Hubble Space Telescope characterized by using phase-retrieval algorithms

J. R. Fienup, J. C. Marron, T. J. Schulz, and J. H. Seldin

We describe several results characterizing the Hubble Space Telescope from measured point spread functions by using phase-retrieval algorithms. The Cramer–Rao lower bounds show that point spread functions taken well out of focus result in smaller errors when aberrations are estimated and that, for those images, photon noise is not a limiting factor. Reconstruction experiments with both simulated and real data show that the calculation of wave-front propagation by the retrieval algorithms must be performed with a multiple-plane propagation rather than a simple fast Fourier transform to ensure the high accuracy required. Pupil reconstruction was performed and indicates a misalignment of the optical axis of a camera relay telescope relative to the main telescope. After we accounted for measured spherical aberration in the relay telescope, our estimate of the conic constant of the primary mirror of the HST was $-1.0144$.

Key words: Hubble Space Telescope, phase retrieval, aberrations.

1. Introduction

Soon after the Hubble Space Telescope (HST) was launched, researchers discovered that its optics were seriously aberrated. Several groups mounted efforts to characterize accurately the aberrations and alignment of the HST. This characterization is important so that we can, first, implement a future optical correction to the aberration; second, know how to align the secondary mirror of the telescope to minimize astigmatism and coma; and third, compute analytically the point spread functions (PSF's) of the system for optimum deblurring of the degraded images currently being collected by the telescope. The latter is important because the PSF changes significantly from one camera to the next and with the position of the field of view, wavelength, and focus setting.

This paper describes elements of a multifaceted program that characterizes the HST performed in the Optical and Infrared Science Laboratory of the Environmental Research Institute of Michigan. Parallel efforts by other groups are described elsewhere in this feature issue of Applied Optics and in Refs. 1 and 2.

Most of the characterization approaches were derived from earlier phase-retrieval approaches that have been in use for several years in the applications of image reconstruction with astronomical interferometry and of wave-front sensing in electron microscopy and optics.

In this paper we begin with the more theoretical and then proceed to the specifics of the HST. In Section 2 we calculate Cramer–Rao (CR) lower bounds on the error in estimating the coefficients of a polynomial expansion of the phase-error (aberration) estimate. This proves useful not only for determining how accurate an estimate to expect but also for selecting a focus setting for the telescope that yields a PSF from which the most accurate estimate can be made. In Section 3 we show the results of computer simulating the HST and performing aberration estimates from the simulated data. This also proves useful for determining the accuracy that can be expected and shows the potential sensitivity of the estimate to imperfections in our model of the HST. Included are the effects of multiple-plane (diffraction) propagation and an imprecisely known spatial scale and quadratic phase factor. This section ends with the results of a blind test that was distributed to several groups. Section 4 describes some of the parameters of the HST that are important in the determination of the aberrations by phase retrieval. We calculate sampling requirements that limit the utility of PSF's measured at wavelengths that are too short. We use the $ABCD$ matrix method to determine the parameters of the propagation integrals.
needed to simulate the optical system. In Section 5 we give the results of processing images taken with the HST. We reconstruct both the unknown phase errors and the pupil function, which also differed from the design. We describe the preliminary results in deconvolving the effects of telescope jitter, which also blurs the images. We verify that the dust artifacts in the images originate from particles on a field flattener. In Section 7 we summarize our results and draw conclusions. Among our conclusions is our estimation that the spherical aberration is somewhat larger than that for which the correction in the replacement cameras is being designed.

2. Lower Bound on Estimating Polynomial Coefficients

The CR bound is an information-theoretical lower bound on the mean-squared error of an unbiased estimate of an unknown parameter. The CR bound for the retrieval of point-by-point phase maps has been computed previously. Here we calculate the CR bound on the mean-squared error of estimates of the coefficients of a polynomial expansion of the phase error. We begin by defining a model of the system, derive the lower bound, compute the bound, and draw conclusions.

A. System Model

In this section we employ a simplified model of the HST. The wave front \( F(u) \) in the detector plane is assumed to be the Fourier transform of the wave front \( f(x) \) in the aperture (pupil) plane, where \( u \) and \( x \) are each two-dimensional spatial coordinates. The wave front in the pupil is given by

\[
\theta(x) = \sum_{j=1}^{J} a_j Z_j(x). \tag{2.2}
\]

For the HST the commonly used phase polynomials are modified Zernike polynomials orthonormal over an annual aperture of inner radius that is 0.330 times the outer radius. They are given in Table 17 in Appendix A. The eleventh polynomial \( Z_{11} \) is the \( (r^4) \) spherical aberration term, which is the aberration in the HST of greatest interest; and the fourth, \( Z_4 \), is the \( (r^2) \) focus term, which can be varied in the optical system by moving the secondary mirror. We define the unknown coefficients as the vector \( \mathbf{a} = [a_1, \ldots, a_J]^T \), where \([\cdot]^T \) denotes the vector transpose. The measured data \( D(u) \) are modeled as independent, Poisson-distributed random variables with an expected value of

\[
E[D(u)] = I_a(u). \tag{2.3}
\]

Here the intensity, which is given in units of the average number of photons per detector, is

\[
I_a(u) = \alpha(u)[F_a(u)F_{a}^*(u) + I_B(u)] + \sigma^2, \tag{2.4}
\]

where \( \alpha(u) \) is a spatially varying scale factor that is determined by the light level, detector efficiency, nonuniform gain, etc. \( I_B(u) \) corresponds to the background intensity, possibly because of the detector dark current, preflash, or background light from the sky, and \( \sigma^2 \) is the variance associated with the CCD readout noise. The subscript \( a \) denotes that \( f_a(x), F_a(u), \) and \( I_a(u) \) are functions of the aberration coefficients \( \mathbf{a} \).

Since \( D(u) \), the number of photons detected at pixel \( u \), is a Poisson-distributed random variable with an expected value of \( I_a(u) \), it has the probability distribution

\[
Pr[D(u)] = \left[ I_a(u) \right]^{D(u)} \exp[-I_a(u)]. \tag{2.5}
\]

Furthermore \( D(u) \) and \( D(v) \) are assumed to be statistically independent for all \( u \neq v \). If we denote the entire measurement data as \( \{D(u)\}_{u \in U} \), where \( U \) is an \( N \times N \) array of pixels, then

\[
Pr[\{D(u)\}] = \prod_u Pr[D(u)]. \tag{2.6}
\]

where all products and summations are over \( u \in U \). The log-likelihood of the measured data, as a function of the unknown parameters \( \mathbf{a} \), is then

\[
L(\mathbf{a}) = \ln Pr[\{D(u)\}] = \sum_u \ln Pr[D(u)] = -\sum_u I_a(u) + \sum_u D(u) \ln I_a(u) - \sum_u D(u)!. \tag{2.7}
\]

Omitting terms that do not depend on \( \mathbf{a} \), we may rewrite the log-likelihood as

\[
L(\mathbf{a}) = -\sum_u I_a(u) + \sum_u D(u) \ln I_a(u). \tag{2.8}
\]

B. Cramer–Rao Lower Bound on the Estimation Error

We denote \( \hat{\mathbf{a}} \) as an estimator of the unknown parameter vector \( \mathbf{a} \). Since \( \hat{\mathbf{a}} \) depends on the measured data, it is a random vector. Suppose that the expectation of \( \hat{\mathbf{a}} \) is equal to the true parameter vector \( \mathbf{a} \). That is, \( \hat{\mathbf{a}} \) is an unbiased estimator of \( \mathbf{a} \). Now denote the error covariance matrix as

\[
\Sigma = \mathbb{E}[(\hat{\mathbf{a}} - \mathbf{a})(\hat{\mathbf{a}} - \mathbf{a})^T]. \tag{2.9}
\]

The CR inequality\(^{12}\) states that \( \Sigma = \mathcal{F}^{-1}(\mathbf{a}) \) is a nonnegative-definite matrix, where \( \mathcal{F}(\mathbf{a}) \) is the Fisher information matrix defined as

\[
\mathcal{F}(\mathbf{a}) = -\mathbb{E}[\mathcal{F}_\mathbf{a}(\mathbf{a})], \tag{2.10}
\]
and \( \mathcal{J}(a) \) is the observed Fisher information matrix:

\[
[\mathcal{J}(a)]_{ik} = \frac{\partial^2 L(a)}{\partial \alpha_i \partial \alpha_k},
\]

(2.11)

An important consequence of the CR inequality is that

\[
E[(\hat{a}_j - a_j)^2] \geq [\mathcal{J}^{-1}(a)]_{jj}.
\]

(2.12)

Now the partial derivative of \( L \) with respect to \( a_j \) is

\[
\frac{\partial L(a)}{\partial a_j} = \sum_u \frac{\partial I_u(a)}{\partial a_j} \left[ \frac{D(u)}{I_u(a)} - 1 \right].
\]

(2.13)

It then follows that

\[
[\mathcal{J}(a)]_{jk} = -E\{[\mathcal{J}(a)]_{jk}\}
\]  

(2.14)

where we have used the fact that \( E[D(u)] = I_a(u) \).

Since

\[
\frac{\partial I_u(a)}{\partial a_j} = \alpha(u) \left[ \frac{\partial F_a^*(u)}{\partial a_j} F_a(u) + \frac{\partial F_a(u)}{\partial a_j} F_a^*(u) \right],
\]

(2.15)

where \( \alpha \) is an inner product and \( \mathcal{I}_a[\cdot] \) denotes the discrete Fourier transform of its argument evaluated at \( u \), we have that

\[
\frac{\partial I_u(a)}{\partial a_j} = \alpha(u) \left( -[\mathcal{I}_a[Z_j f_a]] F_a(u) + [\mathcal{I}_a[Z_j f_a]] F_a^*(u) \right),
\]

(2.16)

where \( \text{Im}\{\cdot\} \) denotes the imaginary part of its argument.

