18 March 1988

IMAGE RECONSTRUCTION FOR AN ABERRATED AMPLITUDE INTERFEROMETER WITH A PARTIALLY-FILLED APERTURE

J. R. Fienup and J. D. Gorman

Optical Science Laboratory
Environmental Research Institute of Michigan
P.O. Box 8618, Ann Arbor, Michigan 48107-8618, USA

1. Introduction

Measurements obtained with an aperture-plane amplitude interferometer [1,2] utilizing a 180° rotational shear through a telescope having a partially-filled aperture can have missing spatial frequency bands corresponding to the aperture-plane regions where there is no aperture fill. The system transfer function in this case is a scaled version of the telescope aperture function, and the missing Fourier-domain data causes the resulting images to be highly distorted [Figures 1(e) and 1(f) for example]. This is in contrast to conventional focal-plane imaging systems where the system transfer function is the autocorrelation of the telescope aperture function, in which case Wiener filtering can often be used to level the transfer function. A further complication arising in the image formation process is that for realistic imaging systems, the phase of the Fourier data can be corrupted or completely lost in the presence of atmospheric turbulence or optical aberrations. Thus there are two difficulties which complicate the reconstruction of images from aperture-plane amplitude interferometer measurements: the absence of particular spatial frequency bands and the possible corruption of the phase of the data. This paper examines an application of the iterative Fourier transform algorithm [3,4,5] to the problem of reconstructing missing Fourier-domain information from aberrated aperture-plane amplitude interferometer measurements to obtain diffraction-limited imagery corresponding to a filled aperture.

Common examples of collection systems having partially-filled apertures are telescopes with a central obscuration, for which the low and middle spatial-frequency bands are blocked by the secondary mirror; and segmented or multiple-mirror telescopes, for which certain middle and high spatial-frequency bands are lost. Two types of aperture functions were considered in this study: an annular aperture which will be denoted as aperture A, and a segmented aperture consisting of a hexagonal arrangement of seven smaller circular apertures, which shall denoted as aperture H. Figure 1(d) shows the original object used in the simulations. Its Fourier transform, the magnitude of which is shown in Figure 1(a), was multiplied by aperture A to obtain the aperture plane data of Figure 1(b) and corresponding image, Figure 1(e). The transform was also multiplied by aperture H to obtain the aperture plane data of Figure 1(c)

and corresponding image, Figure 1(f). The dynamic range of the Fourier magnitude data in Figures 1(a-c), 3(a-c) and 4(a-c) is quite large, so the square root of the Fourier magnitude is displayed.

Three scenarios were investigated: (i) the Fourier magnitude is measured over a filled aperture, (ii) the Fourier magnitude and phase are measured over a partial aperture, and (iii) the Fourier magnitude is measured over a partial aperture and no phase information is measured. The iterative Fourier transform is used to reconstruct the missing data for all these cases.

Case (i) corresponds to the situation in which the Fourier magnitude is known over an entire filled aperture. Here, the image reconstruction problem is equivalent to reconstructing the Fourier phase over the aperture and the problem is that of *phase retrieval*. The iterative transform algorithm is robust in this case. Examples of such reconstructions will not be given here since they can be found in References [3,4,5,6], including the case where large amounts of noise are present [6]. Case (ii) corresponds to the situation where there are no aberrations, but the complex Fourier data is incomplete due to missing frequency bands. The reconstruction of an image from such data requires that the Fourier magnitude and phase be reconstructed within the missing frequency bands to obtain an estimate of a filled aperture plane. Hence the problem is equivalent to that of *interpolation*. The iterative algorithm is used to interpolate the missing spatial frequency bands. One could also consider *extrapolating* the Fourier domain data out to higher spatial frequencies; however this problem is known to be very ill-posed and it is not considered here. Case (iii) corresponds to the situation in which there is no phase information at all and the Fourier magnitude is known only over a partial aperture. Here the image reconstruction problem requires both phase retrieval and interpolation. This case is perhaps the most realistic setting, in which aberrated measurements are taken through a telescope with a partially-filled aperture. Unfortunately, out of the three cases investigated it also poses the most difficult reconstruction problem.

In the following discussion, reconstruction examples for cases (ii) and (iii) will be described. In each of these cases, the iterative Fourier transform algorithm was applied, each iteration consisting of the following four steps, as illustrated in Figure 2:

1. The current image estimate is Fourier transformed to produce an estimate of the object's Fourier transform over the entire Fourier domain.

2. Fourier-domain constraints corresponding to the measured data are satisfied by: (a) replacing the magnitude and phase of the current estimate with the measured magnitude and phase within the region of the aperture plane corresponding to the telescope aperture

function [note that in cases (i) and (iii) the phase is not measured and only the magnitude is replaced], (b) leaving the Fourier transform unaltered over the missing frequency bands within the filled aperture, and (c) setting the Fourier transform to zero outside the filled aperture.

3. The result is inverse Fourier transformed.

4. The object-domain constraints of positivity and object support are satisfied using one of two methods: Error Reduction (ER), which is a Gerchberg-type algorithm [7] or Hybrid Input/Output (HIO) [3,4,5].

The object-domain support constraint is determined from the measured data in one of two ways. If the phase is known over part of the Fourier domain, then one can form a degraded image from the partial Fourier magnitude and phase data. An initial support constraint can then be formed by thresholding the magnitude of the degraded image. To minimize the ringing effects due to the partial fill of the aperture, it is necessary to first apply a weighting function to the Fourier magnitude data. If there is no measured phase, then an object-domain support is determined from the Fourier magnitude as follows. The magnitude is squared and inverse Fourier transformed to obtain the autocorrelation of the object. Again, weighting of the Fourier-domain squared magnitude may be necessary to avoid excessive ringing in the autocorrelation. The autocorrelation is then thresholded to obtain an estimate of the autocorrelation support. An initial estimate of the object support is then obtained from the autocorrelation support by using a triple-intersection rule [8,9]. For future reference, the object support estimate determined according to this rule will be called the *triple-intersection support*. It is important to note that in both cases the object support estimates described above rely on thresholded values and thus may exclude parts of the actual object. Hence as the iterations progress, the support constraint is enlarged by including neighboring pixels, thus ensuring that the whole object is eventually contained within the support constraint.

2. Case (ii), Partial Fourier Magnitude and Phase

Figures 3 and 4 show examples of the iterative transform algorithm applied to the problem of interpolation. The measured data was assumed to consist of the Fourier magnitude and phase over a partial aperture. Figure 3(b) shows the simulation of measurements over aperture A, for which the Fourier data over a central disk 1/3 the diameter of the filled aperture was blocked. Hence the ratio of the area of the blocked region to the entire filled aperture was 1/9. Figure 4(b) shows the data corresponding to aperture H, for which Fourier data was only collected over seven small circular subapertures. The ratio of the area where there was

no Fourier information to the area of a filled aperture encompassing aperture H was 1/2. The images corresponding to the data collected in apertures top A and H are shown in Figures 3(e) and 4(e) respectively. These images were used as the initial object estimates for the iterative transform algorithm.

For the case of aperture A, Figure 3(g) shows the initial support constraint, which was a thresholded version of the degraded image shown in Figure 3(e). Enlarged support constraints which were used as the iterations progressed are shown in Figures 3(h) and 3(i). With the support constraint of Figure 3(i) in place, the algorithm converged quite quickly to a solution consistent with the support constraint and the measured Fourier data, yet it did not converge to the true solution. The resulting reconstructed image, shown in Figure 3(f), still has some distortion; nevertheless it appears to be better than the initial estimate, shown in Figure 3(e). Similar reconstruction results were obtained for case of aperture H and are shown in Figure 4.

An examination of the Fourier magnitude of the reconstructed image, shown in Figure 3(c), indicates that part of the problem in the reconstruction may be that the magnitude in the interpolated region of the Fourier plane is underestimated. Figure 5 shows a plot of cuts through the filled-aperture Fourier magnitude and the interpolated Fourier magnitude. Over the blocked central region, the peaks of the estimated Fourier magnitude appear to be in the right place but they are smaller and show less contrast than the true Fourier magnitude.

Thus, for the case of interpolation only, the algorithm converged quickly, but the reconstructed image was of mediocre quality. The fast convergence is due to the fact that the constraints in each domain form a convex set. The ER algorithm for this case is a projection onto convex sets (POCS) algorithm. POCS algorithms are known to have strong convergence properties [10]. However, the poor quality of the reconstructed images, despite the absence of noise in the measurements, can be an indication that the interpolation problem is ill-posed.

3. Case (iii), Partial Fourier Magnitude and No Phase

Figure 6 shows an example of the iterative transform algorithm applied to the problem of simultaneous phase retrieval and interpolation. In this case, an aberrated aperture-plane measurement was simulated for a centrally-blocked aperture in which the central obscuration was a circle with 1/8th the diameter of the filled aperture. The phase was assumed to be too corrupted to be useful, so that the only input data to the algorithm was the Fourier magnitude over a partial aperture having an annular shape.

Figure 6(a) shows the original object used in the simulation. For reference, Figure 6(b) shows the image corresponding to error-free magnitude and phase measurements over the centrally-obscured aperture. This image was assumed to be unavailable since the Fourier phase is unknown. The initial triple-intersection object support constraint computed from the given Fourier magnitude is shown in Figure 6(d). Enlarged versions of the support constraint are shown in Figures 6(e) and 6(f). The initial estimate for the object was obtained by filling the support shown in Figure 6(d) with uniformly distributed random numbers. A partially-reconstructed image was obtained from the partial Fourier magnitude data using the support constraints shown in Figures 6(d-f). The algorithm was then rerun using a different sequence of random numbers, yielding a second partially-reconstructed image. Two more partially-reconstructed images were obtained similarly, using a second initial support constraint. This second support constraint was generated by applying a triple-intersection rule to an autocorrelation support computed with a different threshold value. The four partially-reconstructed images then were combined to form a composite image by using the stripe-removal methods described in Reference [5]. The resulting reconstructed image, shown in Figure 6(c), still has some stripe artifacts but is otherwise a faithful representation of the true object. The experiment was repeated with much larger central obscurations but the quality of the resulting reconstructed images was significantly degraded.

4. Conclusions

In practical optical systems, the measurements made in aperture-plane amplitude interferometry can have missing spatial frequency bands. Moreover, the phase of these measurements can be corrupted by atmospheric turbulence or aberrations present in the optical system. The reconstruction of an extended object from these measurements thus involves the interpolation of the missing frequency bands and the retrieval of the missing or aberrated phase. In this paper we demonstrated that the iterative transform algorithm can be used for *phase retrieval* or *interpolation* or both simultaneously.

It was found that, for the phase-retrieval problem of reconstructing an image from filled-aperture magnitude and no phase, the algorithm converges reasonably quickly to the correct solution. For the interpolation problem it was found that the algorithm converged quickly to a solution, but that the solution is not necessarily close to the original object, indicating that the problem of interpolation is not a well-posed problem. The most realistic problem is the case where the magnitude is measured over a partial aperture and the phase is not available at all. In this case, the problem is that of simultaneous phase retrieval and interpolation. For the case where the missing Fourier magnitude covered a region about the origin with 1/64th the area of

the filled aperture, a good reconstruction was obtained using the iterative transform algorithm augmented by the stripe-removal methods of [5]. Thus it is possible to combine phase retrieval and interpolation in the reconstruction of an image from partial Fourier magnitude information if the interpolation is confined to a small region of the aperture plane.

Acknowledgements

This research was supported by the Office of Naval Research.

References

[1] D. G. Currie, S. L. Knapp, and K. M. Liewer, "Four Stellar-Diameter Measurements by a New Technique: Amplitude Interferometry," Astrophys. J. **187**, 131-144 (1974).

[2] F. Roddier, C. Roddier, and J. DeMarcq, "A Rotation Shearing Interferometer with Phase-Compensated Roof-Prisms," J. Optics (Paris) **9**, 145-149 (1978); J. B. Breckinridge,"Obtaining Information through the Atmosphere at the Diffraction Limit of a Large Aperture," J. Opt. Soc. Am. **65**, 755-759 (1975).

[3] J. R. Fienup, "Reconstruction of an Object from the Modulus of Its Fourier Transform," Opt. Lett. **3**, 27-29 (1978).

[4] J. R. Fienup, "Reconstruction and Synthesis Applications of an Iterative Algorithm," in *Transformations in Optical Signal Processing*, W. T. Rhodes, J. R. Fienup, and B. E. A. Saleh, eds., Proc. SPIE **373**, 147-160 (1981).

[5] J. R. Fienup and C. C. Wackerman, "Phase Retrieval Stagnation Problems and Solutions," J. Opt. Soc. Am. A **3**, 1897-1907 (1986).

[6] G. B. Feldkamp and J. R. Fienup, "Noise Properties of Images Reconstructed from Fourier Modulus," in *1980 International Optical Computing Conference*, W. T. Rhodes, ed., Proc. SPIE **231**, 84-93 (1980).

[7] G. W. Gerchberg, "Super-Resolution through Error Energy Reduction," Optica Acta **21**, 709-720 (1974).

[8] J. R. Fienup, "Image Reconstruction from Fourier Modulus Samples," Proceedings of the ESO-NOAO Workshop on High Angular Resolution Imaging from the Ground Using Interferometric Techniques (National Optical Astronomy Observatories), pp. 67-70, Oracle, Arizona, 12-15 January 1987.

[9] T. R. Crimmins, J. R. Fienup, and B. J. Thelen, "Improved Object Support Reconstruction from Autocorrelation Support and Application to Phase Retrieval," in preparation.

[10] D. C. Youla, "Generalized Image Restoration by Method of Alternating Orthogonal Projections," IEEE Trans. Circ. Sys. **CAS-25**, 694-702 (1978).

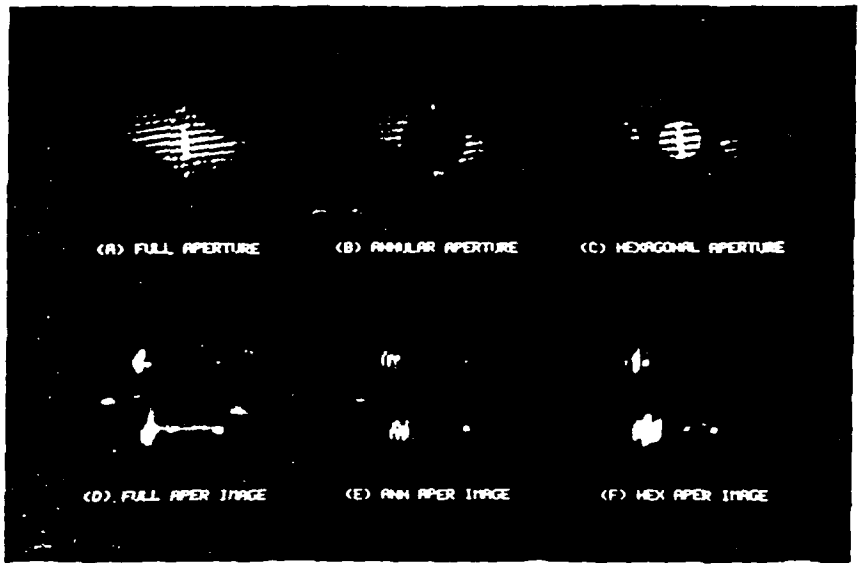


Figure 1. Aperture-plane measurements and corresponding images: (a) filled-aperture Fourier magnitude. (b) Fourier magnitude over aperture A. (c) Fourier magnitude over aperture H. (d) filled-aperture image. (e) aperture A image. (f) aperture H image.

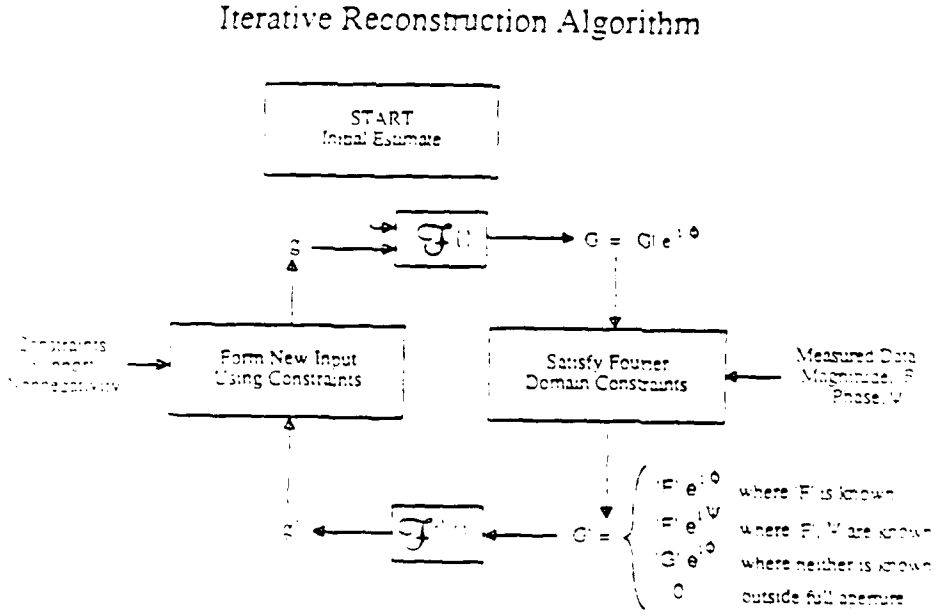


Figure 2. Block diagram of the iterative transform algorithm.

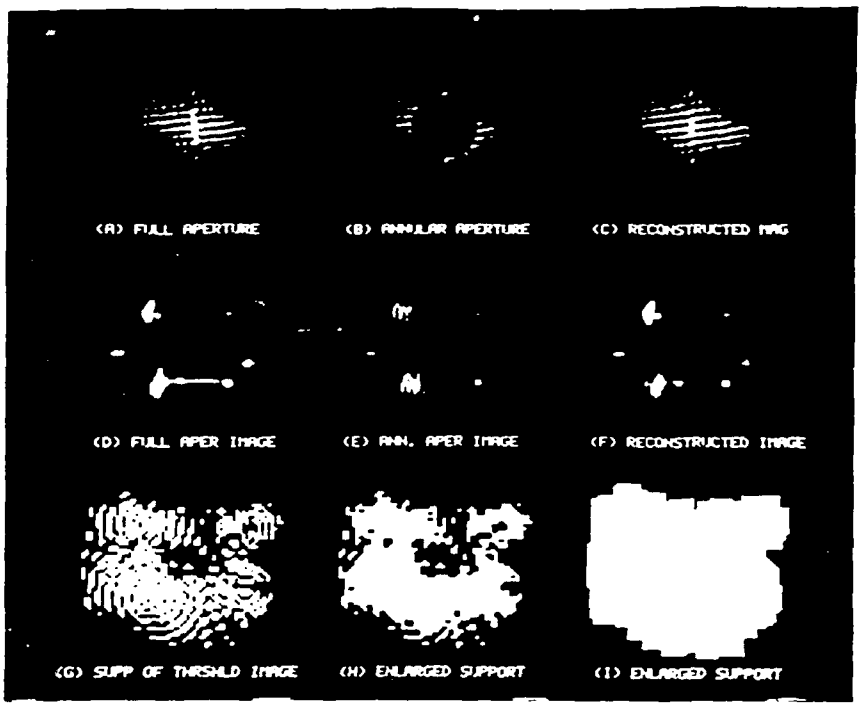


Figure 3. Interpolation from Fourier magnitude and phase over aperture A: (a) filled-aperture Fourier magnitude. (b) Fourier magnitude over aperture A, (c) Fourier magnitude of reconstructed image. (d) filled-aperture image, (e) aperture A image, (f) reconstructed image. (g) support formed from thresholding aperture A image, (h) enlarged support constraint. (i) further-enlarged support constraint.