C. Numerical Results

In this section the lower bound on the estimation error is evaluated numerically for some special cases. For these situations \( \alpha(u) \) was constant over \( U \) and \( I_b(u) \) and \( \alpha^2 \) were zero. Two pupil functions were used: aperture 1 shown in Fig. 1(a) and aperture 2 shown in Fig. 1(b). The array size is 256 \( \times \) 256 with aperture 1 having a diameter of 145 pixels and a central obscuration diameter of 60 pixels and aperture 2 having a diameter of 100 pixels and a central obscuration diameter of 33 pixels. For the computations the value of \( a_4 \) was permitted to vary while the remaining coefficients were set to the following values:

\[
a_5 = 0.01, \quad a_6 = -0.01, \quad a_{10} = 0.03, \quad a_{11} = 0.30, \quad a_7 = a_8 = a_9 = a_{22} = 0.00.
\]

In this section the coefficients are in units of radians rms of the wave-front error.

Fig. 1. Pupil functions used for the numeric computation of CR lower bounds on aberration estimates: (a) aperture 1, (b) aperture 2.
performed with much greater accuracy when \( a_4 \) is known.

As an example of how these tables might be used, suppose that all the coefficients are unknown, as is the case for the HST. This situation is addressed in Tables 1 and 2. Suppose that the \( N_{\text{max}} \) the maximum number of detected photons in any pixel, is 16,000 and that no pixels are saturated. Then, when Eq. (2.20) is used, for aperture 1 the lower bounds on the rms error of \( a_1 \) for various focus settings are as follows: 0.0166 for \( a_1 = 0.0 \), 0.0105 for \( a_1 = 1.0 \) and 0.0017 for \( a_1 = 3.0 \). From this we see that, for well-exposed, far out-of-focus PSF’s, the CR lower bound allows for acceptably low estimation errors, while the performance for PSF’s that are close to focus is marginal at best at this light level.

### 3. Computer Simulation Results

In this section we explore the accuracy with which phase-error polynomial coefficients can be estimated from HST PSF’s by means of the digital simulation of data and reconstruction experiments. We first show the importance of a multiple-plane diffraction model for imaging with the planetary camera (PC) mode of the wide-field/planetary camera (WF/PC) of the HST. Then we demonstrate the effect of imperfectly known system parameters. The section ends with the results of a blind test of our phase-retrieval algorithms.

#### A. Effect of Multiple-Plane Propagation

Figure 2 shows our model for the optical system for these simulations. The simplified model replaces mirrors with ideal thin lenses and eliminates several elements and folding mirrors. The main telescope, the optical telescope assembly (OTA), consisting of a positive and a negative lens (in reality a concave primary mirror and a convex secondary mirror), forms an image of a star in plane \( x_4 \). The image is formed near the surface of a four-faceted mirror in the shape of a pyramid, which divides the image plane into four quadrants. The pyramid facets also have optical power. A PC reimages each quadrant onto a CCD array in plane \( x_4 \) by a relay telescope, which is also depicted by two lenses. Only one of the quadrants and relay telescopes is shown in the figure.

The aberrated wave front, \( U_1(x_1) = m_1(x_1)\exp[i\theta(x_1)] \), in the input plane of the OTA is spatially limited by the transmittance function \( m_1(x_1) \), which includes the effect of the aperture diameter, the obscuration of the secondary mirror and the spiders (struts) holding it in place, and the three pads (bolts) on the primary mirror that hold it in place. On its way through the PC relay telescope, the wave front is multiplied by the transmittance function \( m_2(x_2) \), which represents the central obscuration and spiders of the relay telescope in plane \( x_3 \).

The relationship between the wave fronts in planes \( x_1 \) and \( x_2 \) is given by

\[
U_2(x_2) = \frac{1}{i\lambda B} \int U_1(x_1) \exp \left[ \frac{i\pi}{\lambda B} (Ax_1^2 - 2x_1x_2 + Dx_2^2) \right] dx_1
\]

and similarly for propagations between other pairs of planes. In the above \( x_1 \) and \( x_2 \) are both two-

### Tables

#### Table 1. Normalized Lower Bounds on \( E(\delta_j - a_j)^2 \) for Aperture 1 when \( a_4 \) is Known

<table>
<thead>
<tr>
<th>( j )</th>
<th>( a_4 = 0.0 )</th>
<th>( a_4 = 1.0 )</th>
<th>( a_4 = 3.0 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>96.460854</td>
<td>40.463634</td>
<td>1.013765</td>
</tr>
<tr>
<td>5</td>
<td>0.032564</td>
<td>0.021035</td>
<td>0.000830</td>
</tr>
<tr>
<td>6</td>
<td>0.047582</td>
<td>0.035003</td>
<td>0.000560</td>
</tr>
<tr>
<td>7</td>
<td>0.026248</td>
<td>0.011118</td>
<td>0.000346</td>
</tr>
<tr>
<td>8</td>
<td>0.029077</td>
<td>0.009690</td>
<td>0.000346</td>
</tr>
<tr>
<td>9</td>
<td>0.013022</td>
<td>0.018398</td>
<td>0.000153</td>
</tr>
<tr>
<td>10</td>
<td>0.013116</td>
<td>0.016509</td>
<td>0.000423</td>
</tr>
<tr>
<td>11</td>
<td>4.385449</td>
<td>1.764731</td>
<td>0.045521</td>
</tr>
<tr>
<td>12</td>
<td>0.019322</td>
<td>0.007839</td>
<td>0.000197</td>
</tr>
</tbody>
</table>

#### Table 2. Normalized Lower Bounds on \( E(\delta_j - a_j)^2 \) for Aperture 1 when \( a_4 \) is Known

<table>
<thead>
<tr>
<th>( j )</th>
<th>( a_4 = 0.0 )</th>
<th>( a_4 = 1.0 )</th>
<th>( a_4 = 3.0 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>Known</td>
<td>Known</td>
<td>Known</td>
</tr>
<tr>
<td>5</td>
<td>0.032412</td>
<td>0.020977</td>
<td>0.000829</td>
</tr>
<tr>
<td>6</td>
<td>0.047395</td>
<td>0.034598</td>
<td>0.000560</td>
</tr>
<tr>
<td>7</td>
<td>0.025872</td>
<td>0.010686</td>
<td>0.000345</td>
</tr>
<tr>
<td>8</td>
<td>0.028869</td>
<td>0.009541</td>
<td>0.000345</td>
</tr>
<tr>
<td>9</td>
<td>0.013019</td>
<td>0.018265</td>
<td>0.000414</td>
</tr>
<tr>
<td>10</td>
<td>0.013089</td>
<td>0.016304</td>
<td>0.000423</td>
</tr>
<tr>
<td>11</td>
<td>0.025298</td>
<td>0.010237</td>
<td>0.000368</td>
</tr>
<tr>
<td>12</td>
<td>0.000000</td>
<td>0.000000</td>
<td>0.000000</td>
</tr>
</tbody>
</table>

#### Table 3. Normalized Lower Bounds on \( E(\delta_j - a_j)^2 \) for Aperture 2 when \( a_4 \) is Known

<table>
<thead>
<tr>
<th>( j )</th>
<th>( a_4 = 0.0 )</th>
<th>( a_4 = 1.0 )</th>
<th>( a_4 = 3.0 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>Known</td>
<td>Known</td>
<td>Known</td>
</tr>
<tr>
<td>5</td>
<td>0.024147</td>
<td>0.053788</td>
<td>0.000745</td>
</tr>
<tr>
<td>6</td>
<td>0.024038</td>
<td>0.048589</td>
<td>0.000990</td>
</tr>
<tr>
<td>7</td>
<td>0.013642</td>
<td>0.007164</td>
<td>0.000171</td>
</tr>
<tr>
<td>8</td>
<td>0.013659</td>
<td>0.006924</td>
<td>0.000174</td>
</tr>
<tr>
<td>9</td>
<td>0.007838</td>
<td>0.009878</td>
<td>0.000153</td>
</tr>
<tr>
<td>10</td>
<td>0.006186</td>
<td>0.003008</td>
<td>0.000157</td>
</tr>
<tr>
<td>11</td>
<td>0.079072</td>
<td>1.784691</td>
<td>0.040705</td>
</tr>
<tr>
<td>12</td>
<td>0.004710</td>
<td>0.008210</td>
<td>0.000181</td>
</tr>
</tbody>
</table>

#### Table 4. Normalized Lower Bounds on \( E(\delta_j - a_j)^2 \) for Aperture 2 when \( a_4 \) is Known

<table>
<thead>
<tr>
<th>( j )</th>
<th>( a_4 = 0.0 )</th>
<th>( a_4 = 1.0 )</th>
<th>( a_4 = 3.0 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>Known</td>
<td>Known</td>
<td>Known</td>
</tr>
<tr>
<td>5</td>
<td>0.024147</td>
<td>0.053788</td>
<td>0.000745</td>
</tr>
<tr>
<td>6</td>
<td>0.024038</td>
<td>0.048589</td>
<td>0.000990</td>
</tr>
<tr>
<td>7</td>
<td>0.013642</td>
<td>0.007164</td>
<td>0.000171</td>
</tr>
<tr>
<td>8</td>
<td>0.013659</td>
<td>0.006924</td>
<td>0.000174</td>
</tr>
<tr>
<td>9</td>
<td>0.007838</td>
<td>0.009878</td>
<td>0.000153</td>
</tr>
<tr>
<td>10</td>
<td>0.006186</td>
<td>0.003008</td>
<td>0.000157</td>
</tr>
<tr>
<td>11</td>
<td>0.079072</td>
<td>1.784691</td>
<td>0.040705</td>
</tr>
<tr>
<td>12</td>
<td>0.004710</td>
<td>0.008210</td>
<td>0.000181</td>
</tr>
</tbody>
</table>
dimensional coordinates, $x_1x_2$ is interpreted as a dot (inner) product, and the integrals are over the interval $(-\infty, \infty)$ in each dimension. The coefficients $A, B,$ and $D$ are determined most conveniently by the $ABCD$ matrix method as described in Section 4. As can be seen from the second line of the equation above, this propagation can be performed with a single Fourier transform with the coefficient of the Fourier kernel being $1/(\lambda B)$.

In the digital computer we approximate the propagation given above by using the discrete Fourier transform:

$$U_2(x_2) = \exp(i\alpha_1x_2^2) \frac{1}{N} \sum_{x_1=0}^{N-1} U_1(x_1) \exp(i\beta_1x_1^2)$$

$$\times \exp(-i2\pi x_1x_2/N) = \exp(i\alpha_1x_2^2/2) [U_1(x_1) \exp(i\beta_1x_1^2)], \quad (3.2)$$

where $x_1$ and $x_2$ are now redefined to be integer sample (pixel) numbers for pixels separated by physical distances $\Delta x_1$ and $\Delta x_2$ for the two respective planes (which are assumed to be the same in each of the two dimensions). The $N \times N$ discrete Fourier transform is computed by using the fast Fourier transform (FFT) algorithm. Comparison of the discrete and continuous transforms reveals the following relationships:

$$\frac{1}{N} = \frac{\Delta x_1\Delta x_2}{\lambda B} \quad \text{or} \quad \Delta x_2 = \frac{\lambda B}{N\Delta x_1}, \quad (3.3)$$

$$\alpha_1 = \frac{\pi D\Delta x_2^2}{\lambda B}, \quad (3.4)$$

$$\beta_1 = \frac{\pi A\Delta x_1^2}{\lambda B}, \quad (3.5)$$

and similarly for the other propagations.

The requirements on sampling are as follows. To capture the OTA primary mirror, we need $N\Delta x_1 \geq 2.4$ m. To represent the OTA spiders, of 25.6-mm width, we need $\Delta x_1 \leq 25.6$ mm (assuming the gray-level masks as described below). To capture the spiders of 0.834-mm width in the relay telescope, we need $\Delta x_2 \leq 0.894$ mm. To match the CCD pixel spacing in the PC, we need $N\Delta x_3$ to exceed the sensible diameter of the PSF. The quadratic phase factors determined by the $A$ and $D$ parameters must not be aliased. The difference in a quadratic phase from one pixel to the next should not exceed $\pi$ radians:

$$\alpha_1 \left[ \frac{|N|^2}{2} - \frac{|N-1|^2}{2} \right] = \alpha_1(N-1) \leq \pi, \quad (3.6)$$

and similarly $\beta_1(N-1) \leq \pi$.

It is possible to accomplish the transformation from plane $x_1$ to plane $x_3$ with a single FFT. However, to satisfy the requirements above, we found that $N$ of $\sim 2000$ (2048 for an efficient FFT) would be required. This would be computationally expensive. By propagating from plane $x_1$ to plane $x_2$ and then propagating from plane $x_2$ to plane $x_3$, we can use much smaller FFT's. Furthermore we kept $\alpha_1(N-1) \leq \pi$ for the second propagation by keeping track of the quadratic phase term $\alpha_2\Delta x_3^2$ analytically for the second and third propagations. This is conceptually identical to inserting a thin lens of focal length $d$ just before plane $x_3$ and a thin lens of focal length $-d$ just after plane $x_3$. This pair of imaginary lenses cancel one another in terms of their effect on the optical system but allow the use of FFT's of smaller size in the digital propagations.

We can use a binary mask to represent the obscurations. However, for extremely fine features, such as the spiders, two or more samples across a spider would be necessary to represent it accurately. This
would make the sampling requirements more demanding. An alternative is to represent binary obstructions with gray-level transmittance functions. Think of each pixel as a square with the center at the sample value. If there is no obstruction within the square (it is all in the clear aperture), the transmittance of the sample is 1.0, and if the square is totally obscured, the transmittance of the sample is 0.0. For a partially obscured square the transmittance of the sample is the fraction of the square that the clear aperture covers. Transmittance functions \( m_i(x_i) \) made in this way should more accurately represent the obstructions than binary masks using the same sampling rate. Functions made in this fashion have the appearance of finer resolution and avoid much of the staircasing effects seen in binary representations. With the computational tricks of using gray-level transmittance functions, the additional propagation, and the imaginary pair of lenses, we were able to use FFT's of size \( N = 512 \) and even \( N = 256 \) with confidence.

For the model of the HST that uses three propagations including the imaginary pair of lenses at the plane of the PC obscurations \( x_3 \), most of the quadratic phase coefficients, \( \alpha_i \) and \( \beta_i \), are negligibly small, except for those in the first focal plane, in plane \( x_2 \). There the sum of all the quadratic phase terms is \( 2\pi BQ x_2^2 \), where the quadratic coefficient is given by

\[
BQ = \frac{1}{2\pi} \left( \alpha_1 - \frac{\pi \Delta x_2^2}{\lambda f_{py}} + \beta_2 \right) = \frac{D_{12}}{B_{12} - f_{py} + A_{23}} \Delta x_2^2 \frac{2\lambda}{\lambda^2},
\]

(3.7)

where \( \alpha_1, D_{12}, \) and \( B_{12} \) are from the propagation from plane \( x_1 \) to plane \( x_2; \beta_2, A_{23}, \) and \( B_{23} \) are from the propagation from plane \( x_2 \) to plane \( x_3; \) and \( f_{py} \) is the focal length of the pyramid in plane \( x_3 \). The units of \( BQ \) are waves per square pixel.

We explored the importance of the multiple-plane propagation model over a simpler single-FFT model of the optical system in a first set of simulation experiments. The values of the system parameters, as indicated in Fig. 2, for these simulations are given in Table 5. (Later work in retrieving the aberrations from HST data used updated parameters, as described in Section 4.) For the purpose of wave-front propagation the OTA secondary obstructions and the primary mirror all can be considered to be in the same plane, making the exact value of \( S_1 \) unimportant.

In these simulations it was taken to be zero, whereas in Section 4 it was taken to be 4.907 m. The values of \( A, B, C, \) and \( D \) for the three propagations, computed from the values in Table 5, are given in Table 6. We performed the simulations using \( N = 256 \) and a wavelength of \( \lambda = 0.889 \text{ m} \). The sample spacings and array widths for this set of simulations is given in Table 7. The aberration coefficients for the focus and spherical-aberration Zernike polynomials were assumed to be \( a_4 = -2.0 \) waves and \( a_{11} = -0.25 \) waves. Simulations were performed for two cases: (1) where OTA and WF/PC obscurations were both in their respective planes and (2) where all obscurations were in a single plane. Case 1 represents the more accurate, multiple-plane propagation simulation of the HST, and case 2 is essentially the single-Fourier-transform approximation that is used in the basic phase-retrieval algorithm. Figures 3(a) and 3(b) show, respectively, the OTA and PC obscurations used for case 1, and Fig. 3(c) shows the composite obscuration mask used for case 2. (Because of the number of FFT's involved, the OTA mask must be rotated by 180° for the multiple-plane propagation compared with the single-plane propagation for one to arrive at a PSF that is not rotated.) The field intensity \( |U_j(x_j)|^2 \) at the CCD camera after a plane wave is propagated through the HST model for case 1 is shown in Fig. 4(a). This PSF, called PSF\(_{\text{both}}\) (where the obscurations are in both planes), is difficult to distinguish visually from the PSF for case 2, PSF\(_{\text{OTA}}\) (where all obscurations are in the OTA plane). However, the differences caused by multiple-plane diffraction effects are evident in Fig. 4(b), which shows the difference image (PSF\(_{\text{both}}\) - PSF\(_{\text{OTA}}\)), where each PSF was normalized to have a peak value of unity prior to subtraction. The difference image has a mean value of \(-0.0002\) with a range of \([-0.067, 0.073]\). Another measure of the significance of multiple-plane diffraction effects is the absolute rms error:

\[
E = \left( \frac{1}{\sum_c^x} \sum_{x_4}^y |C(\text{PSF}_{\text{both}}) - \text{PSF}_{\text{OTA}}|^2 \right)^{1/2} \approx 0.085,
\]

(3.8)

where the constant \( C \) compensates for image-scaling differences. The 8.5% error observed is significant enough to effect the phase-retrieval estimate of the HST aberrations.

### Table 5. Parameters of the HST (PC) for Simulations

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Description</th>
<th>Value (mm)</th>
</tr>
</thead>
<tbody>
<tr>
<td>( S_1 )</td>
<td>OTA input plane to primary</td>
<td>Arbitrary</td>
</tr>
<tr>
<td>( f_1 )</td>
<td>OTA primary focal length</td>
<td>5520.0</td>
</tr>
<tr>
<td>( S_2 )</td>
<td>OTA element separation</td>
<td>4906.071</td>
</tr>
<tr>
<td>( f_2 )</td>
<td>OTA secondary focal length</td>
<td>-679.0</td>
</tr>
<tr>
<td>( S_3 )</td>
<td>OTA secondary to pyramid</td>
<td>6406.200</td>
</tr>
<tr>
<td>( f_{py} )</td>
<td>Pyramid focal length</td>
<td>1534.2</td>
</tr>
<tr>
<td>( d )</td>
<td>Pyramid to PC obscuration</td>
<td>895.35</td>
</tr>
<tr>
<td>( S_4 )</td>
<td>PC obscuration to primary</td>
<td>234.851</td>
</tr>
<tr>
<td>( f_3 )</td>
<td>PC primary focal length</td>
<td>249.840</td>
</tr>
<tr>
<td>( S_5 )</td>
<td>PC element separation</td>
<td>234.851</td>
</tr>
<tr>
<td>( f_4 )</td>
<td>PC secondary focal length</td>
<td>-112.380</td>
</tr>
<tr>
<td>( S_6 )</td>
<td>PC secondary to detector</td>
<td>364.441</td>
</tr>
</tbody>
</table>

### Table 6. ABCD Values for Simulations

<table>
<thead>
<tr>
<th>Propagation</th>
<th>( A )</th>
<th>( B ) (mm)</th>
<th>( C ) (mm(^{-1}))</th>
<th>( D )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( x_1 \rightarrow x_2 )</td>
<td>-1.3 \times 10^{-8}</td>
<td>57,600</td>
<td>-1.74 \times 10^{-5}</td>
<td>8.22544</td>
</tr>
<tr>
<td>( x_2 \rightarrow x_3 )</td>
<td>1</td>
<td>895.35</td>
<td>-0.0011169</td>
<td>0.0</td>
</tr>
<tr>
<td>( x_3 \rightarrow x_4 )</td>
<td>-3.096 \times 10^{-5}</td>
<td>1078.1</td>
<td>-0.0009278</td>
<td>2.27516</td>
</tr>
</tbody>
</table>

*The pyramid is handled separately.*
Table 7. Sample Spacings and Array Widths for Simulations
\(N = 256, \lambda = 0.689 \mu m\)