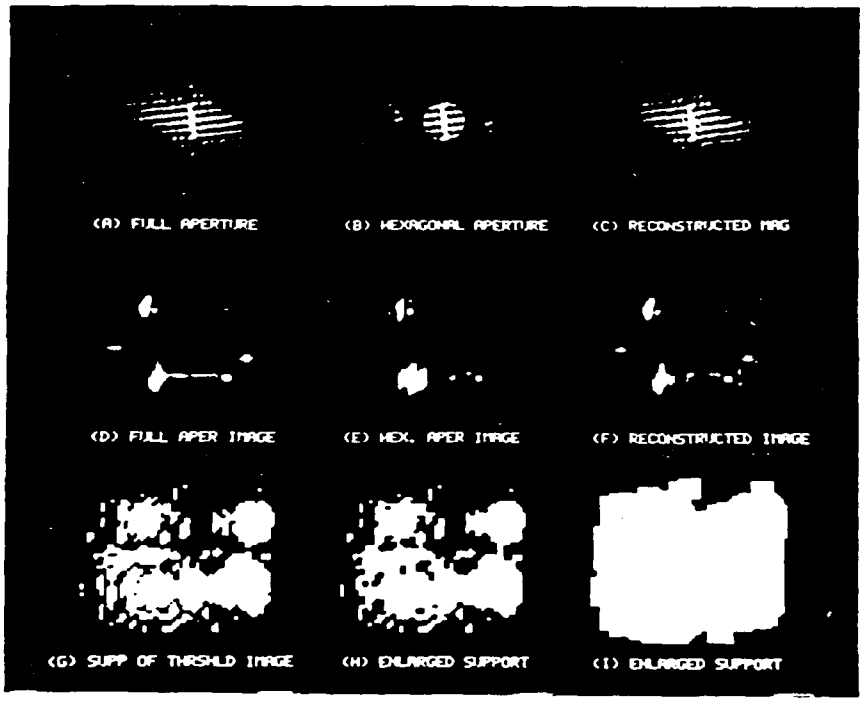


Figure 4. Interpolation from Fourier magnitude and phase over aperture H. (See caption to Figure 3)

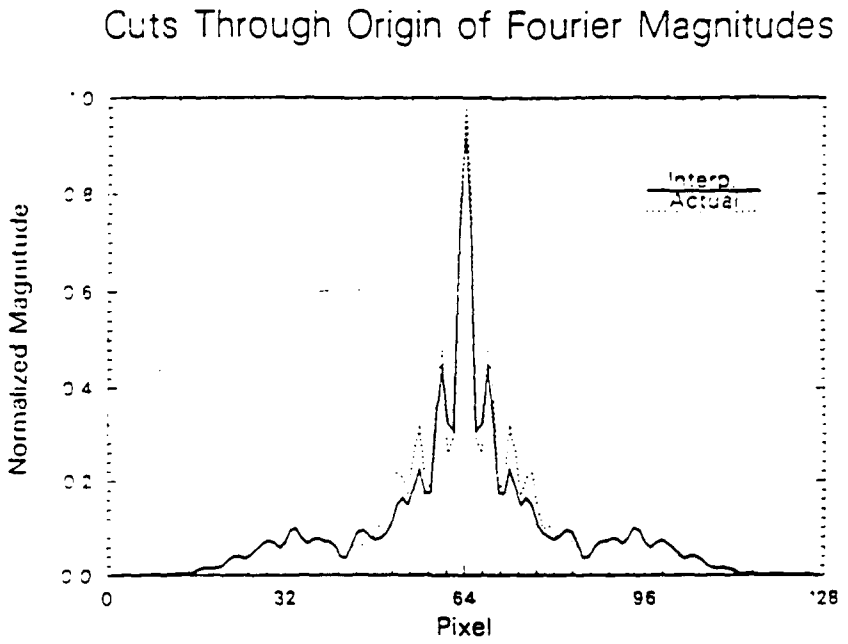


Figure 5. Cuts through the filled-aperture Fourier magnitude (dotted line) and Fourier magnitude of the reconstructed image of Figure 3 (solid line).

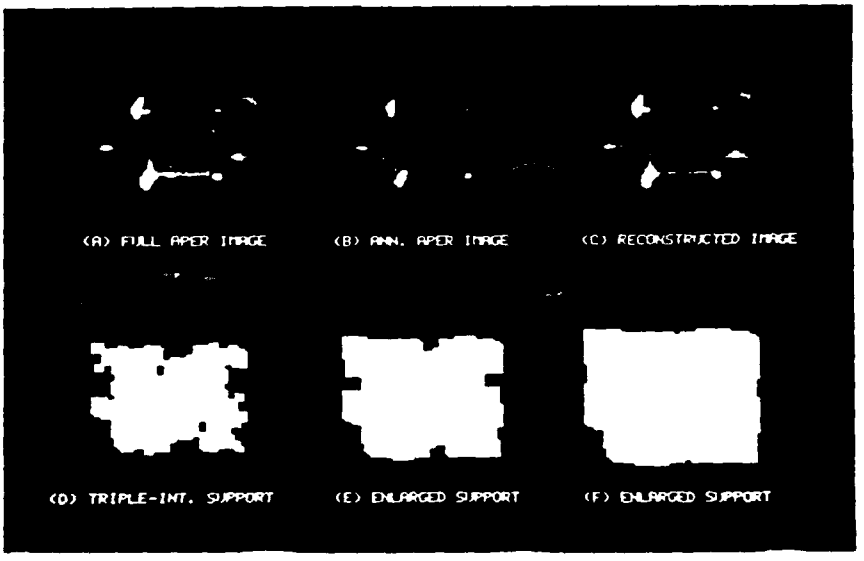


Figure 6. Interpolation and phase retrieval from partial-aperture Fourier magnitude: (a) filled-aperture image, (b) partial-aperture image with correct Fourier phase. (c) reconstructed image, (d) triple-intersection support constraint, (e) enlarged support constraint, (f) further-enlarged support constraint.

APPENDIX E

Iterative blind deconvolution algorithm applied to phase retrieval

J. H. Seldin and J. R. Fienup

Optical Science Laboratory, Advanced Concepts Division, Environmental Research Institute of Michigan, P.O. Box 8618, Ann Arbor, Michigan 48107

Received July 29, 1989; accepted October 24, 1989

The iterative blind deconvolution algorithm proposed by Ayers and Dainty [Opt. Lett. 13, 547 (1988)] and improved on by Davey *et al.* [Opt. Commun. 69, 353 (1989)] is applied to the problem of phase retrieval, which is a special case of the blind deconvolution problem. A close relationship between this algorithm and the error-reduction version of the iterative Fourier-transform phase-retrieval algorithm is shown analytically. The performance of the blind deconvolution algorithm is compared with the error-reduction and hybrid input–output versions of the iterative Fourier-transform algorithm by reconstruction experiments on real-valued, nonnegative images with and without noise.

1. INTRODUCTION

Blind deconvolution is the problem of finding two unknown functions, $f(\bar{x})$ and $g(\bar{x})$, from a noisy measurement, $c(\bar{x})$, of the convolution of these functions, defined as

$$c(\bar{x}) = \int_{-\infty}^{\infty} f(\bar{x}')g(\bar{x} - x')\mathrm{d}x' + n(\bar{x})$$

$$= f(\bar{x}) * g(\bar{x}) + n(\bar{x}), \tag{1}$$

or in the Fourier domain as

$$C(\bar{u}) = F(\bar{u})G(\bar{u}) + N(\bar{u}), \tag{2}$$

where C, F, G, and N are the Fourier transforms of c, f, g, and n, respectively. Ayers and Dainty[1] recently proposed a practical, two-dimensional blind deconvolution algorithm for the noise-free case, where the additive noise term $n(\bar{x}) = 0$.

In this paper we apply the Ayers–Dainty (AD) algorithm to the phase-retrieval problem, in which we desire to recover an image, $f(\bar{x})$, from the modulus, $|F(\bar{u})|$, of its Fourier transform:

$$F(\bar{u}) = |F(\bar{u})| \exp[i\psi(\bar{u})] = \mathcal{F}[f(\bar{x})]$$

$$= \int_{-\infty}^{\infty} f(\bar{x})\exp[-i2\pi(\bar{u} \cdot x)]\mathrm{d}x. \tag{3}$$

Phase retrieval is equivalent to the reconstruction of the Fourier phase, $\psi(\bar{u})$, from the Fourier modulus and to the reconstruction of $f(\bar{x})$ or $\psi(\bar{u})$ from the autocorrelation function:

$$r(\bar{x}) = \int_{-\infty}^{\infty} f(\bar{x}')f^*(\bar{x}' - x)\mathrm{d}x'$$

$$= \mathcal{F}^{-1}[F(\bar{u})F^*(\bar{u})] = \mathcal{F}^{-1}[|F(\bar{u})|^2]. \tag{4}$$

The phase-retrieval problem arises in several disciplines including optical and radio astronomy, wave-front sensing, holography, and remote sensing.

Comparing Eqs. (1) [with $n(\bar{x}) = 0$] and (4), we find that phase retrieval can be considered a special case of blind deconvolution, in which we deconvolve $f(\bar{x})$ and $f^*(-\bar{x})$ from $r(\bar{x})$. Because the AD algorithm represents a new, practical algorithm for blind deconvolution, we will apply it to phase retrieval and compare it with two existing phase-retrieval algorithms. We will begin by describing the AD algorithm and adaptations of the algorithm appropriate for phase retrieval. Because its structure closely resembles that of the error-reduction (ER) algorithm commonly used for phase retrieval,[2–4] the AD algorithm is compared both analytically and experimentally with ER. The performance of both of these algorithms is compared with the faster hybrid input–output (HIO) algorithm[2–4] for real, nonnegative objects for the cases of known and unknown support, using Fourier intensity data with different levels of additive Gaussian noise.

2. DESCRIPTION OF THE ALGORITHM

A. Blind Deconvolution

The AD blind deconvolution algorithm[1] (Fig. 1) alternates between the object domain and the Fourier domain, enforcing known constraints in each domain. Object-domain constraints such as support and nonnegativity are combined with the Fourier-domain constraint of Eq. (2) to produce new estimates of f and g, $\tilde{f}_k$ and $\tilde{g}_k$, respectively, at each iteration. Note that each AD loop produces two estimates of F (and G): (1) $\tilde{F}_k$, the Fourier transform of $\tilde{f}_k$, and (2) the estimate obtained by imposing the Fourier-domain constraint of Eq. (2). These two estimates are averaged by using the scalar β $(0 < \beta < 1)$ to form F_k, a composite estimate of F. Ayers and Dainty proposed the following estimate of F from $\tilde{F}_k$ and $\tilde{G}_k$, the Fourier transform of $\tilde{g}_k$:

if $|C(\bar{u})| <$ noise level,

$$F_k(\bar{u}) = \tilde{F}_k(\bar{u}); \tag{5a}$$

0740-3232/90/030428-06$02.00

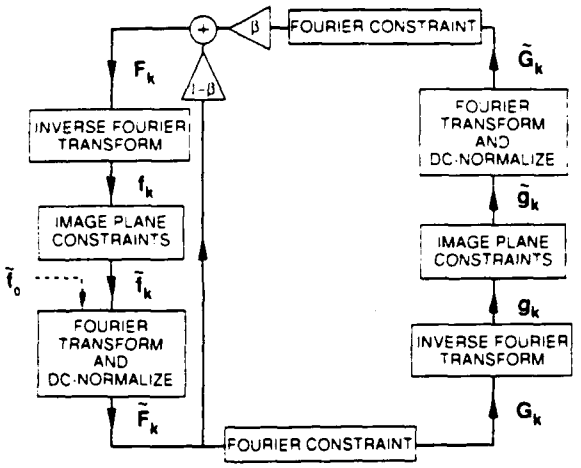


Fig. 1. AD blind deconvolution algorithm.

if $|G_k(\tilde{u})| > |C(\tilde{u})|$,

$$F_k(\tilde{u}) = (1-\beta)\tilde{F}_k(\tilde{u}) + \beta\frac{C(\tilde{u})}{\tilde{G}_k(\tilde{u})}; \tag{5b}$$

if $|G_k(\tilde{u})| < |C(\tilde{u})|$,

$$\frac{1}{F_k(\tilde{u})} = \frac{1-\beta}{\tilde{F}_k(\tilde{u})} + \beta\frac{\tilde{G}_k(\tilde{u})}{C(\tilde{u})}. \tag{5c}$$

Rather than implementing Eqs. (5), we use a Wiener-type filter based on the following imaging model:

$$c(\tilde{x}) = s(\tilde{x}) * f(\tilde{x}) + n(\tilde{x}), \tag{6}$$

or in the Fourier domain

$$C(\tilde{u}) = S(\tilde{u})F(\tilde{u}) + N(\tilde{u}), \tag{7}$$

where c is the measured image, f is the object, s is the impulse response [the Fourier transform of which is $S(\tilde{u})$, the optical transfer function], and n is the noise. Assuming that f and n are independent, zero-mean, Gaussian random processes, the minimum mean-squared-error linear estimator for $f(\tilde{x})$ is[5] $\hat{f}(\tilde{x}) = \mathcal{F}^{-1}[\hat{F}(\tilde{u})]$, where

$$\hat{F}(\tilde{u}) = W(\tilde{u})C(\tilde{u}), \tag{8}$$

the Wiener–Helstrom filter is

$$W(\tilde{u}) = \frac{S^*(\tilde{u})}{|S(\tilde{u})|^2 + \langle|N(\tilde{u})|^2\rangle/\langle|F(\tilde{u})|^2\rangle}, \tag{9}$$

and $\langle|N(\tilde{u})|^2\rangle$ and $\langle|F(\tilde{u})|^2\rangle$ are the ensemble-averaged energy spectra of the noise and the object, respectively. Although the images generally will not satisfy the statistical assumptions stated above, the filter is still effective and simple to implement. The Wiener–Helstrom filter of Eq. (9) is often used for image restoration.

To apply Eq. (9) to the problem of estimating F from C and $\tilde{G}$, we relate Eq. (2) to Eq. (7) [and, hence, Eq. (1) to Eq. (6)] by allowing $G(\tilde{u})$ to play the role of $S(\tilde{u})$. The resulting Fourier-domain constraint (with $\beta = 1$) is

$$F_k(\tilde{u}) = \frac{\tilde{G}_k^*(\tilde{u})}{|\tilde{G}_k(\tilde{u})|^2 + \sigma^2/|\tilde{F}_k(\tilde{u})|^2}C(\tilde{u}), \tag{10}$$

where $\tilde{G}_k$ is the latest estimate of G, the constant σ^2 is an estimate of $\langle|N|^2\rangle$, and $|\tilde{F}_k|^2$ is used to estimate $\langle|F|^2\rangle$. A filter similar to this was used with the AD algorithm by Davey *et al.*[6] for the blind deconvolution of noisy, complex-valued images. We have approximated $\langle|N|^2\rangle$ with a constant based on the assumption that $n(\tilde{x})$ is a delta-correlated, Gaussian random process. If the ensemble-averaged energy spectrum of the noise is known, it should replace σ^2 in Eq. (10).

To estimate G from C and $\tilde{F}_k$, the latest estimate of F, in Eq. (10) we replace F_k with G_k, $\tilde{G}_k$ with $\tilde{F}_k$, and, following the indexing of Fig. 1, $\tilde{F}_k$ with $\tilde{G}_{k-1}$:

$$G_k(\tilde{u}) = \frac{\tilde{F}_k^*(\tilde{u})}{|\tilde{F}_k(\tilde{u})|^2 + \sigma^2/|\tilde{G}_{k-1}(\tilde{u})|^2}C(\tilde{u}). \tag{11}$$

We have also used an even simpler Wiener-type filter, formed by replacing the term $\sigma^2/|\tilde{F}_k|^2$ in the denominator of Eq. (10) with a constant, α:

$$F_k(\tilde{u}) = \frac{\tilde{G}_k^*(\tilde{u})}{|\tilde{G}_k(\tilde{u})|^2 + \alpha}C(\tilde{u}). \tag{12}$$

We will refer to this simpler filter as AD Filter 1, and the filter in Eq. (10) as AD Filter 2. We make the same substitutions that are made for Eq. (10) to obtain the following expression for $G_k(\tilde{u})$ from Eq. (12):

$$G_k(\tilde{u}) = \frac{\tilde{F}_k^*(\tilde{u})}{|\tilde{F}_k(\tilde{u})|^2 + \alpha}C(\tilde{u}). \tag{13}$$

B. Phase Retrieval

As we noted in Section 1, phase retrieval can be viewed as the process of blindly deconvolving a function $f(\tilde{x})$ and its twin, $f^*(-\tilde{x})$. Thus for phase retrieval the noisy measurements of $r(\tilde{x})$ and $|F(\tilde{u})|^2$ take on the roles of $c(\tilde{x})$ and $C(\tilde{u})$, respectively, and $F_k(\tilde{u})$ and $G_k(\tilde{u})$ become estimates of $F(\tilde{u})$ and $F^*(\tilde{u})$, respectively. Because the two convolution factors are twins, the AD algorithm actually produces two estimates of f per iteration. Therefore we need only consider half of the AD loop (Fig. 2); i.e., instead of estimating $F^*(\tilde{u})$ and $f^*(-\tilde{x})$ we forego the second half of the loop and find a new estimate of $F(\tilde{u})$ by conjugating $G_k(\tilde{u})$, the estimate of $F^*(\tilde{u})$. Replacing C with $|F|^2$, we conjugate Eq. (13) to obtain the AD Filter 1 phase-retrieval Fourier-domain constraint:

$$F_k(\tilde{u}) = \tilde{G}_k^*(\tilde{u}) = \frac{\tilde{F}_k(\tilde{u})}{|\tilde{F}_k(\tilde{u})|^2 + \alpha}|F(\tilde{u})|^2. \tag{14}$$

AD Filter 2 is modified in a similar manner by conjugating Eq. (11) and substituting $|\tilde{F}_k|^2$ for $|\tilde{G}_{k-1}|^2$:

$$F_k(\tilde{u}) = \frac{\tilde{F}_k(\tilde{u})}{|\tilde{F}_k(\tilde{u})|^2 + \sigma^2/|\tilde{F}_k(\tilde{u})|^2}|F(\tilde{u})|^2. \tag{15}$$

Note that for photon (shot) noise in the measurement of $C(\tilde{u})$, which would have a variance proportional to the mean of $|F|^2$, the quantity $\sigma^2/|\tilde{F}(\tilde{u})|^2$ is equivalent to α in Eq. (14).

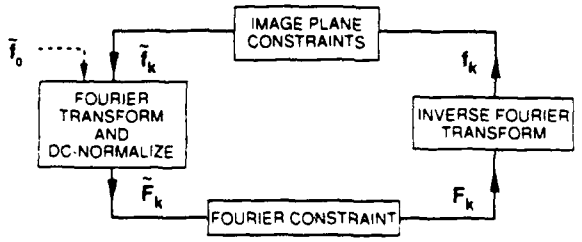


Fig. 2. AD blind deconvolution algorithm applied to phase retrieval.

C. Comparison with Error Reduction

The flow chart in Fig. 2 of the AD algorithm applied to phase retrieval is identical in form to the ER algorithm. The difference between the ER algorithm and the AD algorithm lies with the Fourier-domain constraint. In the ER algorithm the Fourier-domain constraint is imposed by substituting the known modulus, $|F(\bar{u})|$, for $|\tilde{F}_k(\bar{u})|$, the modulus of the Fourier transform of $\tilde{f}_k(\bar{x})$, the estimate of the object. If we write $\tilde{F}_k(\bar{u}) = |\tilde{F}_k(\bar{u})| \exp[i\Phi_k(\bar{u})]$, then the Fourier-domain step in the ER algorithm gives

$$F_k(\bar{u}) = |F(\bar{u})|\exp[i\Phi_k(\bar{u})] = \tilde{F}_k(\bar{u}) \frac{|F(\bar{u})|}{|\tilde{F}_k(\bar{u})|}. \tag{16}$$

If for simplicity we assume that we are using an inverse filter [which corresponds to the noise-free case and is obtained by setting $\alpha = 0$ in Eq. (14) or $\sigma = 0$ in Eq. (15)], then the AD Fourier-domain constraint can be written as

$$F_k(\bar{u}) = \tilde{F}_k(\bar{u}) \frac{|F(\bar{u})|^2}{|\tilde{F}_k(\bar{u})|^2}. \tag{17}$$

Comparison of Eqs. (16) and (17) shows that, for the noise-free case, the Fourier-domain constraint of the AD algorithm is similar to that of the ER algorithm: they both produce estimates with the same phase, and the magnitudes of both estimates are boosted (or attenuated) where $|F|/|\tilde{F}_k| > 1$ (or < 1). Because the object-domain operations are identical and the Fourier-domain constraints are so similar, we expect the AD and ER algorithms to behave similarly.

3. EXPERIMENTAL SIMULATIONS

The two versions of the AD algorithm (AD Filters 1 and 2) were compared experimentally with each other, with ER, and with a combination of HIO and ER (HIO/ER) for two cases: (1) a real-valued, nonnegative object with *a priori* known triangular support of side 128 pixels embedded in a 256 × 256 array and (2) a real-valued, nonnegative object with unknown support (approximately 40 × 60 pixels) in a 128 × 128 array. The triangular support in case (1) was chosen to allow for rapid convergence even for the slower algorithms.[7] For case (1) we also added Gaussian noise to the Fourier intensity data. The reconstructions for case (2) are more difficult because the support is unknown and because it is of a less-favorable shape.[7] For each case, the same initial guess is used to begin all the algorithms.

A useful error metric for measuring the success of the reconstruction is the normalized root-mean-squared (NRMS) error with the original object. This error metric takes advantage of the fact that, in a simulation like this, we know the original object, $f(\bar{x})$. Recalling that the estimate of $f(\bar{x})$ after the kth iteration is $\tilde{f}_k(\bar{x})$, we define the NRMS error,

$$\text{ABSERR} \equiv \left[\frac{\sum_{\bar{x}} |\alpha \tilde{f}_k(\bar{x} - \bar{x}_0) - f(\bar{x})|^2}{\sum_{\bar{x}} |f(\bar{x})|^2} \right]^{1/2}, \tag{18}$$

where x_0 maximizes the cross correlation between f and $\tilde{f}_k$ and

Fig. 3. Comparison of phase-retrieval using AD blind deconvolution with the HIO and ER iterative transform algorithms for a real-valued, nonnegative object with known support and no Fourier modulus error. Reconstructed images: (A) HIO/ER (indistinguishable from the original object); (B) ER; (C) AD with the Fourier constraint of Eq. (14); (D) AD with the Fourier constraint of Eq. (15).

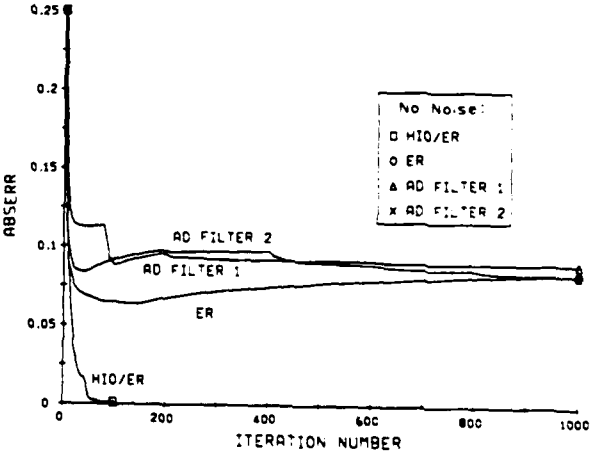


Fig. 4. ABSERR versus iteration number for the reconstructions of Fig. 3.

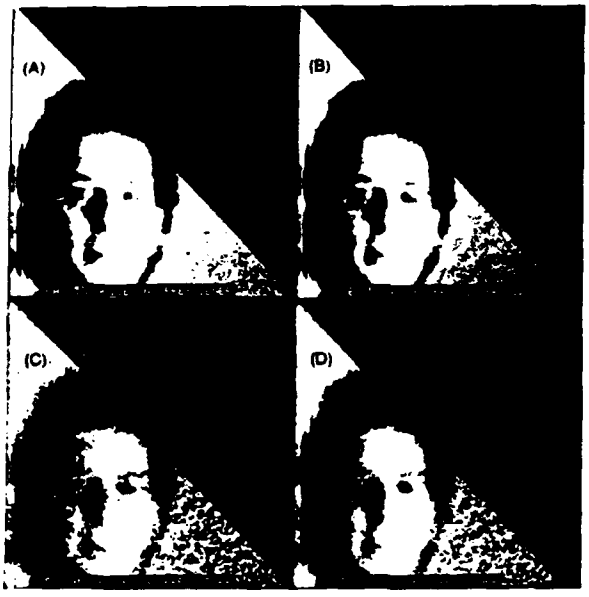


Fig. 5. Comparison of the effect of the pre-Wiener filtering of noisy Fourier intensity data on reconstructions with the ER algorithm. Reconstructed images after 1000 iterations: (A) 5% FME, no pre-Wiener filtering; (B) 5% FME, pre-Wiener filtering; (C) 20% FME, no pre-Wiener filtering; (D) 20% FME, pre-Wiener filtering.

Fig. 6. Comparison of phase retrieval using AD, HIO, and ER for a real-valued, nonnegative object with known support and 5% FME. Reconstructed images: (A) HIO/ER, (B) ER, (C) AD with the Fourier constraint of Eq. (14), (D) AD with the Fourier constraint of Eq. (15).

$$\alpha = \frac{\sum_{\underline{x}} f(\underline{x}) f_k^*(\underline{x} - \underline{x}_0)}{\sum_{\underline{x}} |f_k(\underline{x})|^2} \tag{19}$$

is a scalar that can be shown to minimize ABSERR.

The reconstructions for case (1) with noise-free Fourier intensity data are shown in Fig. 3 [AD Filter 1 corresponds to Eq. (14), and AD Filter 2 to Eq. (15)]. The ER and AD images exhibit similar striping artifacts, which are frequently seen in iterative reconstruction.[4] Methods developed for eliminating the stripes[4] were not attempted here. The HIO/ER image avoids this stagnation effect and converges more quickly to a solution indistinguishable from the original object. Figure 4 is a plot of ABSERR versus iteration number for the reconstructions of Fig. 3. The AD and ER algorithms stagnated after approximately 50 iterations, while HIO/ER converged to the solution in fewer than 100 iterations. Because we used filter parameters α and σ^2 that were

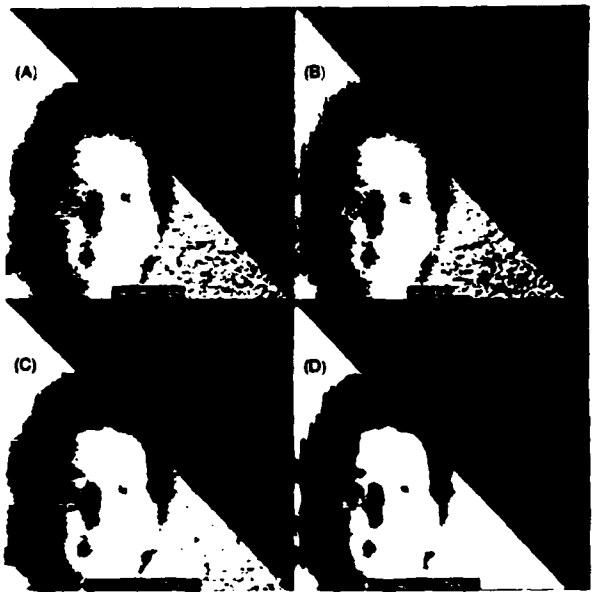


Fig. 7. Comparison of phase retrieval using AD, HIO, and ER for a real-valued, nonnegative object with known support and 20% FME. Reconstructed images: (A) HIO/ER, (B) ER, (C) AD with the Fourier constraint of Eq. (14), (D) AD with the Fourier constraint of Eq. (15).

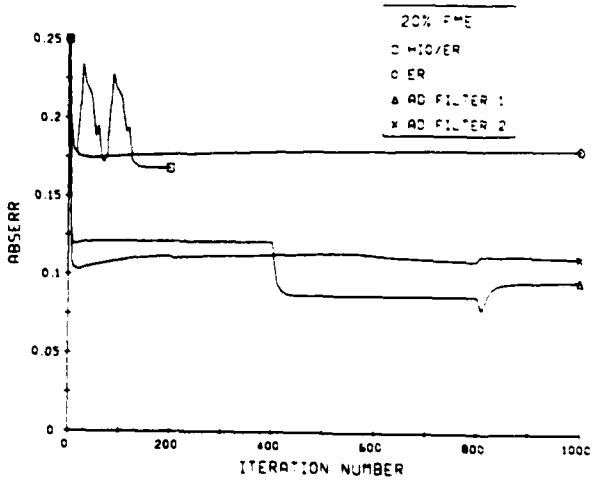


Fig. 8. ABSERR versus iteration number for the reconstructions of Fig. 7.

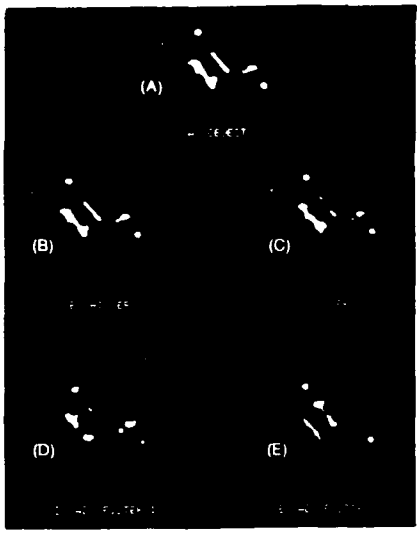


Fig. 9. Comparison of phase retrieval using AD, HIO, and ER for a real-valued, nonnegative object with unknown support and no FME. (A) Object. Reconstructed images: (B) HIO/ER, (C) ER, (D) AD with the Fourier constraint of Eq. (14), (E) AD with the Fourier constraint of Eq. (15).

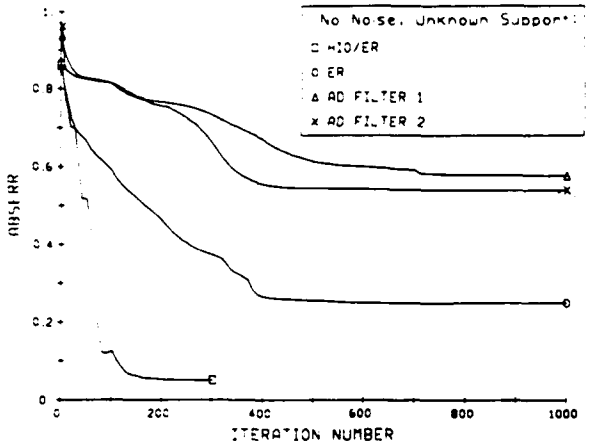


Fig. 10. ABSERR versus iteration number for the reconstructions of Fig. 9.

small (to account for computer roundoff error) for the noiseless case, there is little difference between the two AD filters, and the corresponding reconstructions are almost identical. We expect the differences between the filters to become more apparent for the case of noisy Fourier intensity data.

We now consider the same image with Gaussian noise added to the Fourier intensity. When the noisy Fourier intensity is denoted by $|F(\bar{u})|_n^2$, the Fourier-modulus error (FME) with respect to the original Fourier intensity, $|F(\bar{u})|^2$, is

$$\text{FME} \equiv \left\{ \frac{\sum_{\bar{u}} [|F(\bar{u})|_n - |F(\bar{u})|]^2}{\sum_{\bar{u}} |F(\bar{u})|^2} \right\}^{1/2}. \tag{20}$$

We performed reconstructions for single realizations of $|F|_n^2$ with 5% and 20% FME. Because the AD algorithm has a Wiener-type filter built into it, a less-prejudiced comparison between algorithms is obtained if we filter the noisy Fourier intensity before use with the ER and HIO algorithms. The pre-Wiener-filtered modulus that is used in this case is

$$|\hat{F}(\bar{u})| = \left[\frac{1}{1 + \sigma^2/|F(\bar{u})|_n^4} |F(\bar{u})|_n^2 \right]^{1/2}, \tag{21}$$

where σ^2 is the variance of the noise added to the Fourier intensity. Figure 5 demonstrates the effect of Eq. (21) on ER reconstructions for the two noisy cases. The smoothing of the pre-Wiener filter has a negligible effect for the 5% FME data but is more significant for the 20% FME data.

The reconstructions from all four algorithms for the case of 5% FME are shown in Fig. 6. Since the pre-Wiener filtering of Eq. (21) was insignificant at the 5% FME noise level, it was not used in these HIO and ER reconstructions. The 5% level of noise has little effect on visual image quality, and the performance of the algorithms relative to one another is similar to that for the noiseless case. Reconstructions with 20% FME are shown in Fig. 7. This level of noise significantly degrades the visual image quality, and the pre-Wiener filtering was implemented for the HIO and ER reconstructions. The AD Filter 1 image of Fig. 7(C) has no striping artifacts and is comparable in quality with the HIO/ER reconstruction of Fig. 7(A), whereas AD Filter 2 stagnates with stripes after starting with the same initial guess. The low-pass nature of the Wiener-type filter has a smoothing effect that is evident in the AD reconstructions. The amount of smoothing depends on the filter parameters α and σ^2: the larger these parameter are, the larger the attenuation of high frequencies and the smoother the reconstruction. In this case the two AD reconstructions achieve a smaller ABSERR than either ER or HIO/ER (Fig. 8) but at the expense of image sharpness. The reconstructions stagnate almost immediately, but a change in α after 400 iterations moves the AD Filter 1 image out of stripe stagnation. The ability to vary the built-in Wiener-type filter parameters may be an advantage of the AD algorithm. The AD algorithm also may be making better use of the Wiener filter, and a few iterations of AD Filter 1 on the HIO/ER image of Fig. 7(A) yields an image that is similar to th[illegible] Fig. 7(C).

Figure 9 shows the reconstructions from all four algorithms for case (2), a real-valued, nonnegative image with unknown support in a 128 × 128 array. The support was estimated from the support of the autocorrelation, $r(\bar{x})$, using a triple-intersection algorithm.[8] Figure 10 is a plot of ABSERR versus iteration number for the reconstructions of Fig. 9. The HIO/ER algorithm converged close to the solution in fewer than 200 iterations, whereas AD and ER both converged more slowly and stagnated after approximately 400 iterations. The error of the ER reconstruction is significantly lower than that of the AD algorithms. For this more-difficult case, we find again that the AD and ER algorithms perform comparably (ER somewhat better than AD), and HIO/ER is still more effective than either.

4. CONCLUSION

We have shown that the Ayers–Dainty (AD) blind deconvolution algorithm applied to phase retrieval is similar to the error-reduction (ER) iterative Fourier-transform algorithm, both in form and in performance. A nice feature of the AD

algorithm is a built-in Wiener-type filter, which seems to perform slightly better than the pre-Wiener filter used with hybrid input–output (HIO) and ER for the noisier case. The two different Wiener-type filters considered here performed comparably, and the significant difference between them is that Filter 1 [Eq. (14)] is simpler to implement than Filter 2 [Eq. (15)]. For the more difficult case of reconstructing an object with unknown support, the AD algorithm was not quite so effective as ER and did not converge close to a solution as did the combination of HIO and ER (HIO/ER). HIO/ER is still the most effective reconstruction algorithm at low noise levels, and at higher levels of noise the AD algorithm can be used in conjunction with HIO to improve the quality of the reconstruction.

ACKNOWLEDGMENTS

This research was supported by the U.S. Office of Naval Research under contract N00014-86-C-0587.

Portions of this paper were presented at the Optical Society of America Topical Meeting on Signal Recovery and Synthesis III, North Falmouth, Massachusetts, June 14–16, 1989.[9]

REFERENCES

1. G. R. Ayers and J. C. Dainty, "An iterative blind deconvolution method and its applications," Opt. Lett. **13**, 547–549 (1988).
2. J. R. Fienup, "Reconstruction of an object from the modulus of its Fourier transform," Opt. Lett. **3**, 27–29 (1978).
3. J. R. Fienup, "Phase retrieval algorithms: a comparison," Appl. Opt. **21**, 2758–2769 (1982).
4. J. R. Fienup and C. C. Wackerman, "Phase-retrieval stagnation problems and solutions," J. Opt. Soc. Am. A **3**, 1897–1907 (1986).
5. C. W. Helstrom, "Image restoration by the method of least squares," J. Opt. Soc. Am. **57**, 297–303 (1967).
6. B. L. K. Davey, R. G. Lane, and R. H. T. Bates, "Blind deconvolution of noisy complex-valued image," Opt. Commun. **69**, 353–356 (1989). In Eq. (15) of that paper, by our logic, the term $\alpha/|F_{i-1}(u)|^2$ in the denominator should be $\alpha/|H_{i-1}(u)|^2$.
7. J. R. Fienup, "Reconstruction of a complex-valued object from the modulus of its Fourier transform using a support constraint," J. Opt. Soc. Am. A **4**, 118–123 (1987).
8. T. R. Crimmins, J. R. Fienup, and B. J. Thelen, "Improved bounds on object support from autocorrelation support and application to phase retrieval," J. Opt. Soc. Am. A **7**, 1–13 (1990).
9. J. H. Seldin and J. R. Fienup, "Phase retrieval using Ayers/Dainty deconvolution," in *Digest of Topical Meeting on Signal Recovery and Synthesis III* (Optical Society of America, Washington, D.C., 1989), pp. 124–127.

APPENDIX F

Numerical investigation of the uniqueness of phase retrieval

J. H. Seldin and J. R. Fienup

Optical Science Laboratory, Advanced Concepts Division, Environmental Research Institute of Michigan, P O Box 8618, Ann Arbor, Michigan 48107

Received July 29, 1989; accepted October 24, 1989

Both a new iterative grid-search technique and the iterative Fourier-transform algorithm are used to illuminate the relationships among the ambiguous images nearest a given object, error metric minima, and stagnation points of phase-retrieval algorithms. Analytic expressions for the subspace of ambiguous solutions to the phase-retrieval problem are derived for 2×2 and 3×2 objects. Monte Carlo digital experiments using a reduced-gradient search of these subspaces are used to estimate the probability that the worst-case nearest ambiguous image to a given object has a Fourier modulus error of less than a prescribed amount. Probability distributions for nearest ambiguities are estimated for different object-domain constraints.