<table>
<thead>
<tr>
<th>Spacing (\Delta x_i) (mm)</th>
<th>Width (N/\Delta x_i) (mm)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>16.656</td>
</tr>
<tr>
<td>2</td>
<td>0.01201</td>
</tr>
<tr>
<td>3</td>
<td>0.25891</td>
</tr>
<tr>
<td>4</td>
<td>0.01446</td>
</tr>
</tbody>
</table>

We quantified the significance of multiple-plane diffraction further by performing phase retrieval with single-plane propagation on the two different PSF’s, fitting Zernike coefficients through \(a_{22}\). The gradient search algorithms for retrieving the phase coefficients are described in Ref. 13. The algorithms minimize the mean squared error:

\[
E = \sum_u W(u) \left| |G(u)| - |F(u)|^2 \right|
\]

where \(|F(u)|\) is the square root of the measured PSF, \(|G(u)|\) is the magnitude of the model of the aberrated wave front in the pupil digitally propagated to the PSF plane (the CCD plane), and \(W(u)\) is a weighting function (needed for the real data), which is zero wherever there are bad pixels or wherever the signal-to-noise ratio drops too low. When we report results, we give the normalized rms error

\[
\text{err} = \left( \frac{\sum_u W(u) |G(u)| - |F(u)|^2}{\sum_u W(u) |F(u)|^2} \right)^{1/2}
\]

For PSF\(_{\text{OTA}}\) the single-plane phase-retrieval algorithm yielded the correct Zernike coefficients to within 0.001 waves rms, since the system model used in the phase-retrieval algorithm matches that used to simulate the data. For the PSF\(_{\text{PSF}}\) both the single-plane phase-retrieval algorithm produced an estimate of \(-1.977\) waves for \(a_4\) (off by 0.023 wave), \(-0.2545\) wave for \(a_{11}\) (the magnitude of the spherical aberration was overestimated by 0.0045 wave), and up to 0.004 wave for the other coefficients (which all had true values of 0).

This is not a large error, but the single-plane phase-retrieval algorithm ultimately prevents us from obtaining a more accurate estimate of the coefficients. From Eq. (3.7) and Tables 6 and 7 the value of the quadratic phase factor, which determines the amount by which the OTA obscurations are out of focus at the plane of the PC obscurations, is \(BQ = 0.00005\), and this is the value used for the experiments described above. As discussed below this has approximately the same magnitude but the opposite sign compared with the true value for the HST, which was determined later, as described in Section 5. For this reason an algorithm using a single-plane propagation and a value of \(BQ\) with the correct (negative) sign might be expected to underestimate the magnitude of the spherical aberration for HST data.
A. Incorrect Spatial Scale and Obscuration Position

The plate scale refers to the relationship between the arcseconds of the angle in the sky and the number of pixels of the CCD detector. It can be derived from the $ABCD$ matrix values describing the optical system, but it can also be determined independently by measuring the distances between the images of stars of known separation. The plate scale also affects the spatial scale in the pupil function that we model, which is related to the spatial scale in the CCD plane by equations of the form $\Delta x_2 = \lambda B/(N\Delta x_1)$. The correct scale factors were known in practice only approximately, and they vary as a function of the location of the PSF on the CCD chip in a way that was not completely characterized at the time of this work. Brewer later performed ray tracing to satisfy this need, as described in Section 4. We determined the sensitivity to the plate scale by simulating data using one spatial scale factor and using an incorrect scale factor in the retrieval algorithm. For these simulations the phase-retrieval algorithm used multiple-plane propagation, and the (incorrectly) assumed scale factors were independently changed for the OTA obscurations and the PC obscurations, or the PC obscurations were translated to simulate an unknown shift of the PC obscuration.

Selected results of these simulation experiments are summarized as follows. If the scales of both sets of obscurations were changed by up to $\pm 3\%$, the error in $a_{11}$ was up to 0.004 wave (assuming that the quadratic phase $BQ$ was correct). As will be seen below, with the real data the sensitivity to the plate scale was considerably greater. When a single parameter was changed, an error in $a_{11}$ of 0.01 wave rms (which would be of considerable concern) was caused by (1) an OTA spatial scale error of $+4\%$ or (2) a PC spatial scale error of $\pm 6\%$. A PC shift of $\pm 2$ pixels caused an error in $a_{11}$ of only $\pm 0.002$ wave. In all the cases above the normalized rms error of the fit, given by Eq. (3.10), was in the range of 0.09–0.12.

It is unlikely that spatial scale errors exceeding 1% will occur when our best estimates of the actual system parameters are used; therefore spatial scale errors should not ultimately be a limiting factor in estimating $a_{11}$. However, for determining the other aberrations a spatial scale error of 1% could be significant. In particular a retrieval where the scale of the OTA is too small causes little change in $a_{11}$, but a large change in $a_{22}$, which could be of concern. It does not appear to be likely that shifts of the PC obscurations will be the limiting factor in estimating $a_{11}$ either as long as we use our best estimates of the actual shifts. However, for determining other aberrations, in particular asymmetric aberrations such as coma, we expect a greater sensitivity to shifts in the PC obscurations.

The results above are for the case of knowing a priori the correct value of the quadratic phase factor $BQ$ used in the multiple-plane propagation. If $BQ$ is not known exactly, its value should be optimized along with the plate scale and the aberration coefficients. For the simulations described above, the value of $BQ$ was 0.00005. In another set of simulation experiments a range of values of the plate scale was assumed during different applications of the phase-retrieval algorithm, and multiple values of $BQ$ were used for each case to determine the value of $BQ$ that minimized the fitting error for each of the plate scales. Then, with the optimum values of $BQ$ for each plate scale, the aberrations were reoptimized. The results are given in Table 8. The values of the estimated $a_{11}$ are plotted in Fig. 5 as a function of the plate scale for both cases: the true value of $BQ$ was used and when the optimum value of $BQ$ was used. The estimated $a_{11}$ varied much more for the optimum value of $BQ$ (which is optimum in the sense that it minimizes the error metric for a given assumed plate scale) than for the true value of $BQ$. When the optimum value of $BQ$ was used, the estimated $a_{11}$ decreased in magnitude as the value of the plate scale was underestimated.

B. Blind-Test Results

J. Holtzman (Lowell Observatory, Flagstaff, Ariz.) simulated PSF's with realistic amounts of noise and aberrations for a variety of focus settings and distributed the data for blind tests through several phase-retrieval groups.

We processed only the data with a spectral filter corresponding to a wavelength of 889 nm, since at only that wavelength was the Nyquist simulated data Nyquist sampled for the optical fields. The data were crudely filtered by zeroing out the data outside a circle beyond which the signal-to-noise ratio appeared (by eye) to decrease below unity. These results were obtained early in our effort, before we had implemented a weighting function in the phase-retrieval algorithm. All reconstructions reported here used $N \times N = 256 \times 256$ arrays for the FFT's. Limited reconstructions with $512 \times 512$ arrays were also performed, and those results differed little from the $256 \times 256$ case. The pupil function used was the one circulated in late August 1990 but with the PC secondary obscurations moved 2 pixels in both dimensions (i.e., diagonally). This pupil shift was determined by the iterative transform algorithm. Our results are summarized in Table 9, which shows recovered Zernike coefficients $a_i$ for PSF's simulated for five different (unknown) focus settings, designated $A$–$E$.

<table>
<thead>
<tr>
<th>Root-Mean-Squared Error</th>
<th>Optimum $BQ$</th>
<th>$a_4$</th>
<th>$a_{11}$</th>
<th>$a_{22}$</th>
<th>Value of Scale $s$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.119</td>
<td>0.000117</td>
<td>-2.435</td>
<td>-0.292</td>
<td>-0.0011</td>
<td>3% Too small</td>
</tr>
<tr>
<td>0.000</td>
<td>0.00005</td>
<td>-2.500</td>
<td>-0.300</td>
<td>-0.0020</td>
<td>Exact</td>
</tr>
<tr>
<td>0.064</td>
<td>0.000031</td>
<td>-2.527</td>
<td>-0.300</td>
<td>-0.0014</td>
<td>1% Too large</td>
</tr>
<tr>
<td>0.116</td>
<td>-0.000014</td>
<td>-2.566</td>
<td>-0.304</td>
<td>-0.0023</td>
<td>3% Too large</td>
</tr>
<tr>
<td>0.141</td>
<td>-0.000069</td>
<td>-2.568</td>
<td>-0.313</td>
<td>-0.0097</td>
<td>5% Too large</td>
</tr>
</tbody>
</table>
Fig. 5. Estimated spherical aberration \( a_{11} \) as a function of assumed spatial scale \( s \) for simulated data. The solid curve shows the quadratic phase factor \( BQ \) optimized for the given spatial scale; the dashed curve indicates when the true value of \( BQ \) is used.

From Table 9 we see that the only coefficients for which the mean value is significantly greater than the standard deviation among the results are (with their mean values) \( a_7 = -0.014 \), \( a_8 = 0.005 \), and \( a_{11} = -0.227 \) (and \( a_8 \) differs from 0 by only 2\( \sigma \), and thus our confidence level in \( a_8 \) was not high). All these values agree well with the actual values given in the last column, which were revealed to us at a later date. Note that for the nominally in-focus data, given by image \( C \), the value of spherical aberration \( a_{11} \) is farthest from the true value and is substantially underestimated. This is consistent with the results obtained with CR lower bounds and is one of the reasons why we prefer PSF's that are far out of focus.

4. Parameters for the HST

In this section we discuss the specific parameters of the HST that are important to the phase-retrieval algorithms.

A. Sampling Requirement

To avoid aliasing,

\[ p_s < \frac{\lambda}{2D}, \]

whereas for optical fields the sampling requirement is milder:

\[ p_s < \frac{\lambda}{D}. \]

The latter requirement is pertinent to the phase-retrieval algorithms that we used, since we must digitally propagate optical fields back and forth within the system. Phase-retrieval algorithms that require only a forward propagation through the system do not have this restriction. Table 10 shows the as-designed sample spacings (in arcseconds and in microradians) of the detector arrays for the various cameras in the HST and the shortest wavelengths for which the optical intensity and optical field are adequately sampled at the detector. For both the wide-field camera (WFC) and the PC, the center wavelength of the narrow-band spectral filter with the longest wavelength is 889 nm. From Table 10 we see that even the optical field at this wavelength is undersampled in the WFC. Consequently we did not use any data from the WFC. There were suitable spectral filters for the faint-object camera (FOC). However, the images from the FOC are quite noisy compared with those from the WFC and PC since the FOC is count-rate limited. For these reasons we restricted our attention to PC images taken through narrow-band filters with wavelengths above 500 nm, which are the most suitable for characterizing the aberrations of the HST.