1. INTRODUCTION

The phase-retrieval problem considered in this paper is the reconstruction of an object function $f(x, y)$ from the modulus $|F(u, v)|$ of its Fourier transform:

$$F(u, v) = |F(u, v)|\exp[i\psi(u, v)] = \mathcal{F}[f(x, y)] = \iint f(x, y)\exp[-i2\pi(ux + vy)]\,dx\,dy. \quad (1)$$

It is equivalent to the reconstruction of the Fourier phase $\psi(u, v)$ from the Fourier modulus and to the reconstruction of $f(x, y)$ or $\psi(u, v)$ from the autocorrelation function

$$r(x, y) = \mathcal{F}^{-1}|F(u, v)|^2. \quad (2)$$

This problem arises in several disciplines, including optical and radio astronomy, wave-front sensing, holography, and remote sensing.

There are the omnipresent ambiguities: that the object $f(x, y)$, any translation of the object $f(x - x_0, y - y_0)$, the twin image $f^*(-x - x_0, -y - y_0)$, and any of these multiplied by a constant of unit magnitude $\exp(i\phi_c)$ all have exactly the same Fourier modulus. These ambiguities change only the object's position or orientation, not its appearance. If they are the only ambiguities, then we refer to the object as being unique. A solution is considered to be ambiguous only if it differs from the object in ways other than these omnipresent ambiguities.

If nothing is known about the object, then reconstruction from its Fourier modulus is generally ambiguous except for special cases. Fortunately, for many applications one has additional *a priori* knowledge about or constraints on the object. In the astronomy application, for example, the object's spatial brightness distribution, $f(x, y)$, is a real, nonnegative function. For several applications, one has a support constraint, i.e., the object is known to be zero outside some finite area. Even if the support constraint is not known *a priori*, upper bounds can be placed on the support of the object since it can be no larger than half the diameter of the autocorrelation along any direction. Additional measurements or other forms of *a priori* information may be available for specific applications; in this paper we consider real-valued objects with known support, both with and without a nonnegativity constraint.

Until the late 1970's, there was much doubt that the phase-retrieval problem could be solved or that the solution would be useful, because the one-dimensional theory of analytic functions available at the time indicated that there were ordinarily a huge number of ambiguous solutions.[1-3]

The first indications that the two-dimensional (2-D) case is usually unique, despite the lack of uniqueness in one dimension, came from empirical reconstruction results[4,5]: images that were reconstructed resembled the original simulated objects used to compute the Fourier modulus data. These results gave hope that 2-D phase-retrieval problems might be solvable and unique. (Other phase-retrieval problems, such as in electron microscopy in which one has squared-modulus measurements in each of two domains[6] and in x-ray crystallography in which one has the *a priori* information that the object consists of a finite collection of atoms,[7] had been solved; but those earlier successes depended on much greater object-domain constraints than just nonnegativity and support.) Those empirical results gave impetus to attempts to extend the one-dimensional (1-D) theory to two dimensions. Although progress has been made,[8-13] the level of understanding of the 2-D problem has not yet matched that of the 1-D problem.

One of the most enlightening developments has been the work of Bruck and Sodin,[14] who modeled the object distribution as an array of delta functions on a regular grid. Then the continuous Fourier transform becomes the discrete Fourier transform (DFT),

$$F(u, v) = |F(u, v)|\exp[i\psi(u, v)] = \mathrm{DFT}[f(x, y)] = \sum_{x=0}^{M-1}\sum_{y=0}^{N-1} f(x, y)\exp\left[-j2\pi\left(\frac{ux}{2M} + \frac{vy}{2N}\right)\right], \quad (3)$$

where the DFT is taken over a $2M \times 2N$ array but $f(x, y)$ is zero outside an $M \times N$ array in order to avoid aliasing in the computation of $r(x, y)$ and $|F(u, v)|^2$. For this discrete case the Fourier transform given in Eq. (3) can then be expressed

0740-3232/90/030412-16$02.00

as a polynomial of two complex variables, $z = \exp(j\pi u/M)$ and $w = \exp(j\pi v/N)$. It is also equivalent to the z transform. Then the presence of ambiguity in the phase-retrieval problem is equivalent to the factorability of the polynomial. This explains the vast difference between the 1-D and 2-D cases, because polynomials (of degree 2 or greater) of a single complex variable are always factorable, whereas polynomials of two (or more) complex variables are rarely factorable.[14-16] Other interesting results have been obtained by exploiting this discrete model. Fiddy *et al.*[17] and Nieto-Vesperinas and Dainty[18] described an object support that, by virtue of Eisenstein's irreducibility theorem, guarantees uniqueness. Brames[19] showed that any discrete object having a support whose convex hull has no parallel sides is unique among objects with supports having the same convex hull; so if the convex hull of the support of such an object is known *a priori*, then it is unique. For these cases, there also exists a closed-form recursive reconstruction algorithm.[20,21]

Whether the objects are discrete or continuous, it is easy to make up cases that are ambiguous. If $g(x, y)$ and $h(x, y)$ are two functions of finite support with Fourier transforms $G(u, v)$ and $H(u, v)$ respectively, then the convolutions

$$f_1(x, y) = g(x, y) * h(x, y) \tag{4}$$

and

$$f_2(x, y) = g(x, y) * h^*(-x, -y) \tag{5}$$

are different objects as long as neither g nor h is conjugate centrosymmetric, they have Fourier transforms

$$F_1(u, v) = G(u, v)H(u, v) \tag{6}$$

and

$$F_2(u, v) = G(u, v)H^*(u, v) \tag{7}$$

that have the same modulus,

$$|F_1(u, v)| = |F_2(u, v)| = |G(u, v)||H(u, v)|, \tag{8}$$

and the objects f_1 and f_2 are ambiguous. This demonstrates the equivalence of phase-retrieval ambiguity to convolutions in the object domain [Eqs. (4) and (5)] and factorability in the Fourier domain [Eqs. (6) and (7)]. Furthermore, if there are K irreducible Fourier factors, then there are 2^{K-1} ambiguous solutions. By this convolutional (products or factors in the Fourier domain) method, it is possible to make up an uncountably infinite number of ambiguous cases even though the theory indicates that ambiguity is rare (of zero probability) in two dimensions. Consider that it is also true that any randomly chosen real number has probability zero of being a rational number (almost all are irrational numbers). Yet any real number, even if irrational, can be approximated arbitrarily well by a rational number. Thus the fact that the probability of any given object's being ambiguous (the Fourier transform being factorable) is zero is not necessarily comforting.

Sanz *et al.* have shown that the "uniqueness condition is stable in the sense that it is not sensitive to noise."[22] However, their analysis does not shed light on a more practical definition of uniqueness. If a given nonfactorable polynomial is near enough (in an integrated mean-squared difference sense) to a factorable polynomial, then the ambiguous solutions associated with the factorable polynomial will be consistent (to within the noise) with the noisy Fourier-modulus data. Under this circumstance the object may be considered to be ambiguous in a practical sense, even though it may be unique, traditionally speaking. Up to this point it was not known how close an arbitrary polynomial is, on the average, to a factorable polynomial. Furthermore, the existence of ambiguous objects close to a given object is likely to cause the existence of local minima in which iterative reconstruction algorithms will become trapped. Current theory has not adequately addressed these questions, even for the discrete model. These questions can be answered, though, by numerical means, as will be seen below.

One way to test for practical uniqueness is the use of the iterative Fourier-transform algorithm.[4,23-25] If multiple solutions exist, then the algorithm tends to find all of them if many reconstructions are performed, each starting from a different array of random numbers as the initial estimate.[26] In most instances investigated, when the algorithm is applied to the Fourier modulus of an object of interest, if it does not stagnate[25] it reconstructs essentially the correct object,[27] giving strong evidence of uniqueness for those types of object. Furthermore, when noise is added to the Fourier-modulus data, the result is usually a noisy image of the object rather than a completely different reconstruction,[28,29] contrary to some predictions.[30] While this approach has provided some assurance that the phase-retrieval problem is usually unique in the practical sense even in the presence of noise, it has not yielded any quantitative results on the probability of uniqueness for any given level of noise.

An important consideration in the probability of uniqueness is the set of constraints placed on the object. In all cases we assume that the object has finite support (it is zero outside some finite region). The support of the object plays a crucial role. If the object has a delta function known to satisfy the holography condition,[31] then it is unique. As mentioned above, discrete objects having certain supports are guaranteed to be unique.[17,19] In addition, objects having separated parts are more likely to be unique.[32] Although it is less well understood, nonnegativity also plays an important role in uniqueness.

In this paper we establish a methodology for determining the probability of phase-retrieval uniqueness in the practical sense. We have developed a method, suitable for small images, for answering the questions: Given an arbitrary object and its Fourier polynomial, how close is the nearest factorable polynomial, and does it have an ambiguous solution that is significantly different from the given object? In this paper we explore this question for the case of objects defined within 2×2 and 3×2 supports. A derivation of object-domain conditions for factorability provides a means for finding nearest factorable polynomials through a constrained-minimization search over the space of 2×2 or 3×2 ambiguous images. These searches are implemented with different object-domain constraints in a Monte Carlo simulation to estimate the probability that the nearest factorable polynomial, with an ambiguous solution that is significantly different from a given object, is within some distance of the given polynomial. Before describing these main results, we first define the pertinent error metrics and discuss some preliminary results of a grid-search method for finding local minima in phase retrieval, and relationships among minima, ambiguities, and phase-retrieval stagnation.

2. OBJECT-TO-FREQUENCY-DOMAIN MAPPINGS AND ERROR METRICS

A useful means for visualizing the ambiguity problem is through a mapping between the space of objects (images) and the space of Fourier moduli as illustrated in Fig. 1. In Fig. 1 each domain is a finite-dimensional space in which any one point represents a 2-D function. In this diagram $|F(u, v)|$ represents Fourier-modulus data for a unique object and $|G_a(u, v)|$ modulus data for an ambiguous object, since both g_a and g_{ac} map into it. We refer to g_a and g_{ac} as ambiguous counterparts of each other, gotten by conjugating one or more of the Fourier-domain polynomial factors. For the case depicted in Fig. 1, as indicated by the distances between the points, two widely different images, f and g_{ac}, may have similar, but not identical, Fourier moduli. Thus, although f is unique, one might unknowingly reconstruct g_{ac} by a phase-retrieval algorithm given a noisy measurement of $|F|$.

The following error metrics provide a means for quantifying differences in both domains. These metrics are the focus of the numerical approach presented in this paper. (Other related error metrics are also useful.) Given two real-valued functions $g(x, y)$ and $f(x, y)$ defined on an $M \times N$ support and zero padded to a $2M \times 2N$ array, we define the Fourier-modulus error, the error (distance) between $|F(u, v)|$ and $|G(u, v)|$, as

$$\epsilon(g, f) \equiv \left[\frac{\sum_{u,v} [\alpha_f |G(u, v)| - |F(u, v)|]^2}{\sum_{u,v} |F(u, v)|^2} \right]^{1/2}, \tag{9}$$

where

$$\alpha_f = \left[\frac{\sum_{u,v} |F(u, v)|^2}{\sum_{u,v} |G(u, v)|^2} \right]^{1/2} \tag{10}$$

is an energy normalization factor, $G(u, v) = \mathrm{DFT}\{g(x, y)\}$, and u and v summations are taken over the intervals 0, 1, . . . , $2M - 1$ and 0, 1, . . . , $2N - 1$, respectively.

A similar metric defines the object-domain error between $f(x, y)$ and $g(x, y)$:

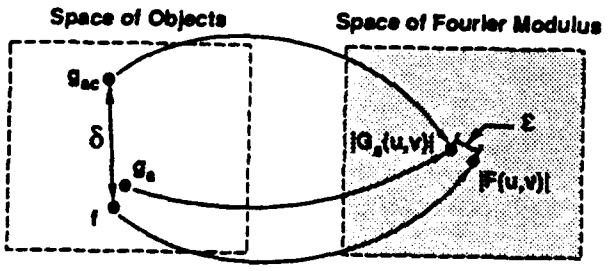


Fig. 1. Object-space to Fourier-modulus-space mappings of a unique object f and a pair of ambiguous images (g_a, g_{ac}), with error metrics δ and ϵ.

$$\delta(g, f) \equiv \left[\frac{\sum_{x,y} [\alpha_0 g(x, y) - f(x, y)]^2}{\sum_{x,y} f^2(x, y)} \right]^{1/2}, \tag{11}$$

where

$$\alpha_0 = \alpha_f \operatorname{sign}\left[\sum_{x,y} f(x, y) g(x, y) \right] \tag{12}$$

and x and y summations are taken over 0, 1, . . . , $M - 1$ and 0, 1, . . . , $N - 1$, respectively. The parameter α_0 takes into account any differences in scaling and polarity between g and f. Translations are ignored here because the support constraint automatically rules them out. Because $g(x, y)$ and its twin, $g(M - 1 - x, N - 1 - y)$, share the same Fourier modulus, we compute $\delta(g, f)$ for both $g(x, y)$ and its twin and use the smaller of the two values of δ. Of particular interest from the point of view of phase retrieval are images that have a small Fourier-modulus error ϵ, but a large object-domain error δ, since these images may be ambiguous in the practical sense.

3. GRID SEARCHES

Our first approach to understanding the relationship between ϵ and δ, for a collection of images g relative to a given object f, was by a grid search. What we mean by a grid search is illustrated as follows for the case of 3×2 ($M = 3, N = 2$) objects. Given a 3×2 object f, we calculate ϵ and δ for all 3×2 images $g = g_{\mathrm{ref}} + g_{\mathrm{inc}}$, where g_{ref} is another 3×2 real-valued image and

$$g_{\mathrm{inc}} = \begin{bmatrix} s_1 & s_2 & s_3 \\ s_4 & s_5 & s_6 \end{bmatrix}, \tag{13}$$

where, given a real-valued increment Δs, each s_i can assume values in the set $\{k\Delta s; k = -L, -L + 1, \ldots, 0, 1, \ldots, L\}$. If we think of both f and g as points in a six-dimensional (6-D) space, then we are calculating ϵ and δ for all g's sampled on a symmetric 6-D grid of step size Δs centered about the point g_{ref}, with the grid width equal to $2L + 1$ steps in each of the six dimensions.

This search can become quite extensive as the grid width increases. Since the number of different g_{inc}'s (grid points) is $(2L + 1)^6$, even a five-step search ($L = 2$) requires 15,625 calculations of ϵ and δ. If the search uses the zero image for g_{ref}, we can cut down on redundant calculations of ϵ by eliminating twin images and images with polarity [sign of $F(0, 0)$] opposite f. Note that the saving is in the calculation of ϵ, which is computationally more expensive than the calculation of δ.

Grid-Search Example

The use of a successively finer grid search to find minima in ϵ (which could constitute a phase-retrieval algorithm) and shed light on the properties of ϵ and δ is illustrated in the following example. An integer-valued image f was chosen:

$$f = \begin{bmatrix} 1 & 2 & -1 \\ 2 & 1 & -2 \end{bmatrix}. \tag{14}$$

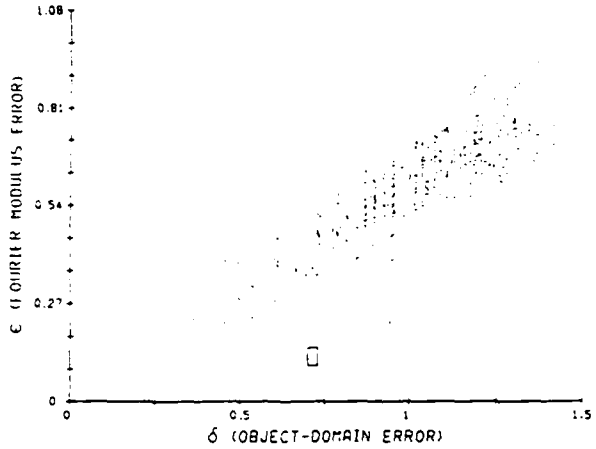


Fig. 2. Fourier-modulus error ϵ versus object-domain error δ for a five-step grid search with step size $\Delta s = 1$. The minimum value of ϵ (excluding $g = f$) is boxed.

A five-step search ($L = 2$) was implemented with g_{ref} equal to the zero function and with $\Delta s = 1$; i.e., the pixel values of g are taken from the set $\{-2, -1, 0, 1, 2\}$. Since the search is centered about the zero function, the twin and polarity search-reduction techniques mentioned above were implemented. The results are displayed in Fig. 2 in the form of a scatter plot of ϵ versus δ. Several features of the scatter plot are noted:

(1) ϵ is less than or equal to δ. The proof of this fact is given in Appendix A.

(2) The vertical striping reflects the discrete nature of the search, i.e., the elements of g take on only integer values.

(3) ϵ and δ can both be greater than unity, despite the normalization that takes place in the denominators of Eqs. (9) and (11).

(4) The scatter plot exhibits a banded type of structure, i.e., the points tend to cluster in a region where both δ and ϵ are large. This is not surprising, since we expect most images that are quite different in the object domain to be quite different in the Fourier-modulus domain as well.

The single point of greatest interest, an outlier with large δ and relatively small ϵ, is outlined by a box in Fig. 2. It corresponds to the image

$$g_0 = \begin{bmatrix} 1 & 1 & -2 \\ 1 & -1 & -2 \end{bmatrix}, \tag{15}$$

with $\delta(g_0, f) = 0.714$ and $\epsilon(g_0, f) = 0.124$. It is the point within the grid search with the lowest value of ϵ aside from $g = f$. Since it represents the point on the grid search closest to being a serious ambiguity, we explored it further by performing another five-step search, with $g_{ref} = g_0$ of Eq. (15) and a step size of $\Delta s = 1/3$. Because g_{ref} is not the zero function, no data reduction was implemented, and ϵ and δ were calculated for the 15,625 different grid points. Figure 3 shows the scatter plot for this second search for $\epsilon < 0.125$. It is apparent that our initial search with unit steps was quite coarse and that, compared with g_0, there are images with significantly smaller values of ϵ and comparably large values of δ. The minimum value of ϵ for this grid search corresponds to the image

$$g_1 = \begin{bmatrix} \frac{2}{3} & \frac{2}{3} & -2 \\ 1 & \frac{-2}{3} & \frac{-7}{3} \end{bmatrix}, \tag{16}$$

with $\delta(g_1, f) = 0.704$ and $\epsilon(g_1, f) = 0.0648$.

We performed a third five-step search, with $g_{ref} = g_1$ and $\Delta s = 1/9$. The image corresponding to the minimum ϵ for this search is

$$g_2 = \begin{bmatrix} \frac{2}{3} & \frac{7}{9} & \frac{-17}{9} \\ \frac{10}{9} & \frac{-2}{3} & \frac{-22}{9} \end{bmatrix}, \tag{17}$$

with $\delta(g_2, f) = 0.666$ and $\epsilon(g_2, f) = 0.0569$.

Iterative Grid Searches

The iterative searching above is an approach for finding minima of ϵ. It is summarized more generally by the following steps for the case of $M \times N = 3 \times 2$.

(1) Initialize. Choose g_{ref}, the number of search steps $(2L + 1)$, the step size (Δs), and a step-size reduction factor (r).

(2) Perform a 6-D $(2L + 1)$-step search with $g = g_{ref} + g_{inc}$, where

$$g_{inc} = \begin{bmatrix} s_1 & s_2 & s_3 \\ s_4 & s_5 & s_6 \end{bmatrix} \tag{18}$$

and each s_j, $j = 1, 2, \ldots, 6$, is from the set $\{k\Delta s; k = -L, -L + 1, \ldots, 0, 1, \ldots, L\}$.

(3) Set g_{ref} equal to the image, g, which has the minimum value of ϵ found in the search of the previous step.

(4) Set Δs equal to $\Delta s/r$.

(5) Stop if the stopping criterion is met; otherwise go to step 2.