When a single-FFT model of the optical system is used, if we compute an FFT of the field in the detector plane that has angular sampling interval \( p_s \), the resulting field in the aperture plane has a physical width of

\[ D_a = \frac{\lambda}{p_s}. \]

Therefore, for an array in the computer of width \( N \) pixels, the scale factor giving the number of pixels per meter of the physical aperture is

\[ s = \frac{N}{D_a} = \frac{Np_s}{\lambda}, \]

where \( Np_s \) is the angular width (projected onto the sky) of the array in the detector plane over which the

<p>| Table 9. Zernike Phase Coefficients Estimated in a Blind Test |
|-------------------|---|---|---|---|---|---|---|</p>
<table>
<thead>
<tr>
<th>( j )</th>
<th>( A )</th>
<th>( B )</th>
<th>( C )</th>
<th>( D )</th>
<th>( E )</th>
<th>( s )</th>
<th>Mean</th>
<th>Actual</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>-2.089</td>
<td>-1.038</td>
<td>0.004</td>
<td>1.051</td>
<td>2.104</td>
<td>N/A</td>
<td>N/A</td>
<td>various</td>
</tr>
<tr>
<td>5</td>
<td>0.002</td>
<td>0.003</td>
<td>-0.004</td>
<td>-0.002</td>
<td>0.001</td>
<td>0.002</td>
<td>0.000</td>
<td>0.001</td>
</tr>
<tr>
<td>6</td>
<td>-0.009</td>
<td>-0.004</td>
<td>0.009</td>
<td>-0.010</td>
<td>0.009</td>
<td>0.008</td>
<td>-0.001</td>
<td>0.007</td>
</tr>
<tr>
<td>7</td>
<td>-0.010</td>
<td>-0.013</td>
<td>-0.013</td>
<td>-0.015</td>
<td>-0.018</td>
<td>0.003</td>
<td>-0.014</td>
<td>-0.012</td>
</tr>
<tr>
<td>8</td>
<td>0.004</td>
<td>0.008</td>
<td>0.002</td>
<td>0.005</td>
<td>0.008</td>
<td>0.002</td>
<td>0.005</td>
<td>0.004</td>
</tr>
<tr>
<td>9</td>
<td>0.004</td>
<td>0</td>
<td>-0.001</td>
<td>0.001</td>
<td>0</td>
<td>0.002</td>
<td>0.001</td>
<td>0</td>
</tr>
<tr>
<td>10</td>
<td>0.005</td>
<td>-0.002</td>
<td>0</td>
<td>-0.001</td>
<td>0</td>
<td>0.002</td>
<td>0.001</td>
<td>0</td>
</tr>
<tr>
<td>11</td>
<td>-0.227</td>
<td>-0.225</td>
<td>-0.209</td>
<td>-0.235</td>
<td>-0.241</td>
<td>0.011</td>
<td>-0.227</td>
<td>-0.236</td>
</tr>
<tr>
<td>22</td>
<td>0</td>
<td>0</td>
<td>0.002</td>
<td>0.001</td>
<td>0</td>
<td>0.001</td>
<td>0.001</td>
<td>0</td>
</tr>
<tr>
<td>err</td>
<td>0.164</td>
<td>0.209</td>
<td>0.341</td>
<td>0.216</td>
<td>0.171</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

| Table 10. Sample Spacings for HST Cameras and Wavelengths for Nyquist Sampling |
|-------------------------------|---|---|---|
| Camera | \( \rho_s \) (arcsec) | \( \rho_s \) (μrad) | \( \lambda \) (nm) |
| PC | 0.043 | 0.2085 | 1000 | 500 |
| WFC | 0.100 | 0.4848 | 2333 | 1667 |
| FOC/48 | 0.044 | 0.2133 | 1023 | 512 |
| FOC/96 | 0.022 | 0.1067 | 512 | 256 |
| FOC/288 | 0.007 | 0.0339 | 163 | 081 |

1 April 1993 / Vol. 32, No. 10 / APPLIED OPTICS 1755
FFT is performed. The width \( D_0 \) in pixels of the aperture of the OTA is given by (2.4 m) s. \( D_0 \) should be large enough to model accurately the features of the aperture function including the spiders and pads. Since the \( D_0 \) is \( \sim 100 \) times the width of a spider and at least 1 or 2 pixels are needed to represent the width of a spider, we prefer to have \( D_0 > 100 \) or 200 pixels.

Table 11 lists the values of \( s \) and \( D_0 \) for all the cameras on the HST, for some of the wavelengths for which there are narrow-band filters, and for \( N = 256 \).

B. \( ABCD \) Matrix Calculation of the Propagation Parameters

The most convenient way to determine the parameters of the digital propagation of wave fronts through the optical system is the \( ABCD \) matrix approach. A paraxial description of any optical system that has no obscurations or vignetting and that may contain multiple elements including lenses, mirrors, and spacings is given by a \( 2 \times 2 \) matrix of four values, \( A, B, C, \) and \( D \), where

\[
M = \begin{bmatrix} A & B \\ C & D \end{bmatrix}.
\]

Each element of the system is also represented by a \( 2 \times 2 \) matrix, and a system described by elements \( M_1, M_2, \ldots, M_n \) is given by the matrix product

\[
M = M_n \cdots M_2 M_1.
\]

The propagation of a wave front through the system is then given by Eq. (3.1). Note that only \( A, B, \) and \( D \) (which together define \( C \)) are needed. One method to determine the \( ABCD \) values for a system would be to compute the matrix product given above, given the parameters of the system that determine the matrix for each element and spacing.

A second approach to determining the \( ABCD \) values is to use the results of a ray-tracing computer program. A ray at the input plane of the system with height (the distance from the optical axis) \( y_k \) and slope \( v_k \) is transformed by the system to have height \( y'_k \) and slope \( v'_k \) at an output plane:

\[
\begin{bmatrix} y'_k \\ v'_k \end{bmatrix} = \begin{bmatrix} A & B \\ C & D \end{bmatrix} \begin{bmatrix} y_k \\ v_k \end{bmatrix}.
\]

Values of the ray heights and the slopes for two rays traced by Brewer with a paraxial ray trace with the PAGOS program are given in Table 12. In Ref. 14, \( y_1, v_1, y_2, \) and \( v_2 \) are denoted as \( \beta, b, \alpha, \) and \( a \). Plane \( x_2 \) was taken to be Brewer’s surface 18, which is in the vicinity of the image plane at the pyramid, at the apex of the pyramid, after interaction with the optical power of the pyramid. The \( ABCD \) values that we computed from these rays by using the equations above are shown in Table 13. Included are the values for the case of the optical system including the imaginary pair of lenses immediately before and after plane \( x_3 \). These propagations are indicated by \( x_2 \rightarrow x_3' \) and \( x_3 \rightarrow x_4 \). Note that the coefficient \( D \) for \( x_2 \rightarrow x_3 \) and \( A \) for \( x_3 \rightarrow x_4 \), which are proportional to the quadratic phase factors in the propagations, are greatly reduced. The last row is for the entire HST that would be used for a single-FFT propagation through the system. The negative sign of \( B \) in this latter case is indicative of a rotation of coordinates in plane \( x_1 \), which is necessary for the single-FFT propagation relative to the multiple-plane propagation.

The key parameters desired from the system analysis for use in our 3-FFT model of the HST OTA + PC are (1) the spatial scale factors relating the four planes of interest and (2) the total quadratic phase factor in the plane \( x_2 \).

Now we compare the plate scale that we determined by using the \( ABCD \) approach with that given by differential ray tracing. When the reported pixel size of 15.24 \( \mu \text{m} \) (reported earlier as 15.0 \( \mu \text{m} \)) at the PC CCD is used, according to this prescription, the
For the PSF's near the center of the CCD, starting with $\Delta x_1 = 0.01524$ mm and working backward, using Eq. (3.3) and the values of $B$ in Table 13, we obtain the spatial scales shown in Table 14, which are evaluated for the particular case of an FFT of length $N = 256$ and $\lambda = 589$ nm. The sample spacing $\Delta x_1 = 16.23$ mm in the input plane is the same whether one or three FFT's is used for the propagation. However, this equality depends on the choice of the locations of the intermediate planes. For example, if plane $x_2$ is taken to be Brewer's surface 20, which is just 20 mm beyond the apex of the pyramid (closer to the point where the chief ray intersects the pyramid), the multiple-plane propagation requires $\Delta x_1 = 16.04$ mm, a difference of 1.16%, which is significant. These parameters were calculated after the retrieval experiments reported in Section 5 were completed. Fortunately we obtained some of our phase-retrieval results while using parameters that were close to these parameters.

The other important parameter is the total quadratic phase factor in the plane $x_2$, which is given by Eq. (3.7). However, the optical power of the pyramid is already included in the matrix coefficient $D$ for the propagation $x_1 \to x_2$, and so we have

$$BQ = \frac{D_{12} \Delta x_2}{2 \lambda B_{12}} + \frac{A_{23} \Delta x_2^2}{2 \lambda B_{23}},$$

which, when the values in Tables 13 and 14 are used, is $5.04 \times 10^{-5}$ waves/pixel$^2$. However, the negative of this number is appropriate because of the sign convention issue discussed next.

C. Sign Conventions for the Phase and Pupil Coordinates

For a period of our effort there was a discrepancy between the sign of the optimum value of the quadratic phase factor $BQ$ used in the multiple-plane diffraction algorithm and that predicted by an analysis of the HST optical system design. We determined that this discrepancy was due to inconsistent sign conventions when the phase of a wave front was described. Two different sign conventions are commonly used for the phase of a wave front. The phase of an expanding spherical wave front is given by

$$\phi(x, y, z) = \frac{2\pi}{\lambda} (x^2 + y^2 + z^2)^{1/2},$$

in Goodman but is given by

$$\phi(x, y, z) = \frac{2\pi}{\lambda} (x^2 + y^2 + z^2)^{1/2}$$

in Born and Wolf but is given by

$$\phi(x, y, z) = \frac{2\pi}{\lambda} (x^2 + y^2 + z^2)^{1/2}$$

in Born and Wolf but is given by
the negative of this expression in Siegman. For our analysis we adopted the former positive sign convention. With this convention the phase of a wavefront in a given plane is greater at points where the wavefront has gone through a greater optical path. Then the transforms are those given in Section 3. We refer to this sign convention as the positive convention and Siegman’s as the negative convention.

Our computer simulations use the positive sign convention described above. Other groups characterizing the HST use the negative convention, for which the spherical aberration has a negative sign. To obtain a spherical aberration with a negative sign, we rotated the pupil function 180° with respect to the measured PSF. Therefore we in effect replaced the aberrated pupil function \( f(x) \) by \( f^*(-x) \). If the Fourier transform of \( f(x) \) is \( F(u) \), the Fourier transform of \( f^*(-x) \) is \( F^*(u) \). Since the measured data are \( |F(u)|^2 = |F^*(u)|^2 \), we can arrive at a consistent solution with the use of the positive sign convention. This same effect makes it necessary to negate the value of the quadratic phase factor, \( BQ \), which is discussed above, in the computer software compared with the theory.

To interpret correctly the coefficients other than \( a_4 \), \( a_{11} \), and \( a_{22} \) in the results given below, it is necessary to establish the orientation of the coordinates. The convention is as follows: the Zernike polynomials we used for Planetary Camera 6 (PC-6) are defined with the \( x \) coordinate in the \(-v_2\) direction and the \( y \) coordinate in the \(+v_3\) direction, where \( v_2 \) and \( v_3 \) are directions defined on the HST. This was determined by comparing our reconstructed pupil functions with the designed pupil functions. In our software \( x \) is horizontal, left to right, and \( y \) is vertical, top to bottom. Therefore to compare these results with those reported by others using a different convention, the appropriate translation of coefficients is necessary. For multiple-plane propagation, which requires three FFT’s to propagate to the image plane, we rotate the entrance pupil by 180° relative to its orientation for the single-FFT propagation case. (Thus the obscurations of the OTA are rotated; however, the obscurations of the PC have the same orientation for both cases.) This is necessary since the first two FFT’s serve to rotate the entrance pupil by 180°. However, in our code we do not change the Zernike coefficients to accommodate this effect. Therefore when reporting our Zernike fitting results from multiple-plane propagation, we negate the coefficients of polynomials 2, 3, 7-10, and 16-21 to retain the convention that \( x = -v_2 \) and \( y = +v_3 \).

### 5. Results with HST Data

Considerable effort went into processing data from the HST over a period of several months, during which time the quality of both the data and the phase-retrieval algorithms improved. Since the earlier results are not judged to be as accurate as the later results, we report here only some of the later results from the Hubble Aberration Recovery Program (HARP) 1A and 1B collections. In this section we give the results of phase retrieval, of optimizing over some unknown system parameters, and of pupil reconstruction.

We used the retrieval algorithms described in Ref. 13. Given a description of the pupil function we retrieved the aberrations by using first a gradient search algorithm to estimate a smooth polynomial approximation to the phase error and then an iterative propagation algorithm to retrieve a detailed point-by-point phase map.

However, the optical axis of the WF/PC was not aligned as intended with the OTA, and this resulted in a combined pupil function that was significantly different from the design. Had there been no aberrations, this misalignment would have had little effect on the system performance; but it did have a substantial effect on the attempt to characterize the telescope. Given an estimate of the aberrations, we estimated the pupil function by using an iterative propagation.

### Table 13. ABCD Matrix Values

<table>
<thead>
<tr>
<th>Propagation</th>
<th>( A )</th>
<th>( B ) (mm)</th>
<th>( C ) (mm(^{-1}))</th>
<th>( D )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( x_1 \rightarrow x_2 )</td>
<td>-0.0002206</td>
<td>57,292.5</td>
<td>-1.73394 \times 10^{-5}</td>
<td>-29.8421</td>
</tr>
<tr>
<td>( x_2 \rightarrow x_3 )</td>
<td>1.0</td>
<td>895.283</td>
<td>0.0</td>
<td>1.0</td>
</tr>
<tr>
<td>( x_2 \rightarrow x_4 )</td>
<td>1.0</td>
<td>895.283</td>
<td>-0.00111696</td>
<td>0.0</td>
</tr>
<tr>
<td>( x_3 \rightarrow x_4 )</td>
<td>-1.24307</td>
<td>1,113.33</td>
<td>-0.0192194</td>
<td>16.4089</td>
</tr>
<tr>
<td>( x_4 \rightarrow x_3 )</td>
<td>-0.000474674</td>
<td>1,113.33</td>
<td>-0.000891212</td>
<td>16.4089</td>
</tr>
<tr>
<td>( x_1 \rightarrow x_4 )</td>
<td>0.000266858</td>
<td>-71,231.5</td>
<td>1.60748 \times 10^{-5}</td>
<td>-1077.32</td>
</tr>
</tbody>
</table>

*Propagations involving \( x_4 \) include the imaginary pair of lenses before and after the PC obscurations. The pyramid is included in propagation \( x_1 \rightarrow x_2 \).

### Table 14. Example Evaluation of Spatial Scale Factors for \( \lambda = 889 \text{ nm} \) and \( N = 256 \)

<table>
<thead>
<tr>
<th>Plane</th>
<th>In terms of ( \Delta x_{\text{prev}} )</th>
<th>In terms of ( \Delta x_4 )</th>
<th>( \Delta x_4 ) (mm)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CCD ( x_4 )</td>
<td>( \lambda B_{34} )</td>
<td>( \lambda B_{34} )</td>
<td>0.01524</td>
</tr>
<tr>
<td>PC obscuration ( x_3 )</td>
<td>( N \Delta x_4 )</td>
<td>( N \Delta x_4 )</td>
<td>0.2537</td>
</tr>
<tr>
<td>Pyramid ( x_2 )</td>
<td>( \lambda B_{34} )</td>
<td>( B_{34} \Delta x_4 )</td>
<td>0.01226</td>
</tr>
<tr>
<td>OTA pupil ( x_1 )</td>
<td>( N \Delta x_2 )</td>
<td>( N B_{23} \Delta x_4 )</td>
<td>16.23</td>
</tr>
<tr>
<td>OTA pupil ( x_1 ) (single-FFT case)</td>
<td>( \lambda B_{14} )</td>
<td>( N \Delta x_4 )</td>
<td>16.23</td>
</tr>
</tbody>
</table>
A difference in all of 0.01

HARPIA PC-6 F889N2P2, which was taken with

The most effort was put into the image designated

A. Aberration Estimation

HST to be accurate to within 0.01 μm rms. Initially the goal was to know all for the

coefficients of the 11th Zernike polynomial. The relationship between all, the

multiple-plane propagation algorithm: the first is for

Because the pupil functions are needed to retrieve

was far from the center of the CCD, where it should have been.

two sets of these first two steps were necessary. With this improved estimate of the pupil function, we could then use the iterative propagation algorithm to

with this estimate of the phase we reconstructed the

position of the PC obscurations to estimate its pupil function. We then repeated these first two steps until no further improvement was made. Usually

coefficient of spherical aberration, and the conic

two sets of these first two steps were necessary. With this improved estimate of the pupil function, we could then use the iterative propagation algorithm to estimate a point-by-point phase map by using the polynomial approximation to the aberrations as the initial estimate. The result of most concern was the coefficient of the 11th Zernike polynomial, which is the spherical aberration that would be corrected in the cameras that would replace the existing cameras in the HST. Initially the goal was to know \( a_{11} \) for the HST to be accurate to within 0.01 μm rms of the wave-front error. The relationship between \( a_{11} \), the coefficient of spherical aberration, and the conic constant \( \kappa \) on the primary mirror of the OTA that would produce that aberration is

\[
\kappa = -1.0023 + 0.043841 a_{11}. \tag{5.1}
\]

A difference in \( a_{11} \) of 0.01 μm is equivalent to a difference in \( \kappa \) of 0.00044.

A. Aberration Estimation

The most effort was put into the image designated HARP1A PC-6 F889N.P2, which was taken with

PC-6 through the narrow-band filter with a center wavelength of 889 nm. It was the most useful of the HARP1A images. It is well out of focus (focus parameter = -260 μm), which is preferred because, without a bright central spike, most of the pixels of the PSF can have a large number of photons without saturating the CCD, i.e., they have a relatively large signal-to-noise ratio. Many other images were available, but, being closer to focus, they were much less suitable for phase retrieval. Our approach was to develop our algorithms while concentrating on a small number of the best images, in which results we could have high confidence, rather than devoting a large amount of effort to a larger number of images of poor quality in which little confidence could be placed.