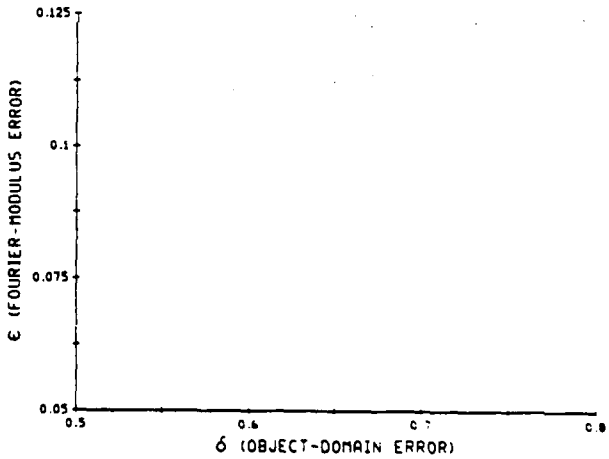


Fig. 3. Fourier-modulus error ϵ versus object-domain error δ for a five-step grid search with $\Delta s = 1/3$ about the minimum of the grid search of Fig. 2. All points satisfying $\epsilon < 0.125$ are shown here.

The stopping criterion is based on the percentage change in the minimum value of ϵ from iteration to iteration, or set for a maximum number of iterations, whichever is satisfied first. For a large value of L, the search time is prohibitive, but the sampling is finer. Also, the initial step-size and step-reduction factor must be chosen carefully, since the step size at the kth iteration is $\Delta s/(r^{k-1})$. If r is chosen too large, the grid may shrink too quickly to progress to a minimum. If Δs is too small, the minimum might not be found because it lies outside the initial grid. The most reliable search uses a slowly shrinking grid with a large number of grid points (large L) that samples the space over a large region. The more finely we sample the space, the more computationally burdensome the algorithm becomes, yet a coarser grid would leave doubt about the reliability of our minimum.

This iterative search could constitute a phase-retrieval algorithm. However, it would be a computationally inefficient algorithm, requiring many thousands of DFT's to converge to a solution for the case of larger objects. Here we are using it only to find a local minimum (the global minimum is at $g = f$ for which $\epsilon = \delta = 0$).

The iterative grid search was tested for f given by Eq. (14) and with the following three sets of parameters: (1) $L = 1$, $\Delta s = 1/2$, $r = 2$; (2) $L = 2$, $\Delta s = 1/3$, $r = 3$; and (3) $L = 3$, $\Delta s = 1/4$, $r = 4$. Each iterative search started with $g_{\text{ref}} = g_0$ given by Eq. (15), corresponding to the minimum ϵ found in the first search described above. Each of these searches found a scalar multiple of the same image, g_{min}, given by

$$g_{\text{min}} = \begin{bmatrix} 0.623 & 0.749 & -1.871 \\ 1.149 & -0.659 & -2.530 \end{bmatrix}, \tag{19}$$

with $\delta(g_{\text{min}}, f) = 0.667$ and $\epsilon(g_{\text{min}}, f) = 0.0558$. This probably represents a deep local minimum for the phase-retrieval problem and could represent a practical ambiguity if the noise in the Fourier modulus data were to exceed $\epsilon(g_{\text{min}}, f)$.

4. MINIMA AND PHASE RETRIEVAL

The minimum in ϵ, represented by g_{min} found in the iterative grid searches described above, represents two potential problems for phase retrieval. First, a relatively small error in the modulus data (5.58%) could cause the data to be consistent with g_{min}, which, if reconstructed, would have a very large object-domain error (66.7%). Second, even when it is performing phase retrieval with error-free modulus data, the algorithm could get trapped and stagnate at this local minimum. In particular, the error-reduction (ER) version of the iterative transform algorithm is equivalent to a steepest-descent gradient search method on a cost function closely related to ϵ.[24] Thus, if the local minimum found in our iterative searches were a true local minimum, the ER algorithm could stagnate at this image, unable to find a direction in which to descend. To visualize how ϵ and δ vary around g_{min}, we plot ϵ and δ along the line joining f [Eq. (14)] and g_{min} [Eq. (19)]. Figure 4 shows $\epsilon(g, f)$ and $\delta(g, f)$ versus t for

$$g = f + t(g_{\text{min}} - f). \tag{20}$$

While Fig. 4 represents only a 1-D slice through a 6-D space, it gives the appearance of a minimum in ϵ at $t = 1$ ($g = g_{\text{min}}$).

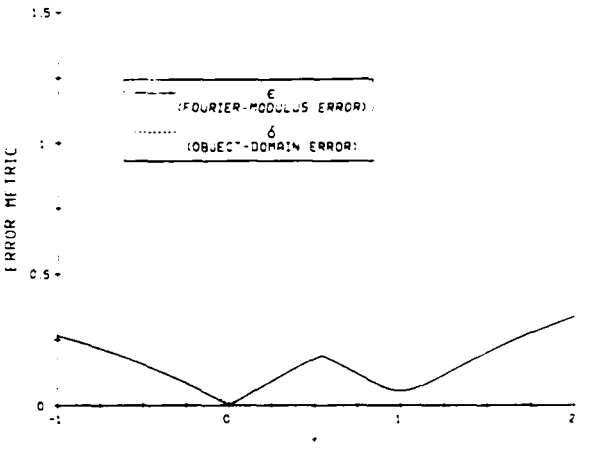


Fig. 4. $\epsilon(g, f)$ and $\delta(g, f)$ versus t for $g = f + t(g_{\text{min}} - f)$, the line joining f and g_{min}.

When ER is performed on $|F|$ with g_{min} as the initial guess, stagnation occurs immediately, giving further evidence of the presence of a local minimum.

As another test of ER's tendency to stagnate at a minimum in ϵ, we use g's corresponding to different values of t in Eq. (20) as initial guesses. These values are selected on both sides of the peak in the ϵ curve in Fig. 4. We might expect values of t chosen on the right-hand side of the peak to correspond to initial guesses that stagnate at g_{min} and guesses chosen to the left of the peak to converge to the correct solution, f. Several values of t were selected on both sides of the peak, and the predicted result was verified for all initial guesses.

The hybrid input–output (HIO) version of the iterative Fourier-transform algorithm[24] is one way of climbing out of local minima. Simulated annealing[33] is another. Cycles of HIO iterations followed by ER iterations[24] were used with a variety of starting points: g_0, g_1, g_2, and g_{min}. In each case the HIO/ER combination converged to the correct solution, f, although ER by itself stagnated in each of these same cases. As we will see below, HIO is not always sufficient to overcome stagnation.

5. MINIMA AND AMBIGUOUS IMAGES

A clue to the understanding of the stagnation point described above is its relationship to ambiguous images. Consider again the object f given by Eq. (14). Using methods that are described below, one can verify that the 3×2 ambiguous image whose Fourier modulus is closest to the Fourier modulus of the object f is

$$g_a = \begin{bmatrix} 0.594 & 1.624 & -1.211 \\ 2.330 & 1.415 & -1.730 \end{bmatrix}, \tag{21}$$

with $\delta(g_a, f) = 0.217$ and $\epsilon(g_a, f) = 0.0859$. The ambiguous counterpart to g_a [gotten by conjugating one of the factors of $G_a(u, v)$] is

$$g_{ac} = \begin{bmatrix} -0.363 & -0.618 & 1.987 \\ -1.422 & 0.600 & 2.837 \end{bmatrix}, \tag{22}$$

with $\delta(g_{ac}, f) = 0.677$ and $\epsilon(g_{ac}, f) = 0.0859$. A comparison of g_{min} [Eq. (19)] with $-g_{ac}$ (which, for our purposes, is equivalent to g_{ac}) reveals a similarity between this pair of images. The error metrics reveal their similarity in both domains: $\delta(-g_{ac}, g_{min}) = \delta(g_{ac}, g_{min}) = 0.113$ and $\epsilon(g_{ac}, g_{min}) = 0.0663$.

Because $-g_{ac}$ and g_{min} are quit similar, we might expect the ER algorithm with an initial guess of $-g_{ac}$ to stagnate at g_{min}. This is indeed the case after approximately 50 iterations. This result, coupled with the similarity between f [Eq. (14)] and its nearest ambiguity, g_a, might lead us to conclude that ER would find the correct solution if it were started with an initial guess of g_a. This is not the case, however, and the algorithm stagnates after fewer than 20 iterations at

$$g_{stag} = \begin{bmatrix} 0.694 & 1.778 & -1.010 \\ 2.235 & 1.355 & -1.856 \end{bmatrix}. \tag{23}$$

with $\delta(g_{stag}, f) = 0.152$ and $\epsilon(g_{stag}, f) = 0.0631$. This stagnation point is close to g_a, with $\delta(g_a, g_{stag}) = 0.0828$ and $\epsilon(g_a, g_{stag}) = 0.0577$. Because g_{stag} is not in the range of the iterative grid searches that found g_{min}, it was not found earlier. A plot of ϵ and δ along the line joining f and g_{stag} is shown in Fig. 5. Despite the difference in vertical scaling, the minimum in Fig. 5 does not appear to be as deep as that in Fig. 4, so one would suspect there might be a good chance of perturbing g_{stag} enough to get the algorithm out of stagnation. As with g_{min}, it was verified that the HIO is able to move out of stagnation at g_{stag} and to the solution.

Figure 6 depicts the possible relationships in both domains between f, its nearest ambiguous image and counterpart, and the two stagnation points. From the previous results we form the following conjecture: For a given object f and its Fourier modulus $|F|$, stagnation points of the iterative transform algorithm (particularly ER) tend to be near ambiguous images that have Fourier moduli close to $|F|$. This conjecture is supported more strongly by the following example.

Consider the following image f and its nearest ambiguity, g_a, with ambiguous counterpart g_{ac}:

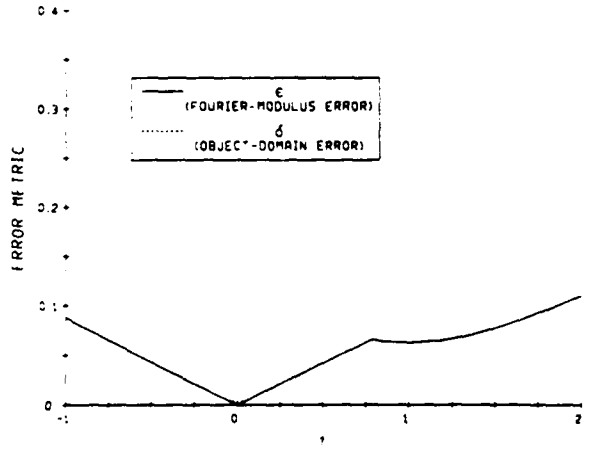


Fig. 5. $\epsilon(g, f)$ and $\delta(g, f)$ versus t for $g = f + t(g_{stag} - f)$, the line joining f and g_{stag}.

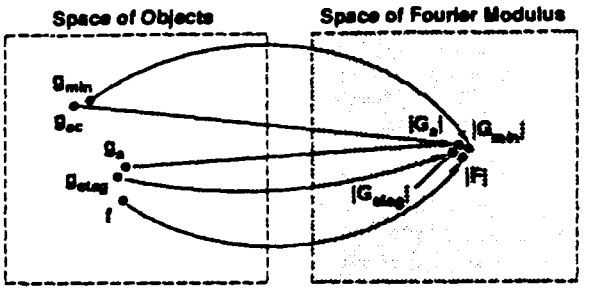


Fig. 6. Object-space to Fourier-modulus-space mappings of an object f, two stagnated images g_{min} and g_{stag}, and the nearest ambiguous image to f with respect to the Fourier-modulus error (g_a, g_{ac}).

$$f = \begin{bmatrix} 0.476 & 3.244 & 1.379 \\ 1.659 & 2.939 & 1.102 \end{bmatrix}. \tag{24}$$

$$g_a = \begin{bmatrix} 0.867 & 3.521 & 1.278 \\ 1.679 & 2.651 & 0.796 \end{bmatrix}. \tag{25}$$

$$g_{ac} = \begin{bmatrix} 0.350 & 2.146 & 3.171 \\ 0.677 & 2.475 & 1.974 \end{bmatrix}. \tag{26}$$

with $\delta(g_a, f) = 0.128$, $\delta(g_{ac}, f) = 0.502$, and $\epsilon(g_a, f) = \epsilon(g_{ac}, f) = 0.00861$. This is a case of a close ambiguity; i.e., the object, f, would be ambiguous in the practical sense unless the data, $|F|$, were low in noise. The ER algorithm was run close to 900 times on $|F|$ using a nonnegativity constraint, each time with a different random initial start. The algorithm converged to the correct solution f of Eq. (24) only 10% of the time. The algorithm stagnated near g_a approximately 9% of the time and at several images close to g_{ac} the rest of the time (81%). When a combination of HIO and ER was used with the same set of random starts, convergence to the solution f was improved to a 26% rate. 74% of the time the algorithm stagnated at one of two different minima, g_{s1} and g_{s2}, each close to the image g_{ac} in Eq. (26):

$$g_{s1} = \begin{bmatrix} 0.353 & 2.143 & 3.172 \\ 0.684 & 2.470 & 1.976 \end{bmatrix} \tag{27}$$

35% of the time, with $\delta(g_{s1}, g_{ac}) = 0.00195$ and $\epsilon(g_{s1}, g_{ac}) = 0.00144$, and

$$g_{s2} = \begin{bmatrix} 0.266 & 1.876 & 2.971 \\ 0.746 & 2.711 & 2.222 \end{bmatrix} \tag{28}$$

39% of the time, with $\delta(g_{s2}, g_{ac}) = 0.0978$ and $\epsilon(g_{s2}, g_{ac}) = 0.0115$. The images g_{s1} and g_{s2} are analogous to g_{min} in Fig. 6. While convergence to g_{s1} is bad in the sense that g_{s1} is different from the solution f [$\delta(g_{s1}, f) = 0.502$], it is still consistent with the given data [$\epsilon(g_{s1}, f) = 0.00848$] and could be considered a solution (albeit the wrong one). The stagnation at g_{s2} is even more troublesome since it is not only similarly consistent with the given data [$\epsilon(g_{s2}, f) = 0.00869$] and far from f [$\delta(g_{s2}, f) = 0.511$] but also is not so close to g_{ac} [$\delta(g_{s2}, g_{ac}) = 0.0978$].

A complete understanding of phase-retrieval stagnation points and their relationship to ambiguous images is not yet available. However, from the limited number of experiments of the type described above, we can say that stagnation points are often related to ambiguous images.

6. NEAREST AMBIGUITIES

In this section we investigate the space of ambiguous images in order to gain some insight into just how close the nearest ambiguous image is to a typical image. This may in turn have implications about how nearest ambiguities relate to stagnation points encountered in iterative phase retrieval. It also will tell us the probability of an ambiguity in the practical sense, as a function of the noise in the Fourier-modulus data.

Object-Domain Conditions for Ambiguity

As described above, ambiguous images are characterized in the Fourier domain by factorable Fourier transforms and in the object domain by being expressible as the convolution of two or more smaller images. We choose the object-domain relationship to characterize the space of ambiguous images. We begin by deriving the ambiguity condition for the smallest possible 2-D ambiguous image (2 × 2 support) and then similarly derive it for a 3 × 2 support.

2 × 2 Ambiguity Conditions

Consider the case of a real-valued image on a 2 × 2 support. It is ambiguous if it can be expressed as the convolution of two 1-D sequences:

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix} = \begin{bmatrix} e \\ f \end{bmatrix} * [g \quad h] = \begin{bmatrix} eg & eh \\ fg & fh \end{bmatrix}, \tag{29}$$

where e, f, g, and h are all nonzero (for simplicity only the nonzero rows and columns of the arrays are shown). This gives the following equations for a, b, c, and d:

$$a = eg, \tag{30a}$$
$$b = eh, \tag{30b}$$
$$c = fg, \tag{30c}$$
$$d = fh. \tag{30d}$$

Multiplying Eq. (30a) with Eq. (30d) and Eq. (30b) with Eq. (30c), we arrive at the following 2 × 2 convolution condition:

$$ad = bc. \tag{31}$$

In this case a single ambiguous counterpart to an image satisfying Eq. (31) is generated by convolving one of the 1-D sequences by the flip (rotation by 180°) of the other (equivalent to conjugating the corresponding Fourier factor). However, if $e = f$ and/or $g = h$ (i.e., one of the 1-D sequences is symmetric), then flipping the factor has no effect, and the image is still unique. Furthermore, if $e = -f$ and/or $g = -h$, then a flip of either convolution factor becomes the negative of the original factor. Since we do not consider two images that differ by a scalar multiple (-1 in this case) as ambiguous counterparts, we must also rule out this special case of negative symmetric factors. Therefore the image is unique if $|e| = |f|$ or if $|g| = |h|$. From Eqs. (30) we see that, if $|a| = |c|$ or $|b| = |d|$, then $|e| = |f|$, and if $|a| = |b|$ or $|c| = |d|$, then $|g| = |h|$. When these special cases are combined with Eq. (31), the ambiguity condition for the case of 2 × 2 support becomes

$$ad = bc, \tag{32a}$$
$$|b| \neq |a| \neq |c|. \tag{32b}$$

Note that the inequalities of relation (32b) combined with Eq. (32a) imply that $|b| \neq |d| \neq |c|$.

Equation (32a) describes a three-dimensional surface in the four-dimensional space of real-valued 2 × 2 images. While it is accepted that there is zero probability that an arbitrarily selected object will land on this surface, i.e., the phase-retrieval problem is almost always (with probability 1) unique, in this paper we are concerned with how close the Fourier modulus of a given object is likely to be to the Fourier moduli of images lying upon this surface.

3 × 2 Ambiguity Conditions

The same approach is used to formulate object-domain ambiguity conditions for 3 × 2 images. A 3 × 2 image results from convolving either (a) a 3 × 1 sequence with a 1 × 2 sequence or (b) a 2 × 1 sequence with a 2 × 2 image. Since it is known that any 1-D sequence can always be written as the convolution of smaller sequences, we can write the 3 × 1 sequence of case (a) as the convolution of two 2 × 1 sequences. We can then combine one of these factors with the 1 × 2 factor to give case (b). Thus we need only consider case (b), and our 3 × 2 image is ambiguous if

$$\begin{bmatrix} a & b & c \\ d & e & f \end{bmatrix} = [g \quad h] * \begin{bmatrix} i & j \\ k & l \end{bmatrix} = \begin{bmatrix} gi & hi + gj & hj \\ gk & hk + gl & hl \end{bmatrix}, \tag{33}$$

where g and h are nonzero and none of the pairs (i and j) or (i and k) or (j and l) or (k and l) is zero. This gives six nonlinear equations for a, b, c, d, e, and f in terms of g, h, i, j, k, and l. As is shown in Appendix B, these equations can be solved to give the following ambiguity condition:

$$(af - cd)^2 - (ae - bd)(bf - ce) = 0. \tag{34}$$

Equation (34) describes a five-dimensional surface in the 6-D space of real-valued 3 × 2 images. In comparison, for the 2 × 2 case the ambiguity surface describes a three-dimensional surface embedded within a four-dimensional space. Appendix B also shows that Eq. (34) can be solved to give, for example, b in terms of the remaining five values:

$$b = \frac{1}{2}\left[e\left(\frac{c}{f} + \frac{a}{d}\right) \pm (e^2 - 4df)^{1/2}\left(\frac{c}{f} - \frac{a}{d}\right)\right]. \tag{35}$$

An ambiguous, real-valued 3 × 2 image arising from the convolution of a 2 × 1 sequence with a nonfactorable 2 × 2 image can be shown to have an ambiguous counterpart that must also be real valued. However, if the 2 × 2 convolution factor of Eq. (33) can itself be factored, then we have the case of a 3 × 2 image resulting from the convolution of a 3 × 1 sequence with a 1 × 2 sequence. An ambiguous, real-valued image formed in this way will have rows that are scalar multiples of one another; i.e., $a = Kd$, $b = Ke$, and $c = Kf$ for some scalar K. This condition makes each difference term in Eq. (34) equal to zero. It is straightforward to show that if $b^2 < 4ac$, then this real-valued ambiguity will have a complex-valued ambiguous counterpart. If the image is constrained to be real valued, then this complex-valued image does not constitute an ambiguity within the space of real-valued images. Furthermore, because this special case is a small subset of the entire ambiguity surface, we expect it to have a relatively minor effect on the likelihood of stagnation due to nearby ambiguities.