For these data, by using the single-plane diffraction algorithm, we arrived at a value of \( a_{11} = -0.28 \) μm rms earlier in the effort. Then later, by using the multiple-plane diffraction algorithm, we obtained \( a_{11} = -0.295 \) μm, a change of -0.015 μm. Still later, not yet knowing the most accurate estimate of the system parameters, we optimized the error metric over the poorly known system parameters and found the optimized parameters to be the plate scale = 0.0442 arcsec/pixel and the quadratic phase factor \( BQ = -0.000054 \). Using these optimized parameters, we arrived at \( a_{11} = -0.299 \) μm rms. Later still, we found from Brewer's ray-tracing results that these optimized parameters were close to the true parameters. Therefore our best estimate of \( a_{11} \) from the HARP1A imagery is \( a_{11} = -0.299 \) μm rms. These results are summarized in Table 15. The first column under P2 is for a single-plane propagation algorithm (only the largest values for \( a_{12} a_{21} \) are shown). Two values for the coefficients are given for the multiple-plane propagation case: the first is for fitting \( a_2 = a_{11} \) and \( a_{22} \), and the second is for fitting \( a_2 = a_2 \). (Only the value for \( a_{16} \) is listed among \( a_{12} a_{21} \), since the other values were negligibly small.) Note that for this real data the magnitude of \( a_{11} \)

\[
\begin{array}{cccccc}
 j & P2 Single & P2 Multi (11) & P2 Multi (22) & New Parameter & Q2 Single  \\
 4 & -2.212 & -2.227 & -2.223 & -2.306 & 0.73  \\
 5 & -0.018 & -0.003 & 0.006 & -0.003 & 0.06  \\
 6 & -0.025 & 0.025 & 0.026 & 0.031 & -0.02  \\
 7 & 0.004 & 0.001 & 0.005 & -0.001 & 0.01  \\
 8 & 0.017 & 0.010 & 0.009 & 0.013 & 0.06  \\
 9 & -0.022 & -0.020 & -0.009 & -0.020 & -0.00  \\
 10 & 0.002 & 0.008 & 0.010 & 0.005 & 0.01  \\
 11 & -0.280 & -0.292 & -0.295 & -0.299 & -0.281  \\
 12 & 0.008 & (n/a) & (n/a) & (n/a) & 0.01  \\
 16 & -0.009 & (n/a) & -0.004 & (n/a) & 0.01  \\
 20 & 0.006 & (n/a) & (n/a) & (n/a) & 0.00  \\
 22 & 0.005 & 0.006 & 0.007 & 0.008 & 0.04  \\
 \text{Conic } \kappa = & -1.0146 & -1.0151 & -1.0152 & -1.0154 & -1.0146  \\
 \text{Root-mean-squared error} = & 0.1583 & 0.1352 & 0.1353 & 0.1428 & 0.2508  \\
\end{array}
\]

Table 15. Zernike Coefficients (Micrometer rms Wave-Front Error) for HARP1A Images PC-6 F889N.P2 and PC-6 F889N.Q2

1 April 1993 / Vol. 32, No. 10 / APPLIED OPTICS 1759
increased when multiple-plane propagation was used, whereas for the simulated data described in Section 3 the magnitude of \( a_{11} \) decreased with the multiple-plane propagation algorithm. The values of \( a_2-a_{10} \) changed substantially with the multiple-plane propagation algorithm but stayed below 0.03 \( \mu \)m rms in magnitude; \( a_{22} \) varied with the reconstruction, staying below 0.01 \( \mu \)m rms (which, however, could be significant). The column labeled New Parameter shows the values obtained when a larger value of the plate scale (0.0442 arcsec/pixel) and a different value of the quadratic phase coefficient \( BQ(-0.000054) \) were used. These values represent what is now thought to be a more accurate system model. Again, all the values of the Zernike coefficients change somewhat. Notably \( a_{11} \) increased in magnitude to -0.299, corresponding to the conic constant on the primar mirror of the OTA of -1.0154. The error metric is somewhat larger than that obtained for the old parameters since the pupil shift used in both cases was that optimized for the old parameters. The values other than \( a_4 \) and \( a_{11} \) are considered to be unreliable at this point, since they change so much depending on the details of the retrieval algorithm for a given data set. These differences show the importance of modeling the system as accurately as possible.

Also shown in the last column of Table 15 is an example of the results for another image, PC-6 F889N.Q2, which was closer to focus and less reliable (as seen from the value of 0.2508 for the rms error of the fit). Since the results for PC-6 F889N.P2 produce a fit that is so much better than the other images, we tend to ignore the results from the other HARP1A images.

There seems to be a trend toward larger magnitudes of \( a_{11} \) as the accuracy of the modeling increases. This being the case we suspect that the average value of \( a_{11} \) reported by all the groups working on phase-retrieval underestimates the magnitude of the true value.

Point-by-point phase maps have also been reconstructed with the iterative propagation algorithm, which shows the fine structure (sometimes referred to as zones) in the mirror surfaces. However, the reliability of these detailed phase maps has not yet been established and are therefore not included here.

Figure 7 compares a PSF computed from a model of

![Image of two PSF comparisons](image)

Fig. 7. Measured image PC-6F889N.P2 from HARP1A and the images computed from it. Measured PSF (upper left), the PSF deconvolved with the Ayers/Dainty algorithm (upper right), the PSF deconvolved with the Wiener filter by using jitter data (lower left), and the PSF computed from the model by using a polynomial phase estimate (lower right).
the aberrated system with the measured PSF. The computed PSF is for the case of single-plane propagation and when only the polynomial approximation of the phase error is used. As is discussed in Subsection 5.D, one of the largest factors that causes the computed PSF to differ from the measured PSF is the jitter in the pointing of the telescope, which smears out the measured PSF.

B. Optimizing over System Parameters

Before knowing the correct system parameters (which were later supplied by ray tracing performed by Brewer as described in Section 4), we performed a search over the plate scale and quadratic phase factor that minimized the error metric. Since these two variables were not included in our optimization software, we optimized them in a brute force fashion by selecting many different combinations of their values, by optimizing for the Zernike coefficients for each combination, and by selecting the combination that yielded the smallest error metric. The optimization would have to be redone for any other PSF in a different location in the field of view or in a different wavelength band. (Automatic optimization over these parameters would be preferable.)

The first three columns of values in Table 16 show the effect of the plate scale on the Zernike coefficients obtained from image PC-6 F889N.P2 over a wide range of plate scales. As the plate scale increases (more arcseconds per pixel in the image), the size of the features in the modeled entrance pupil (pixels per meter) increases, and the retrieved values of all tend to increase in magnitude. The value of the quadratic phase factor $BQ$ that was used in this case was $-0.000068$, which was found to be optimum by trial and error for a plate scale of 43 arcsec/per pixel. Later ray-tracing results, described in Section 4, predicted a value of $BQ$ of $-0.00005$, for which the results in the fourth column of values were obtained for the larger plate scale. Note that a decrease in the absolute value of $BQ$ decreased $\sigma_{11}$.

Figure 8 shows an example of optimizing the plate scale (by the scale parameter $s$, which represents the number of pixels per meter of the OTA entrance pupil and is proportional to the plate scale). An entire minimization was performed, each with the same parameters except $s$, which was varied. The points plotted in Fig. 8 represent the final rms error at the end of each minimization. In this case the quadratic phase factor for multiple-plane propagation was $BQ = -0.000054$, which is close to the value predicted by the updated ray-trace model. The error metric has two minima, one at $s = 61.74$, where it has a value of 0.14278 and for which $\sigma_{11} = -0.299 \mu m$ rms, and a second at $s = 62.64$, where it has a value of 0.14276 and for which $\sigma_{11} = -0.301 \mu m$ rms. We do not understand why there would be a double-bottom minimum. When performing many optimizations with different values of $BQ$, with each of these two values of $s$, we found that the optimum value of $BQ$ is $-0.000059$ for $s = 61.74$ and is $-0.000075$ for $s = 62.64$. The relationship between the plate scale and $s$ is given by $s = \rho_L(\lambda/n)$, where $\rho_L$ is the plate scale in radians (the number of radians separating the pixels). For $\rho_L = 0.0442$ arcsec $= 0.214 \mu rad$, $N = 256$ pixels, and $\lambda = 0.889 \mu m$: $s = 61.7$ pixels/m. This value, calculated from the now-known plate scale for the center of the CCD, is in quite good agreement with the first of the two minima that we found by minimizing the fitting error as a function of $s$. Furthermore we also found later from the optical system design that an appropriate value of $BQ$ is $-0.000051$ (see Section 4). Therefore both the values of the plate scale and $BQ$ predicted by ray tracing point to the first

<table>
<thead>
<tr>
<th>Plate scale</th>
<th>$s$ (pixels/m)</th>
<th>With $BQ = -0.000068$</th>
<th>With $BQ = -0.000054$</th>
<th>With $BQ = -0.000054$</th>
</tr>
</thead>
<tbody>
<tr>
<td>(arcsec/</td>
<td>41.54</td>
<td>43.00</td>
<td>44.40</td>
<td>44.40</td>
</tr>
<tr>
<td>pixel)</td>
<td>44.40</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$j = 4$</td>
<td>-2.213</td>
<td>-2.227</td>
<td>-2.311</td>
<td>-2.320</td>
</tr>
<tr>
<td>$j = 5$</td>
<td>-0.028</td>
<td>-0.003</td>
<td>-0.001</td>
<td>-0.000</td>
</tr>
<tr>
<td>$j = 6$</td>
<td>-0.025</td>
<td>0.025</td>
<td>0.027</td>
<td>0.026</td>
</tr>
<tr>
<td>$j = 7$</td>
<td>0.004</td>
<td>0.001</td>
<td>0.001</td>
<td>-0.001</td>
</tr>
<tr>
<td>$j = 8$</td>
<td>0.021</td>
<td>0.010</td>
<td>0.014</td>
<td>0.014</td>
</tr>
<tr>
<td>$j = 9$</td>
<td>-0.017</td>
<td>-0.020</td>
<td>-0.023</td>
<td>-0.022</td>
</tr>
<tr>
<td>$j = 10$</td>
<td>0.005</td>
<td>0.008</td>
<td>0.002</td>
<td>0.003</td>
</tr>
<tr>
<td>$j = 11$</td>
<td>-0.291</td>
<td>-0.292</td>
<td>-0.302</td>
<td>-0.299</td>
</tr>
<tr>
<td>$j = 22$</td>
<td>-0.001</td>
<td>0.006</td>
<td>0.007</td>
<td>0.008</td>
</tr>
<tr>
<td>Conic $k$</td>
<td>-1.0151</td>
<td>-1.0151</td>
<td>-1.0155</td>
<td>-1.0154</td>
</tr>
<tr>
<td>Root-mean-</td>
<td>0.1545</td>
<td>0.1442</td>
<td>0.1413</td>
<td>0.1415</td>
</tr>
<tr>
<td>squared</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>err $\sigma_{11}$</td>
<td>0.1428</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

$^a$Using multiple-plane propagation, fit coefficients $1$–$11.22$, $N = 256$, PSF weighted by diameter-220 circle, PC obscuration shifted by $-4.25$, $-3.25$ pixels.
minimum as being the true value. The results in the last column of Table 16 represent our estimate when we used parameters that are close to the best system parameters that we currently know. In this case $a_{11} = -0.299 \mu m$ rms, which is equivalent to a conic constant of $-1.0154$.

C. Pupil Reconstruction

Figure 9 shows an example of pupil reconstruction by one iteration of the iterative transform algorithm (ITA) for HARP1B image 33434. In this case the modeled phase error was placed across only the annular aperture of the OTA without any spiders, pads, or PC obscurations. The resulting image shows a clear indication of the location of all the obscurations derived from the measured data: the pads and spiders of the OTA and the spiders and central obscuration of the PC. Based on this picture, we then constructed the modeled obscurations (illustrated together in a single plane) shown in Fig. 10.