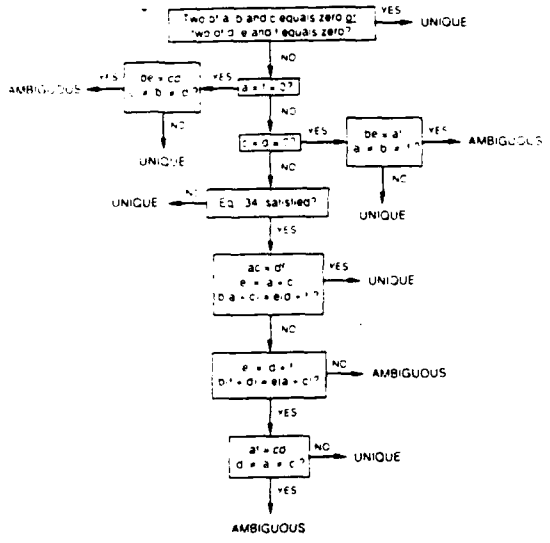


Fig. 7. Flow chart for determining the ambiguity of the 3 × 2 real-valued image of Eq. (33). Multiple conditions in a box must all be satisfied for "YES," except where "or" is specified.

The ability to factor an image into the convolution of two or more images is necessary, but not sufficient, for determining ambiguity. When we discussed the ambiguity condition for 2 × 2 images, we considered the special cases of what we called symmetric and negative-symmetric convolution factors. These special cases, as well as the effect of zero-valued pixels, also must be considered for 3 × 2 images. To save space, rather than discussing these exceptions in detail we summarize them in the ambiguity flow chart in Fig. 7.

Nearest Ambiguity by Means of Constrained Minimization

The mathematical description of ambiguities for 2 × 2 and 3 × 2 images can be used to investigate the nearness of a given object to an ambiguous image. We formulate the task of finding the ambiguous image nearest a given object as a multidimensional constrained-minimization problem. By nearest ambiguity we mean the image on the ambiguity surface for which some objective function involving the ambiguous image and the given object is minimized. For the objective function we choose

$$E(g, f) = \sum_{u,v} [|G(u, v)| - |F(u, v)|]^2, \tag{36}$$

which is just $\epsilon^2(g, f)$ of Eq. (9) with $\alpha_f = 1$ and without the normalization.

Each $M \times N$ image, having $L = M \times N$ pixel values, can be thought of as a single point in an L-dimensional vector space. To emphasize this fact we can denote an image g by the L-dimensional vector $\underline{x}$, where $\underline{x} = (a\ b\ c\ d)^t$ for the 2 × 2 case and $\underline{x} = (a\ b\ c\ d\ e\ f)^t$ for the 3 × 2 case (the ordering of the pixels in the vector $\underline{x}$ is arbitrary). Therefore, for a given image f, we desire to find $\underline{x}$ (or g) on the ambiguity surface that minimizes $E(\underline{x}) \equiv E(g, f)$. (Note that if we did not constrain $\underline{x}$ to be on the ambiguity surface, then we would just be solving the phase-retrieval problem!) If we define the ambiguity surface by $h(\underline{x}) = 0$, then the problem of finding the nearest ambiguity to f can be stated as follows:

Given an object f, find the $\underline{x}$ that minimizes the objective function $E(\underline{x})$ subject to the ambiguity condition $h(\underline{x}) = 0$.

The two image supports for which we have derived ambiguity conditions [Eqs. (32) and (34)] give rise to the following $h(\underline{x})$:

2 × 2 Images (L = 4)

$$h(\underline{x}) = ad - bc = 0, \tag{37}$$

3 × 2 Images (L = 6)

$$h(\underline{x}) = (af - cd)^2 - (ae - bd)(bf - ce) = 0. \tag{38}$$

Iterative Constrained Minimization

Using the mathematical framework developed above, we now implement a generalized reduced-gradient (sometimes referred to as a gradient-projection) method[34] to find the nearest ambiguity to a given image. This method is explained in detail in Appendix C and is summarized below.

In an unconstrained gradient-search method, we search for a minimum to the objective function $E(\underline{x})$ in the direction of $-\nabla E(\underline{x})$, the negative gradient of that function. In a constrained search we still would like to follow the negative gradient, but we are constrained to move along a particular surface within the space, described by the equation $h(\underline{x}) = 0$. We alter the search direction by projecting $-\nabla E(\underline{x})$ onto a tangent plane of $h(\underline{x})$, and we then move along the plane in the direction of the projection, $\underline{p}$, as depicted in Fig. 8. Then, from a point along $\underline{p}$, which is generally not on the constraint surface, we find a nearby (not necessarily the closest) point on the constraint surface. The method used here to return to the constraint surface is detailed in Appendix C. The search for the solution is iterative, and we define our estimate of the solution after the kth iteration as $\underline{x}_k$. At the solution, $\underline{x}_s$, $-\nabla E(\underline{x})$ is perpendicular to the tangent plane to the constraint surface, and the projection onto the tangent plane is zero.

It is difficult to determine whether the minimum found is indeed the global minimum or just a local minimum. In a numerical simulation such as this, one can gain confidence in claiming a minimum as global only through repeated searching with different initial guesses. Our practical criteria for claiming that a minimum, $\underline{x}_s$, is global is that $E(\underline{x}_s)$ is the smallest among all minima found and that it is found more than twice as many times as the total number of minima

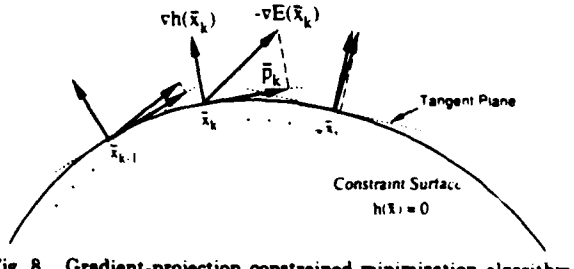


Fig. 8. Gradient-projection constrained-minimization algorithm. The search direction is determined by projecting the negative gradient of the objective function onto the tangent plane to the constraint surface.

found, which must be more than four. If the above criteria are not satisfied after 40 different minima are found, then the one that minimizes $E(\bar{x})$ is chosen (and we simply realize that it may not be the global minimum). It should be noted that at points on the surface where $\nabla h(\bar{x}) = 0$ the tangent plane is not defined. If such a singular point is encountered the search may terminate without satisfying a convergence criterion, but the estimate at the singular point may still minimize the objective function over all other estimates (see Appendix C).

Although the constrained-minimization algorithm minimizes an objective function defined in Fourier-modulus space, the search itself takes place on surfaces in object space. The minima found on the surface of Eq. (37) will always correspond to images with two convolution factors, and that usually will be the case for the minima found on the surface of Eq. (38) as well. Thus the nearest ambiguity in Fourier-modulus space to an object f corresponds in object space to any of four images (not counting scalar multiples of these images): the ambiguity, its ambiguous counterpart, and the twin image of each. So, once we have an estimate of the global minimum with respect to Fourier-domain error [Eq. (36)], denoted by g_1, we calculate the object-domain error δ for g_1 and its twin image, retaining the smaller of the two values. We then find the ambiguous counterpart to g_1, denoted by g_{1c}, by convolving one of the factors of g_1 with the twin of the other. After finding the smaller δ for g_{1c} and its twin, we keep as the worst-case nearest ambiguity the larger of this δ and the one retained for g_1 and its twin. Referring back to Fig. 1, the smaller value of δ corresponds to the nearest ambiguity in the object domain, g_a, and the larger retained value of δ corresponds to its ambiguous counterpart, g_{ac}, the worst-case nearest ambiguity. Although g_a and g_{ac} are both nearest ambiguities to f with respect to Fourier-domain error, we differentiate them by defining the worst-case nearest ambiguity as the one with the larger value of the object-domain error, δ, with respect to f. The worse-case nearest ambiguity corresponds to the point in object space farthest from the true image that either is likely to cause local minima to trap phase-retrieval algorithms or could be confused with the true image if the squared error in the data exceeds $E(\bar{x})$.

Monte Carlo Simulations

To investigate the prevalence of ambiguities we implemented the constrained-minimization nearest-ambiguity search in a Monte Carlo simulation in which nearest ambiguous images were found for a large number of random objects $f(x, y)$. Each pixel of the object was an independent, real-valued random number uniformly distributed on the interval $[-2, 2]$ or $[0, 4]$ for nonnegative objects. The results of the Monte Carlo simulations are presented in the form of scatter plots of ϵ versus δ for the worst-case nearest ambiguity. For each random object f, the value of ϵ for the nearest ambiguity is plotted versus the worst-case δ. The interpretation of these scatter plots should not be confused with that of the grid-search scatter plots shown above. Recall that all the (δ, ϵ) pairs in a grid-search scatter plot are calculated by using a single object f and have nothing to do with ambiguities, while each (δ, ϵ) point in Monte Carlo scatter plot represents metrics for the worst-case nearest ambiguity to a different random object f. We computed these plots for five separate cases: (1) 2×2 objects without a nonnegativity constraint on f, (2) 2×2 objects with a nonnegativity constraint, (3) 3×2 objects without a nonnegativity constraint, (4) 3×2 objects with a nonnegativity constraint, and (5) L-shaped (with $b = c = 0$) 3×2 objects with a nonnegativity constraint. The five cases above represent different constraints on f. The only constraint on the worst-case nearest ambiguity, g_{ac}, is that it lie upon the ambiguity surface corresponding to the support of f.

A typical scatter plot of ~4000 points required ~110 h for the 2×2 objects and ~1500 h for the 3×2 objects on an IBM AT personal computer.

The scatter plots of ϵ versus δ for the 2×2 support cases (1) and (2) are shown in Fig. 9. The points that would cause trouble are those that have small Fourier-modulus error (FME), ϵ, and significantly larger object-domain error (ODE), δ. These troublesome points are likely to induce phase-retrieval algorithm stagnation and/or are ambiguous from a practical point of view when the Fourier-modulus data are sufficiently noisy. One definition of a trouble re-

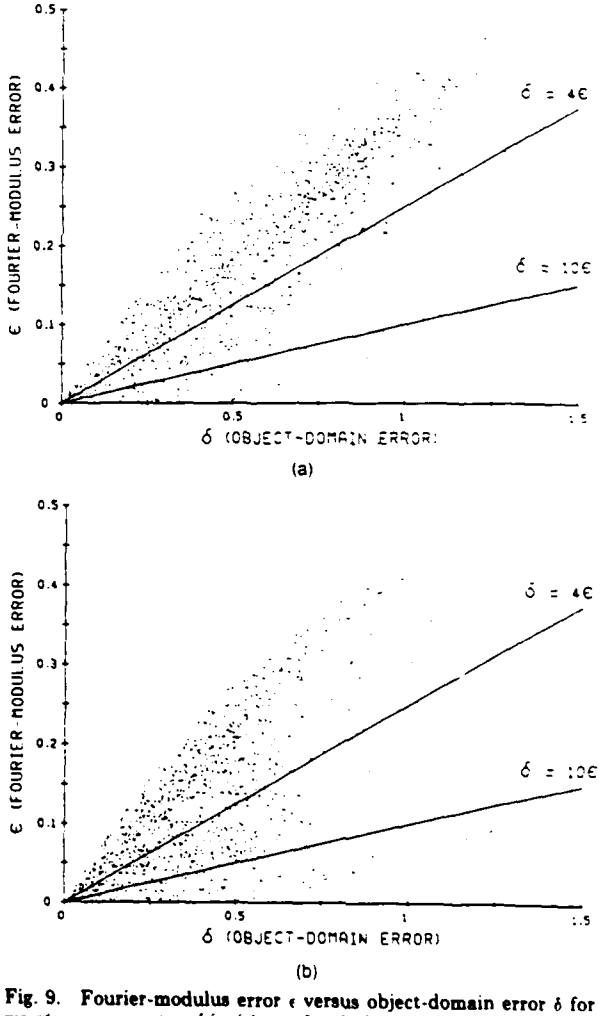


Fig. 9. Fourier-modulus error ϵ versus object-domain error δ for worst-case nearest ambiguities to 2×2 objects. (a) No nonnegativity constraint, 4752 objects; (b) nonnegativity constraint, 4486 objects.

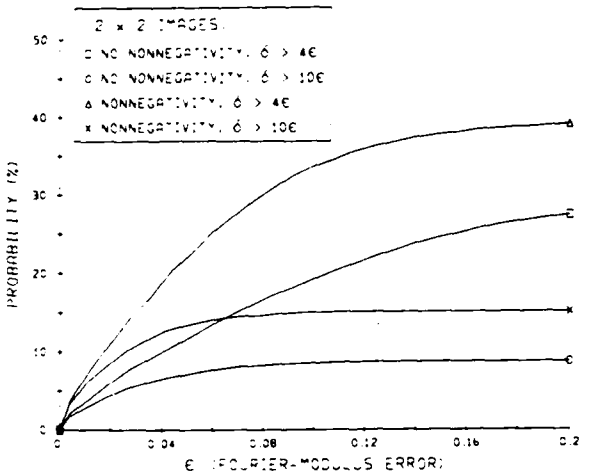


Fig. 10. Monte Carlo estimates of the probability that the worst-case nearest ambiguity to 2 × 2 objects with and without a nonnegativity constraint has a Fourier-modulus error less than ϵ and an object-domain error greater than $K\epsilon$ ($K = 4$ and $K = 10$).

gion is all the points below the line $\delta = K\epsilon$, shown in Fig. 9 for $K = 4$ and $K = 10$. That is, we do not consider the practical ambiguity problem to be significant unless the error, δ, in the ambiguous reconstruction or stagnation point exceeds 4 times (or 10 times) the error in the Fourier-modulus data. Only then would we consider the ambiguity to be significant. (Although it was easy to show in Appendix A that $\delta > \epsilon$ for any pair of images, an analgous relationship for an image and its worst-case nearest ambiguity has not been developed.) Figure 9(a) (no nonnegativity constraint on f) exhibits a banded structure with a higher density of points above the $\delta = 4\epsilon$ line, which effectively reduces the probability of nearest ambiguities in the trouble region. Figure 9(b) (nonnegativity constraint on f) reveals a higher density of points in the trouble region, particularly for $\delta < 0.5$. Thus the nonnegativity constraint on f actually increases the probability that a random object's Fourier modulus is close to that of an ambiguous image for the 2 × 2 case.

One way to estimate the probability of significant ambiguity is to integrate these scatter plots in the trouble region below the line $\delta = K\epsilon$. If we bin the points below this line with respect to ϵ, we can obtain an estimate of the probability-density function of the probability that the worst-case nearest ambiguity has FME ϵ and $\delta > K\epsilon$. Integrating this estimated probability-density function from 0 to ϵ yields an estimate of the probability that the worst-case nearest ambiguity to an arbitrary object has less than ϵ FME and ODE $\delta > K\epsilon$. These cumulative probability distributions define what we mean by the probability of significant ambiguity. These distributions for cases (1) and (2) are shown in Fig. 10 for $K = 4$ and $K = 10$. Figure 10 verifies our previous observation that the nonnegativity constraint actually improves the chance of significant ambiguity. For example, these estimated distributions tell us that, given an arbitrary, real-valued 2 × 2 object, the probability of finding a worst-case nearest ambiguity with FME $\epsilon < 0.04$ and ODE $\delta > 0.16$ is 10% for f without nonnegativity and 18% for f with nonnegativity.

The same analysis for the 3 × 2 object support [cases (3) and (4)] reveals the opposite trend. Figure 11 shows the ϵ versus δ scatter plots for the nearest 3 × 2 ambiguities with and without a nonnegativity constraint on f. With no nonnegativity constraint, the scatter plot of Fig. 11(a) is uniform in appearance, indicating a greater likelihood of nearby ambiguities in the trouble regions. With the nonnegativity constraint, Fig. 11(b) shows a high concentration of points in the large ϵ, large δ region of the plot, away from the trouble region. It is the nonnegativity constraint that creates the favorable banding effect for the 3 × 2 case. Integrating these plots below the $K = 4$ and $K = 10$ lines yields the probability distributions of Fig. 12. In comparison with the example given for the 2 × 2 nonnegative case, the probability of finding a worst-case nearest ambiguity with FME $\epsilon < 0.04$ and ODE $\delta > 0.16$ is increased to 17% without nonnegativity but reduced by approximately one half to 9% with the nonnegativity constraint on f.

One possible reason that nonnegativity reduces the probability of significant ambiguity for the 3 × 2 case is as follows. From Eq. (35) we see that there are no real-valued ambigu-

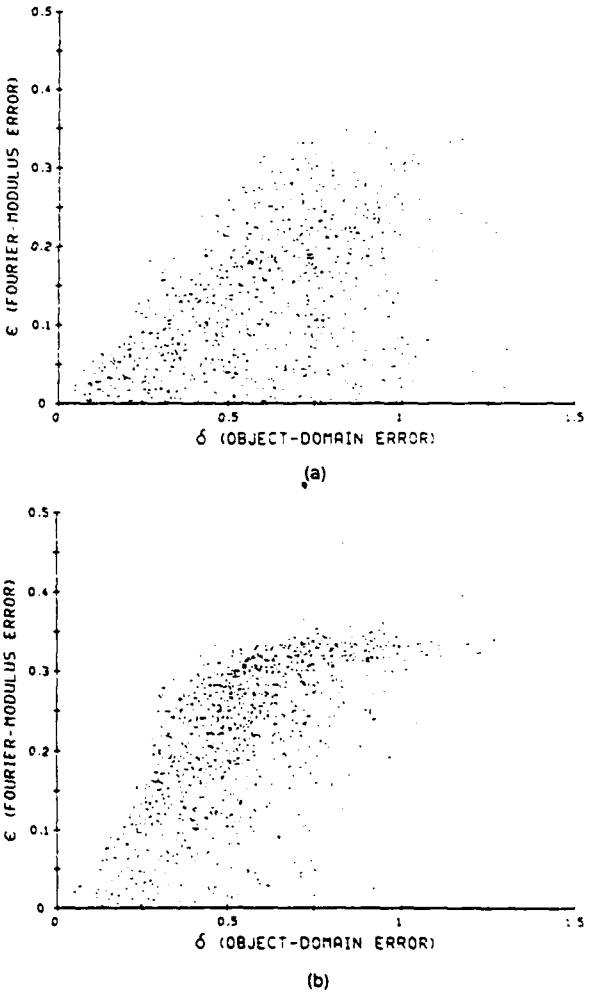


Fig. 11. Fourier-modulus error ϵ versus object-domain error δ for worst-case nearest ambiguities to 3 × 2 objects. (a) No nonnegativity constraint, 4112 objects; (b) nonnegativity constraint, 4601 objects.

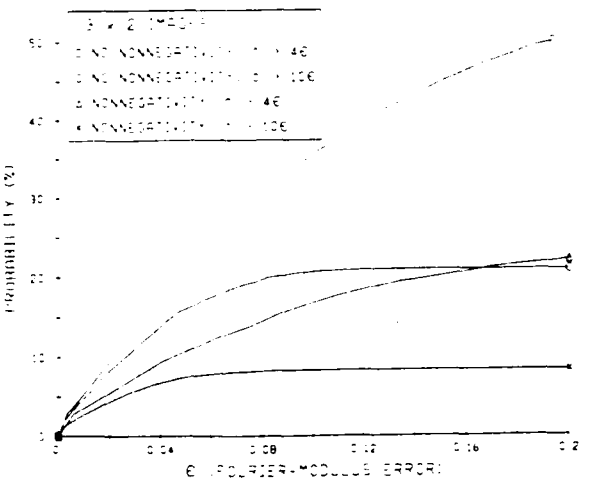


Fig. 12. Monte Carlo estimates of the probability that the worst-case nearest ambiguity to 3×2 objects with and without a nonnegativity constraint has a Fourier-modulus error less than ϵ and an object-domain error greater than $K\epsilon$ ($K = 4$ and $K = 10$).

ous images for which $e^2 - 4df < 0$. Since $-4df$ is negative for positive d and f, but is positive if one of them is negative, $e^2 - 4df$ is more often negative for nonnegative images. Thus nonnegative objects are less likely to have nearest ambiguities that are nearby (in the object domain) than are objects without a nonnegativity constraint. Since objects that are similar in the object domain will tend to be similar in the Fourier-modulus domain, the nearest ambiguities to nonnegative objects are less likely to be nearby with respect to Fourier modulus as well.

An important point that should be stressed is that the nonnegativity constraint discussed in this section is on the object f and not on the nearest ambiguity. Because of this fact, the nearest ambiguous image to a nonnegative object might not be nonnegative itself; it could contain one or two negative-valued pixels. Thus a nonnegativity constraint in a phase-retrieval algorithm may help to move the image away from a stagnation point near the ambiguity, and the probability of ambiguity in the practical sense would be reduced compared with the results shown here.

At this point it is useful to recall the conjecture made in Section 5, i.e., that stagnation points of the iterative Fourier-transform algorithm tend to be near ambiguous images that have Fourier moduli close to the given Fourier modulus, $|F|$. The example given in Section 5 used an object f and its nearest ambiguity [Eqs. (24)–(26)] taken from the Monte Carlo experiment with 3×2 nonnegative objects. Recall that, for the object f of Eq. (24), after numerous trials we found two stagnation points, g_{s1} and g_{s2}, of both the HIO and ER versions of the iterative Fourier-transform algorithm. The closeness in both domains of these stagnation points to the worst-case nearest ambiguity, g_{ac} [Eq. (26)], was shown. A few more simulations of this type were performed for different nonnegative 3×2 objects. Objects were selected based on the locations in Fig. 11(b) of their worst-case nearest-ambiguity error metrics. All objects selected had a worst-case nearest ambiguity with $0.45 < \delta < 0.55$. Three objects with (significant) worst-case nearest ambiguities with $\epsilon < 0.05$ [as was the case for f of Eq. (24)] were selected, and, compared with the 26% success rate for f with HIO, the true solution was found 48%, 49%, and 59% of the time, respectively, by using HIO on these three objects. As with f, when the true solution was not found, the algorithm stagnated near the worst-case nearest ambiguity (g_{ac}) to each of the three objects. Two objects with a worst-case nearest ambiguity with $\epsilon \simeq 0.10$ converged to the true solution 78% and 100% of the time, and another object with a worst-case nearest ambiguity with $\epsilon = 0.30$ converged to the solution 100% of the time. Thus stagnation tends to decrease as the nearest ambiguities move farther away with respect to ϵ (equivalently, as the significance of ambiguity decreases). As mentioned above, the limited number of experiments of this type has not yet provided us with a complete understanding of phase-retrieval stagnation points and their relationship to worst-case nearest ambiguous images. Nevertheless, the correlation of the object's worst-case nearest ambiguity having large δ and small ϵ ($\epsilon < 0.05$ for our experiments) with the presence of stagnation points has been convincingly established.

The final case investigated is nonnegative, 3×2 objects with $b = c = 0$, which we call L-shaped objects. The L-shaped support itself mandates uniqueness; i.e., it is not possible to convolve two nontrivial functions to obtain an image with this support. After running the Monte Carlo simulation for these objects, we discovered a class of L-shaped ambiguities that gives rise to misleading results. Consider the object

$$f = \begin{bmatrix} 1.48155 & 0 & 0 \\ 2.01553 & 3.97050 & 0.16831 \end{bmatrix}, \tag{39}$$

with nearest 3×2 ambiguous image

$$g_a = \begin{bmatrix} 1.48170 & 6.29E\text{-}4 & -2.78E\text{-}3 \\ 2.01419 & 3.97109 & 0.16907 \end{bmatrix} = \begin{bmatrix} 1 & 0.04354 \end{bmatrix} * \begin{bmatrix} 1.48170 & -0.06388 \\ 2.01419 & 3.88340 \end{bmatrix}, \tag{40}$$

with $\delta(g_a, f) = 7.015E\text{-}4$ and $\epsilon(g_a, f) = 4.167E\text{-}4$. The ambiguous counterpart to g_a, obtained by flipping the first convolution factor in Eq. (40), is

$$g_{ac} = \begin{bmatrix} 0.045354 & 1 \end{bmatrix} * \begin{bmatrix} 1.48170 & -0.06388 \\ 2.01419 & 3.88340 \end{bmatrix} = \begin{bmatrix} 0.06451 & 1.47892 & -0.06388 \\ 0.08769 & 2.18326 & 3.88340 \end{bmatrix}. \tag{41}$$

The object-domain error between f and g_{ac} as defined by Eq. (11) is $\delta(g_{ac}, f) = 1.0629$. However, comparison of f and g_{ac} reveals that the image g_{ac} is similar to the image f shifted by one pixel to the right. This is because the first convolution factor of Eq. (41) is nearly a delta function, and the second factor is very similar to the image f without its right-hand column. The first convolution factor causes a tapering of the image, making one column much smaller in value than the other nonzero pixels. Flipping one of the convolution factors simply shifts the significant pixels and moves the tapered column to the other side of the image. Because the object-domain error metric δ does not take such shifts into account, the value of $\delta(g_{ac}, f)$ calculated for this case is much too large, resulting in a misleading point on the scatter plot. (If the calculations were to be redone, then this problem could be accounted for by cross correlating g_{ac} with f and

shifting g_{ac} according to the cross-correlation peak to minimize δ.)

A similar problem may occur if the shorter leg of the L-shaped support is tapered, leading to nearest ambiguities that are close to 1-D sequences. To reduce the misleading effects of tapered images on our analysis, we consider only those images that satisfy a bound on the robustness of the L shape. An L-shaped image $\begin{bmatrix} a & 0 \\ d & f \end{bmatrix}$ has L robustness R%, defined by

$$\frac{R}{100} = \min[a, f]/[(a^2 + d^2 + e^2 + f^2)/4]^{1/2}. \quad (42)$$

Images with large R are robustly L shaped, whereas images with small R (strongly tapered) are only weakly L shaped.

It should be noted that the same taper problem can also cause misleading ODE calculations of worst-case nearest ambiguities for the 2×2 and 3×2 images in cases (1)–(4). In these cases, whole rows or columns would have to be significantly smaller than the rms pixel intensity of the image. Since the images are random, it is much less likely for this to occur in cases (1)–(4) for which two or more pixels must be small simultaneously than for the L-shaped case (5) for which only a single pixel must be small.

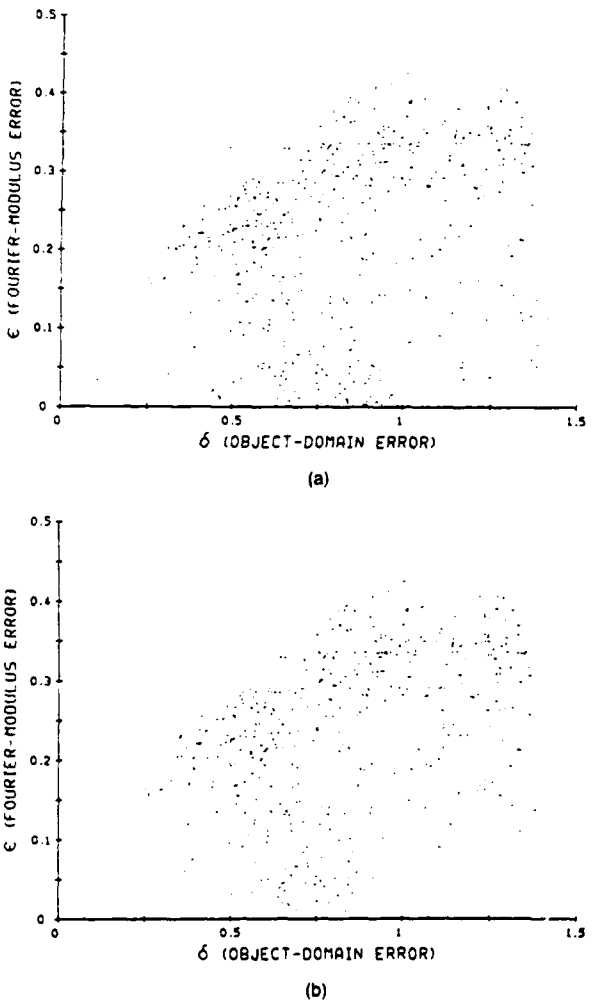


Fig. 13. Fourier-modulus error ϵ versus object-domain error δ for worst-case nearest ambiguities to 3×2, nonnegative, L-shaped objects. (a) L robustness > 10%, 3190 objects; (b) L robustness > 25%, 2714 objects.

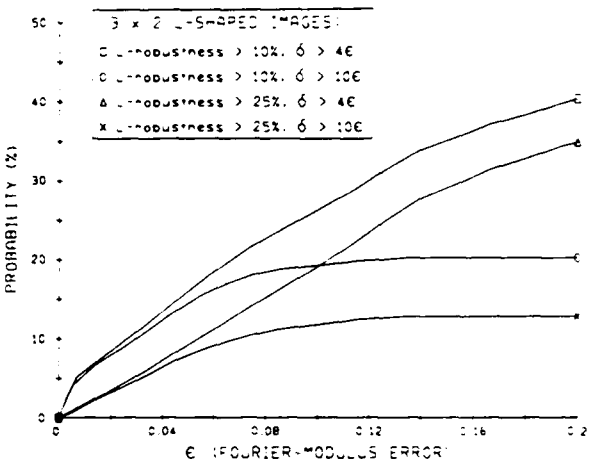


Fig. 14. Monte Carlo estimates of the probability that the worst-case nearest ambiguity to 3×2, nonnegative, L-shaped objects with L robustness greater than R% ($R = 10$ and $R = 25$) has a Fourier-modulus error less than ϵ and an object-domain error greater than $K\epsilon$ ($K = 4$ and $K = 10$).

The worst-case nearest-ambiguity scatter plots for nonnegative, L-shaped images with L robustness greater than 10% and 25% are shown in Fig. 13. As the L-robustness requirement is increased, many points clustered about the δ axis disappear. (Had we been able to calculate δ with image shifts taken into account, we would have found these points moving horizontally into the small δ, small ϵ region of the plot.) Despite the nonnegativity of f, these scatter plots are less banded than for general 3×2 nonnegative objects—case (4) in Fig. 11(b). This is verified by the estimated distributions for both taper percentages (Fig. 14). For the case of L robustness greater than 25%, the distributions of Fig. 14 achieve a lower probability than does case (4) for values of ϵ less than 0.07, reflected by the small number of points near the origin of the plots in Fig. 13. Therefore, for the low-noise case, the L-shaped support constraint not only prevents ambiguity in the absolute sense but it also makes ambiguity less likely in the practical sense.

7. SUMMARY AND CONCLUSIONS

An ambiguous image is one whose Fourier modulus is identical to the Fourier modulus of a second image that is other than a scaled version, a translation, or a twin of the image. Arbitrary objects are almost never (i.e., with probability zero) ambiguous. Nevertheless, the existence of an ambiguous image close to a given object has two harmful effects: it causes stagnation points for phase-retrieval algorithms and, for the case of noisy Fourier-modulus data, it may cause the solution to be ambiguous in the practical sense. Because of the nonlinearity of the phase-retrieval problem, these issues are difficult to characterize analytically. We investigated the prevalence of ambiguous images for the phase-retrieval problem, using numerical approaches. This is practical because we considered the case of small objects defined on 2×2 and 3×2 supports.

Using both a new iterative grid-search algorithm and the iterative Fourier-transform algorithm, multiple phase-retrieval experiments were performed, and stagnation points were found that correspond to local minima in the Fourier-domain error metric. These stagnation points were shown to be close to ambiguous images whose Fourier moduli are close to the modulus of the Fourier transform of the object. The implication is that the existence of the ambiguous images causes the local minima to occur. However, the precise relationship between the local minima and the ambiguous images is not yet understood, and nearest ambiguities may not be the sole cause of stagnation.

The prevalence of ambiguities close (with respect to Fourier modulus) to a given object was explored by a Monte Carlo experiment in which nearest ambiguities were found. First, object-domain analytic expressions for the set of ambiguous images were derived for both the 2×2 and 3×2 supports [Eqs. (37) and (38)]. For the 2×2 case, the set of ambiguous images forms a three-dimensional surface embedded in the four-dimensional space of 2×2 real-valued images. For the 3×2 case, the set of ambiguous images forms a five-dimensional surface embedded in the 6-D space of 3×2 real-valued images. Next, a reduced-gradient search technique was used to search along the surfaces of ambiguous images to find the ambiguous image nearest a given object with respect to Fourier modulus. Of the nearest-ambiguity pair of images, one is usually close to the object f, while its ambiguous counterpart is usually a worse case: it is much farther from the given object, yet it has a Fourier modulus identical to the ambiguous image that is close to f. Histograms of Fourier-modulus-domain versus object-domain errors were accumulated in Monte Carlo experiments involving numerous random objects and their worst-case nearest ambiguities. Integration of the histograms, over the points for which the object-domain error is large relative to the Fourier-modulus error, yielded estimates of the probability that a significant ambiguity would occur within a given Fourier-modulus error tolerance. It was found that nonnegativity of the object decreased the probability of significant ambiguity for the 3×2 case (as anticipated) but increased the probability of significant ambiguity for the 2×2 case. However, since the ambiguous images were allowed to have negative values even when the objects were restricted to be nonnegative, it is likely that the imposition of a nonnegativity constraint in a phase-retrieval algorithm would help to avoid some of those ambiguities. L-shaped images, whose support guarantees uniqueness in the absolute sense, were also investigated. It was found that, for low-noise data, the L-shaped support of the object also makes ambiguity less likely in the practical sense.

Future work should include the application of this approach to objects with larger supports. This is important since it is difficult to extrapolate from these results for 2×2 and 3×2 supports to the case of most interest: supports with many pixels in each dimension. The probability of significant ambiguity for the 3×2 case was of similar magnitude to that of the 2×2 case. This is probably because the ambiguity surfaces in both cases were of dimension one less than the dimension of the space of objects. When larger objects are considered, however, this changes. For example, for 3×3 objects

$$\begin{bmatrix} a & b & c \\ d & e & f \\ g & h & i \end{bmatrix},$$

factoring into a (3×2) convolved with a (1×2), the ambiguity condition is given by the simultaneous equations

$$(ah - bg)^2 - (ae - bd)(dh - eg) = 0 \tag{43}$$

and

$$(ah - bg)(af - cd) - (ae - bd)(ai - cg) = 0. \tag{44}$$

These describe two eight-dimensional surfaces embedded in a nine-dimensional space of 3×3 real-valued objects, the intersection of which would ordinarily be expected to be a seven-dimensional surface embedded in the nine-dimensional space. The ambiguity condition for the factoring of a 3×3 object into a (2×2) convolved with another (2×2) is also given by a pair of simultaneous equations describing two eight-dimensional surfaces embedded in a nine-dimensional space, the intersection of which would ordinarily be a seven-dimensional surface in the nine-dimensional space. Thus for these larger images the dimensionality of the surface of ambiguous images is smaller relative to the space of all objects than for the 2×2 or 3×2 case; consequently one would expect the probability of significant ambiguity to be less for these larger images. The importance of the shape of the support constraint (convex versus nonconvex versus separated parts, etc.) may also reveal itself more forcefully for larger supports. Finally, a better understanding of the precise relationship between local minima and nearest ambiguous images could lead to methods for avoiding phase-retrieval algorithm stagnation at local minima.

APPENDIX A: PROOF THAT $\epsilon \le \delta$

By definition, $|\alpha_0| = \alpha_f$, or $\alpha_0 = \pm\alpha_f$. The proof that $\epsilon(g, f) \le \delta(g, f)$ can be given by using Parseval's theorem with the definition of $\delta(g, f)$:

$$\delta(g,f) = \left[\sum_{x,y} [\alpha_0 g(x,y) - f(x,y)]^2 \Big/ \sum_{x,y} f^2(x,y)\right]^{1/2}$$

$$= \left[\sum_{u,v} |\pm\alpha_f G(u,v) - F(u,v)|^2 \Big/ \sum_{u,v} |F(u,v)|^2\right]^{1/2}. \tag{A1}$$

By the triangle inequality, given two vectors, v_1 and v_2, $|v_1 - v_2|^2 \ge [|v_1| - |v_2|]^2$. Therefore

$$|\pm\alpha_f G(u,v) - F(u,v)|^2 \ge [\alpha_f |G(u,v)| - |F(u,v)|]^2. \tag{A2}$$

Inserting inequality (A2) into Eq. (A1), we have

$$\delta(g,f) \ge \left[\sum_{u,v} [\alpha_f |G(u,v)| - |F(u,v)|]^2 \Big/ \sum_{u,v} |F(u,v)|^2\right]^{1/2}$$

$$\equiv \epsilon(g,f). \tag{A3}$$

APPENDIX B: DERIVATION OF 3 × 2 AMBIGUITY CONDITION

Equation (33) gives us the following six equations:

$$a = gi, \tag{B1a}$$
$$b = hi + gj, \tag{B1b}$$
$$c = hj, \tag{B1c}$$
$$d = gk, \tag{B1d}$$
$$e = hk + gl, \tag{B1e}$$
$$f = hl. \tag{B1f}$$

Multiplying Eqs. (B1a) and (B1f) gives

$$af = ghil, \tag{B2}$$

and multiplying Eqs. (B1c) and (B1d) gives

$$cd = ghjk. \tag{B3}$$

Combining these yields

$$(af - cd)^2 = g^2h^2(il - jk)^2. \tag{B4}$$

From Eqs. (B1b), (B1c), (B1e), and (B1f) we have

$$(bf - ce) = h^2(il - jk), \tag{B5}$$

and from Eqs (B1a), (B1b), (B1d), and (B1e) we have

$$(ae - bd) = g^2(il - jk). \tag{B6}$$

Taking the product of Eqs. (B5) and (B6) yields

$$(ae - bd)(bf - ce) = g^2h^2(il - jk)^2. \tag{B7}$$

From Eqs. (B4) and (B7) we arrive at the result

$$(af - cd)^2 - (ae - bd)(bf - ce) = 0. \tag{B8}$$

This equation is the condition that must be met in order for the 3 × 2 image of Eq. (33) to be ambiguous.

From Eq. (B8) we can solve for any of the six variables in terms of the other five. For example, by expanding and collecting powers of b, we arrive at

$$b^2(df) - b(aef + cde) + ace^2 + (af - cd)^2 = 0. \tag{B9}$$

The solution of Eq. (B9), which is quadratic in b, is given by

$$b = \frac{[e(af + cd) \pm (e^2 - 4df)^{1/2}(af - cd)]}{2df}$$

$$= \frac{1}{2}\left[e\left(\frac{c}{f} + \frac{a}{d}\right) \pm (e^2 - 4df)^{1/2}\left(\frac{c}{f} - \frac{a}{d}\right)\right]. \tag{B10}$$

APPENDIX C: GENERALIZED REDUCED-GRADIENT METHOD

The generalized reduced-gradient method is a gradient-projection technique used to apply a set of constraints to a minimization problem. The application discussed here uses a single nonlinear homogeneous constraint, $h(\bar{x}) = 0$, and the discussion is presented with this assumption. We begin by defining the tangent plane to a surface:

Given a point $\bar{x}^*$ satisfying $h(\bar{x}^*) = 0$, the tangent plane T at that point is $T = \{\bar{y}: \nabla h(\bar{x}^*) \cdot \bar{y} = 0\}$, where ∇ denotes the gradient with respect to $\bar{x}$ and $\cdot$ denotes the dot product.

Simply stated, all vectors $\bar{y}$ in the tangent plane T are perpendicular to the gradient of $h(\bar{x})$ at $\bar{x}^*$.

In an unconstrained gradient-search method, we would search for a minimum to the objective function $E(\bar{x})$ in the direction of the negative gradient of that function, $-\nabla E(\bar{x})$. In a constrained search, however, the solution is constrained to a particular surface within the space, and we must alter the direction of the search to remain on the surface. We do this by projecting $-\nabla E(\bar{x})$ onto a tangent plane of $h(\bar{x})$ and moving along the plane in the direction of the projection, $\bar{p}$. Because points lying along $\bar{p}$ in general will not lie upon the constraint surface, the goal is to move along $\bar{p}$ and then to return to the surface $h(\bar{x}) = 0$ such that there is a sufficient decrease in the objective function. More will be said below about how to return to the surface from the projection onto the tangent plane.

The solution point, $\bar{x}_s$, satisfies the following first-order condition:

All $\bar{y}$ satisfying $\nabla h(\bar{x}_s) \cdot \bar{y} = 0$ (in the tangent plane at $\bar{x}_s$) must also satisfy $-\nabla E(\bar{x}_s) \cdot \bar{y} = 0$.

The above definition implies that, at the solution, $-\nabla E$ is parallel to ∇h, which in turn implies that the projection $\bar{p}$ is zero. Note that the above definition applies to any minimum and not just to the global minimum.

The search is iterative, and we define $\bar{x}_k$ as our estimate of the solution after k iterations. The goal is to find $\bar{x}_{k+1}$ such that $E(\bar{x}_k)$ significantly decreases at each iteration and to continue iterating until the first-order condition above is satisfied with a sufficient degree of confidence.

We now discuss the reduced-gradient method in more specific terms for the case of a single homogeneous constraint. Let us assume we are working in an L-dimensional space. A tangent plane to $h(\bar{x})$ can be thought of as a surface of dimension one less than the space in which it lies. In order to use projection ideas from linear algebra, we define the tangent plane as a space spanned by a set of basis vectors.

A vector that is perpendicular to the tangent plane to $h(\bar{x})$ at a point $\bar{x} = (x_1\, x_2 \ldots x_L)^t$ is

$$\nabla h(\bar{x}) = \left(\frac{\partial h}{\partial x_1} \quad \frac{\partial h}{\partial x_2} \quad \cdots \quad \frac{\partial h}{\partial x_L}\right)^t. \tag{C1}$$

A set of $L - 1$ linearly independent L-dimensional basis vectors that span the space perpendicular to $\nabla h(\bar{x})$ (i.e., the tangent plane) is (assuming that $\partial h/\partial x_1 \neq 0$)

$$\bar{b}_1 = \left[-\left(\frac{\partial h}{\partial x_1}\right)^{-1}\left(\frac{\partial h}{\partial x_2}\right) \quad 1 \quad 0 \quad 0 \ldots 0\right]^t,$$

$$\bar{b}_2 = \left[-\left(\frac{\partial h}{\partial x_1}\right)^{-1}\left(\frac{\partial h}{\partial x_3}\right) \quad 0 \quad 1 \quad 0 \ldots 0\right]^t,$$

$$\vdots \qquad\qquad \vdots$$

$$\bar{b}_{L-1} = \left[-\left(\frac{\partial h}{\partial x_1}\right)^{-1}\left(\frac{\partial h}{\partial x_L}\right) \quad 0 \quad 0 \quad 0 \ldots 1\right]^t. \tag{C2}$$

The set of basis vectors defined in Eqs. (C2) enables us to define a projection onto the tangent plane to $h(\bar{x})$. If we let the $\bar{b}$'s be the columns of an $L \times (L - 1)$ matrix,

$$Z \equiv [\bar{b}_1 \quad \bar{b}_2 \ldots \bar{b}_{L-1}], \tag{C3}$$

then the projection of an arbitrary L-dimensional vector, $\bar{\omega}$, onto the space spanned by the columns of Z is[14]

$$\bar{p} = Z(Z^tZ)^{-1}Z^t\bar{\omega}. \tag{C4}$$

From Eq. (C4), the projection of $-\nabla E(\bar{x})$ onto the tangent plane to $h(\bar{x})$ is just

$$\bar{p} = -Z(Z^tZ)^{-1}Z^t\nabla E(\bar{x}). \tag{C5}$$

For each estimate $\bar{x}_k$ of the solution we have $h(\bar{x}_k) = 0$. The reduced-gradient method calculates $\nabla h(\bar{x}_k)$, Z_k, and $-\nabla E(\bar{x}_k)$ and uses these with Eq. (C5) to determine the new search direction:

$$\bar{p}_k = -Z_k(Z_k^tZ_k)^{-1}Z_k^t\nabla E(\bar{x}_k). \tag{C6}$$

Once $\bar{p}_k$ is determined, we must move from $\bar{x}_k$ in the direction of $\bar{p}_k$ to find the next estimate $\bar{x}_{k+1}$. However, we must have $h(\bar{x}_{k+1}) = 0$, and, in general, it is not possible to find a step size $\gamma_k \neq 0$ along $\bar{p}_k$ such that $h(\bar{x}_k + \gamma_k\bar{p}_k) = 0$. It becomes necessary to deviate from $\bar{p}_k$ to return to the surface for our next estimate. This estimate becomes

$$\bar{x}_{k+1} = \bar{x}_k + \gamma_k\bar{p}_k + \bar{q}_k, \tag{C7}$$

with

$$h(\bar{x}_{k+1}) = 0, \tag{C8}$$

where γ_k and $\bar{q}_k$ are chosen such that $E(\bar{x}_{k+1}) < E(\bar{x}_k)$. Determining the scalar step size γ_k and the direction back to the surface, $\bar{q}_k$, in Eq. (C7) that minimize $E(\bar{x}_{k+1})$ can be a complicated subproblem.

Rather than spending too much computation time determining the optimal γ_k and $\bar{q}_k$, we opt for a simpler approach to finding an $\bar{x}_{k+1}$ that produces a sufficient decrease in the objective function. We do this by (1) selecting a value for γ_k, then (2) using $\bar{x}_k + \gamma_k\bar{p}_k$ for all but one of the components of $\bar{x}_{k+1}$, and then (3) using Eq. (C8) to determine the last component. Equation (35) is an example of Eq. (C8) for solving for the component b. The objective function is evaluated to determine whether there is a sufficient decrease. If we are not satisfied with the new estimate, we choose another value of γ_k and repeat the procedure. Using this procedure, we can think of the objective function as a function of γ and can set γ_k to the value of γ that minimizes $E(\gamma)$. One could use any of a number of standard line search techniques to estimate γ_k, but we used a slightly different method to estimate this minimum and to find $\bar{x}_{k+1}$.

Iterative Quadratic Fit

The technique implemented to minimize $E(\gamma)$ with respect to γ can best be described as an iterative quadratic fit (IQF). It uses quadratic curve fitting to approximate the minimum of $E(\gamma)$ iteratively and thus determine γ_k. The description of the IQF below assumes the ability to fit a quadratic polynomial to three points:

(1) Initialize: $\gamma_1 = \gamma_{m2} = 0$, γ_2, γ_3.
(2) Calculate $E(\gamma_1)$, $E(\gamma_2)$, and $E(\gamma_3)$.
(3) Calculate γ_{m1}, the value of γ that corresponds to the minimum of the quadratic polynomial in γ fit to the points $[\gamma_1, E(\gamma_1)]$, $[\gamma_2, E(\gamma_2)]$, $[\gamma_3, E(\gamma_3)]$.
(4) Calculate $E(\gamma_{m1})$.
(5) If $|\gamma_{m1} - \gamma_{m2}| < \beta$, then $\gamma_k \leftarrow \gamma_{m1}$ and stop; otherwise continue with step (6).
(6) Of γ_1, γ_2, and γ_3, find the two that are closest to γ_{m1}. Call these γ_{c1} and γ_{c2}.
(7) Set: $\gamma_1 \leftarrow \gamma_{c1}$, $E(\gamma_1) \leftarrow E(\gamma_{c1})$,
$\gamma_2 \leftarrow \gamma_{c2}$, $E(\gamma_2) \leftarrow E(\gamma_{c2})$,
$\gamma_3 \leftarrow \gamma_{m1}$, $E(\gamma_3) \leftarrow E(\gamma_{m1})$,
$\gamma_{m2} \leftarrow \gamma_{m1}$.
(8) Go to step (3).

The initial values of γ_2 and γ_3 should be chosen based on experimentation and observation of typical $E(\gamma)$ versus γ curves. These values are not crucial to the success of the quadratic fit but should be spaced well enough to give a reasonable initial fit. The value of the termination parameter β should be based on the degree of accuracy needed and should be chosen large enough to avoid excessive iterations.

The success of the IQF depends largely on the shape of $E(\gamma)$. If $E(\gamma)$ is not fairly smooth, the IQF may not find the actual minimum; this is not a problem if a sufficient decrease in E is achieved. A more difficult problem occurs when the projection onto the tangent plane extends into a region of the 6-D space for which the equation for a return to the surface is not defined. As an example, consider using Eq. (35) to return to the surface by calculating b given the other five variables. If a range of values of γ exists for which $\gamma\bar{p}_k$ extends into the region where $e^2 - 4df < 0$, then b (which is by definition real valued) and hence $E(\gamma)$ will not be defined over this range. When we encountered a case such as this, we implemented a Fibonacci line search[14] to estimate the minimum of $E(\gamma)$ on the interval γ for which $E(\gamma)$ is defined. It should be stressed that these potential problems arise out of the method used here to return to the ambiguity surface, and other methods exist that may circumvent this but that are more computationally burdensome.

SPECIFICS TO THE NEAREST-AMBIGUITY SEARCH

Since we have discussed the constrained-minimization technique in somewhat general terms to this point, let us now mention some details and summarize the procedure.

The gradient of $E(\bar{x})$ of Eq. (36) can be computed by using the following relationship[24]:

$$\frac{\partial E}{\partial g(x, y)} = 2MN[g(x, y) - g'(x, y)], \tag{C9}$$

where

$$\mathrm{DFT}[g'(x, y)] = \frac{|F(u, v)|}{|G(u, v)|}G(u, v). \tag{C10}$$

Since the ordering of the pixels of $g(x, y)$ in the vector $\bar{x}$ is defined, Eq. (C9) can be used to calculate the components of $\nabla E(\bar{x})$ using two DFT's [since $|F(u, v)|$ is given].

The various steps of the reduced-gradient constrained-minimization algorithm are as follows:

1. Initialization
   (a) Determine $|F(u, v)|$.
   (b) Make an initial guess, $\bar{x}_0$, such that $h(\bar{x}_0) = 0$.
   (c) Compute $E(\bar{x}_0)$.
   (d) $k = 0$.
2. Calculating the search direction, $\bar{p}_k$
   (a) Compute $\nabla h(\bar{x}_k)$.

(b) Form Z_k, Z_k^t.
(c) Compute $\nabla E(\bar{x}_k)$.
(d) Compute $\bar{p}_k = -Z_k(Z_k^t Z_k)^{-1} Z_k^t \nabla E(\bar{x}_k)$.

3. Iterative Quadratic Fit to find $\bar{x}_{k+1}$ from $\bar{x}_k$ and $\bar{p}_k$
4. If $[E(\bar{x}_k) - E(\bar{x}_{k+1})]/E(\bar{x}_k) < \alpha$,
   then: Done; estimate of minimum is $\bar{x}_{k+1}$.
   else: (a) $k \leftarrow k + 1$.
   (b) Go to step 2.

The termination condition in step 4 above is based on a percentage change between successive iterations. The bound α is selected to reflect the precision of the estimate of the minimum. While it may be tempting to use the condition that $-\nabla E$ is perpendicular to the tangent plane, that is,

$$-\nabla E(\bar{x}_{k+1}) \cdot \bar{p}_{k+1} < \zeta \tag{C11}$$

for some small ζ, it is also difficult to pick the value of ζ that will consistently give us the same confidence in the precision of our estimate without choosing it so small that it causes needless iterations in many cases.

ACKNOWLEDGMENTS

The authors thank T. R. Crimmins for helpful suggestions. This research was supported by the U.S. Office of Naval Research under Contract N00014-86-C-0587.

Portions of this paper were presented at the Optical Society of America Topical Meeting on Signal Recovery and Synthesis III, North Falmouth, Massachusetts, June 14–16, 1989.[35]

REFERENCES

1. E. Wolf, "Is a complete determination of the energy spectrum of light possible from measurements of the degree of coherence," Proc. Phys. Soc. (London) **80**, 1269–1272 (1962).
2. A. Walther, "The question of phase retrieval in optics," Opt. Acta **10**, 41–49 (1963).
3. E. M. Hofstetter, "Construction of time-limited functions with specified autocorrelation functions," IEEE Trans. Inf. Theory **IT-10**, 119–126 (1964).
4. J. R. Fienup, "Reconstruction of an object from the modulus of its Fourier transform," Opt. Lett. **3**, 27–29 (1978).
5. P. J. Napier and R. H. T. Bates, "Inferring phase information from modulus information in two-dimensional aperture synthesis," Astron. Astrophys. Suppl. **15**, 427–430 (1974).
6. W. O. Saxton, *Computer Techniques for Image Processing in Electron Microscopy* (Academic, New York, 1978).
7. G. H. Stout and L. H. Jensen, *X-Ray Structure Determination* (Macmillan, London, 1968).
8. W. Lawton, "A numerical algorithm for 2-D wavefront reconstruction from intensity measurements in a single plane," in *1980 International Optical Computing Conference*, W. T. Rhodes, ed., Proc. Soc. Photo-Opt. Instrum. Eng. **231**, 94–98 (1980).
9. M. Nieto-Vesperinas, "Dispersion relations in two dimensions: application to the phase problem," Optik (Stuttgart) **56**, 377–384 (1980).
10. I. Manolitsakis, "Two-dimensional scattered fields: a description in terms of the zeros of entire functions," J. Math. Phys. **23**, 2291–2298 (1982).
11. R. Barakat and G. Newsam, "Necessary conditions for a unique solution to two-dimensional phase recovery," J. Math. Phys. **25**, 3190–3193 (1984).
12. J. L. C. Sanz and T. S. Huang, "Unique reconstruction of a band-limited multidimensional signal from its phase or magnitude," J. Opt. Soc. Am. **73**, 1446–1450 (1983).
13. I. S. Stefanescu, "On the phase retrieval problem in two dimensions," J. Math. Phys. **26**, 2141–2160 (1985).
14. Yu. M. Bruck and L. G. Sodin, "On the ambiguity of the image reconstruction problem," Opt. Commun. **30**, 304–308 (1979).
15. L. Carlitz, "The distribution of irreducible polynomials in several indeterminates," Ill. J. Math. **7**, 371–375 (1963).
16. M. H. Hayes and J. H. McClellan, "Reducible polynomials in more than one variable," Proc. IEEE **70**, 197–198 (1982).
17. M. A. Fiddy, B. J. Brames, and J. C. Dainty, "Enforcing irreducibility for phase retrieval in two dimensions," Opt. Lett. **8**, 96–98 (1983).
18. M. Nieto-Vesperinas and J. C. Dainty, "A note on Eisenstein's irreducibility criterion for two-dimensional sampled objects," Opt. Commun. **54**, 333–334 (1985).
19. B. J. Brames, "Unique phase retrieval with explicit support information," Opt. Lett. **11**, 61–63 (1986).
20. J. R. Fienup, "Reconstruction of objects having latent reference points," J. Opt. Soc. Am. **73**, 1421–1426 (1983).
21. T. R. Crimmins, "Phase retrieval for discrete functions with support constraints," J. Opt. Soc. Am A **4**, 124–134 (1987).
22. J. L. C. Sanz, T. S. Huang, and F. Cukierman, "Stability of unique Fourier-transform phase reconstruction," J. Opt. Soc. Am. **73**, 1442–1445 (1983).
23. J. R. Fienup, "Space object imaging through the turbulent atmosphere," Opt. Eng. **18**, 529–534 (1979).
24. J. R. Fienup, "Phase retrieval algorithms: a comparison," Appl. Opt. **21**, 2758–2769 (1982).
25. J. R. Fienup and C. C. Wackerman, "Phase-retrieval stagnation problems and solutions," J. Opt. Soc. Am. A **3**, 1897–1907 (1986).
26. P. VanToorn, A. H. Greenaway, and A. M. J. Huiser, "Phaseless object reconstruction," Opt. Acta **7**, 767–774 (1984).
27. J. R. Fienup, "Experimental evidence of the uniqueness of phase retrieval from intensity data," in *Indirect Imaging*, J. A. Roberts, ed., Proceedings of International Union of Radio Science/International Astronomical Union Symposium, August 30–September 2, 1983, Sydney, Australia (Cambridge U. Press, Cambridge, 1984), pp. 99–109.
28. G. B. Feldkamp and J. R. Fienup, "Noise properties of images reconstructed from Fourier modulus," in *1980 International Optical Computing Conference*, W. T. Rhodes, ed., Proc. Soc. Photo-Opt. Instrum. Eng. **231**, 84–93 (1980).
29. R. G. Paxman, J. R. Fienup, and J. T. Clinthorne, "Effect of tapered illumination and Fourier intensity errors on phase retrieval," in *Digital Image Recovery and Synthesis*, P. S. Idell, ed., Proc. Soc. Photo-Opt. Instrum. Eng. **828**, 184–189 (1987).
30. A. M. J. Huiser and P. VanToorn, "Ambiguity of the phase-reconstruction problem," Opt. Lett. **5**, 499–501 (1980).
31. E. N. Leith and J. Upatnieks, "Reconstructed wavefronts and communication theory," J. Opt. Soc. Am. **52**, 1123–1130 (1962).
32. T. R. Crimmins and J. R. Fienup, "Uniqueness of phase retrieval for functions with sufficiently disconnected support," J. Opt. Soc. Am. **73**, 218–221 (1983).
33. M. Nieto-Vesperinas, R. Navarro, and F. J. Fuentes, "Performance of a simulated-annealing algorithm for phase retrieval," J. Opt. Soc. Am. A **5**, 30–38 (1988).
34. P. E. Gill, W. Murray, and M. H. Wright, *Practical Optimization* (Academic, London, 1981).
35. J. H. Seldin and J. R. Fienup, "Numerical investigation of phase retrieval uniqueness," in *Digest of Topical Meeting on Signal Recovery and Synthesis III* (Optical Society of America, Washington, D.C., 1989), pp. 120–123.