We also determined the shift of the PC obscurations by trying different shifts and picking the one that minimized the error metric. When this was done for the PC-6F689N_P2 image, the optimum shift was found to be $-4.5$ rows and $-3.25$ columns in pupil space when the array size was $N = 256$. (Fractional-pixel shifts of an array can be computed by Fourier transforming the array, multiplying by an appropriate linear-phase complex exponential, and inverse Fourier transforming). One pixel in this case corresponds to $0.0167 \text{ m}$ projected to the primary mirror of the OTA. Therefore the corresponding shift in that plane would be $-0.075 \text{ m}$ along $-v_2$ and $-0.054 \text{ m}$ along $+V_2$. This was thought to be a large shift when we consider how close to the center of the chip the image was located (at CCD pixel 531, 425). This is an indication that the optics of PC-6 are not aligned properly with the OTA, causing a pupil shift. This was explored in more depth as described below.

The center of symmetry, i.e., the CCD pixel for which the OTA and WF/PC central obscurations are aligned with one another, was designed to be at pixel 400, 400. The method we used to determine the actual center of symmetry (and thereby infer the alignment of the WF/PC relative to the OTA) is as follows. For each of several images an estimate of the aberrations is put over the 0.33-obscured doughnut-shaped aperture of the OTA. Then one iteration of the ITA is performed, the output of which is darker where there are obscurations in the pupil function. For the case of good-quality narrow-band data, the pads, spiders, and WF/PC central obscuration can be seen clearly, as shown in Fig. 9. We measured the positions of the reconstructed PC spiders to estimate the shift of the PC obscurations relative to the OTA obscurations. This is done for the images taken at different locations in the field of view. Then we perform a least-squares fit of a linear model of the WF/PC obscuration shift to the set of estimated WF/PC obscurations. The least-squares fit is performed for the collection of several images. This yields one equation for the row of the WF/PC shift as a function of $(x, y)$ and another for the column of the WF/PC shift as a function of $(x, y)$. Lastly, these two simultaneous equations are solved for the value of $(x, y)$ for which the row and column WF/PC shift are both zero. This defines the center of symmetry for any given WF/PC channel.

We performed the procedure described above for a collection of seven images from the HARP1B series on PC-6 at a 631-nm wavelength at despace $= -90 \mu m$. The result was that the center of symmetry was estimated to be $x = 254.5$, $y = 194.0$ pixels. When we recalculated the center of symmetry leaving
out a couple of the images, it changed by \( \sim 20 \) pixels.
The aberrations for each of the images were not optimized individually, and so with greater care the same procedure could yield an answer with greater accuracy and confidence. Further improvements in the estimates of the PC obscuration shifts could be obtained with the gradient search algorithm described in Ref. 13 by using the derivative with respect to the point-by-point magnitude. Another improvement would be to optimize the shift of a model of the PC obscurations during the phase-retrieval algorithm. Still another improvement would be first to deconvolve the jitter from the images before estimating the shift. Until these improvements are made, the center of symmetry results reported by Roddier and Roddier, who used our approach but performed it more carefully, should be taken to be more accurate than the center of symmetry reported above.

**D. Effects of Jitter**

We believe that jitter in the data from the HST is one of the largest components of the lack of agreement between our modeled PSF's and measured PSF's. Lacking jitter-free, long-wavelength data we could account for the jitter in the phase-retrieval algorithms, as described below.

The effect of jitter can be seen in Fig. 7, which compares a measured PSF (upper left) with a PSF computed from our polynomial-phase model of the aberrated wave front (lower right). The modeled PSF has a fine-fringe structure that is smeared out in the measured PSF. The modeled PSF does not include the effects of the finite spectral bandwidth, pixelation, or jitter. The spectral bandwidth of the F889 filter is 0.57% (5.1 nm/889 nm according to the WF/PC Instrument Handbook) and would not yield this degree of blurring. Furthermore the finite spectral bandwidth causes a blurring that is proportional to the distance from the center of the PSF, which would be negligible near the center of the PSF; consequently the blurring of this PSF that can be seen at its center cannot be a result of the spectral bandwidth. We simulated the effect of pixelization (integration over the area of the CCD pixel) in the modeled PSF and found that it yielded a degree of blurring that is much milder than that seen in the measured PSF; therefore pixelization also cannot explain the loss of the fine-detailed structure in the measured PSF. The jitter, which is known to be of the order of 0.1 arcsec in effective width, or \( \sim 2 \) pixels wide, does explain the differences. We model the effect of jitter to be the convolution of the PSF with a jitter spatial density function. The jitter causes a much poorer fit of the model to the measured data than would be the case if there were no jitter (14% rms error being the best fit to date). It can be seen from studies by Lyon et al. that jitter could cause errors in \( a_{11} \) in the range of 0.01 \( \mu \)m rms (which is the entire error budget). At the very least the jitter causes the level of uncertainty in our results to be much higher than if there were no jitter.

Jitter can be accounted for by two routes. The first is to correct the jitter in the measured data and then operate on that jitter-corrected data with the same phase-retrieval algorithms as before. The second route is to incorporate a model for the jitter directly within the phase-retrieval algorithm.

Any number of deconvolution algorithms can be used to remove the jitter from the measured PSF if the jitter density function is known. The fine-guidance sensors provide a measure of the jitter from which the jitter density function can be computed, but the reliability of that data had not been established. Alternatively we can use a blind deconvolution algorithm to estimate both the PSF and the jitter density function. Figure 7 shows the original measured PSF (upper left), PC-6 F889N.P2, and the PSF deconvolved both by the Ayers/Dainty algorithm\(^{21,22}\) (upper right) and by the Wiener filtering using the measured jitter data (lower left). The PSF's deconvolved by using the Ayers/Dainty algorithm had significantly greater contrast than those deconvolved by using the measured jitter data. That is, the jitter predicted by the Ayers/Dainty algorithm was significantly greater than that given by the measured jitter data. This is possible since the measured jitter data are undersampled in time relative to some of the highest temporal frequencies of the jitter. The two approaches were also combined: the jitter data were used as a starting estimate for the jitter in the Ayers/Dainty algorithm, and then we performed the Lucy algorithm\(^{23}\) by using the jitter function computed by the Ayers/Dainty algorithm. We obtained several different deconvolutions for a single PSF by using these various techniques, all of which are in good agreement with the measured PSF. This may indicate a lack of uniqueness in deconvolving an unknown or partially unknown jitter function from a PSF for the amount of noise and undersampling present in the data. At this point it is difficult to judge which is the most believable deconvolution result. The additional details in the deconvolved PSF's look quite believable. Some of the fine fringes seen in the simulated PSF (shown in the lower right of Fig. 7), which were washed out in the measured PSF (upper left), became visible in the jitter-deconvolved PSF's (upper right and lower left). The artifacts from dust (presumably on the field flattener in front of the CCD) became much sharper and more visible, as can be seen in Fig. 7.

Phase retrieval with Zernike coefficients was performed on various jitter-deconvolved versions of PC-6F889N.P2 (from Wiener, Lucy, and Ayers/Dainty deconvolutions). For all of them the \( a_{11} \) Zernike coefficient obtained for the jitter-deconvolved PSF did not change significantly from that of the original PSF. More extensive simulation studies by Lyon, on the other hand, showed significant changes in the retrieved value of \( a_{11} \) when jitter was present. The error in our fit of the modeled PSF to the measured PSF was actually worse for the jitter-deconvolved PSF's than it was for the original PSF. Possibly this
is because the deconvolution enhances the noise as well as the fine structure of the measured PSF. This issue deserves further study.

The second route to accounting for jitter is to include a model for the jitter directly within the phase-retrieval algorithm. To that end we derived analytic expressions for gradients of the phase-retrieval error metric with jitter included in the model of the data and for the partial derivative of the error metric with respect to a shift in the WF/PC obscurations. These are given in Appendix B. These analytic gradients were implemented in software within the phase-retrieval algorithm; however, the effort was concluded before this software could be debugged and exercised.

The circular artifacts seen in the PSF's in Fig. 17, which are more pronounced in the deconvolved PSF's, were reportedly a result of dust on a field flattener in front of the CCD. In Appendix C we verify the source of this artifact.

6. Inclusion of the $Z_{22}$ Contribution to Spherical Aberration

It has been argued that since $Z_{22}(r)$ has an $r^4$ term in it, this contribution to $r^4$ should be added to the spherical aberration $Z_{11}(r)$ for the purpose of determining what spherical aberration to correct. This should not be done. The $r^4$ term in $Z_{22}(r)$ is there merely to make $Z_{22}(r)$ orthogonal to $Z_{11}(r)$ and does not represent the real spherical aberration that one would want to correct. The $r^4$ term in $Z_{22}(r)$ is there only to balance the $r^6$ term, and it should be included only if the entire $Z_{22}(r)$ function were being corrected (which is not the plan for the replacement of WF/PC). In fact $Z_{11}(r)$ and $Z_{22}(r)$ are orthogonal over a 0.33-obliterated annular aperture, not over the actual aperture; nevertheless they are close to orthogonal over the actual aperture, and the coupling between them should be small.

In addition, consider the following. A typical value for the $Z_{22}(r)$ coefficient $a_{22}$ is 0.007 $\mu$m rms (see Section 5). When we convert the $r^4$ term in 0.007 $Z_{22}(r)$ to an equivalent coefficient, $a_{11}$ of $Z_{12}(r)$ yields $-0.052$ $\mu$m rms. If this value were added to an estimate of $a_{11}$ of $-0.299$ $\mu$m rms, the result would be a total estimate of $a_{11}$ of $-0.351$ $\mu$m rms. This large value of spherical aberration is inconsistent with all the other analyses performed on the HST, giving further evidence that the $r^4$ component of $Z_{22}(r)$ should not be added to our estimate of $Z_{11}(r)$. Again that term should be included only if the entire $Z_{22}(r)$ were being corrected.

7. Conclusions

We applied several new algorithms to retrieving the aberrations of the HST by using the blurred images of stars taken by the telescope on orbit. We also reconstructed pupil functions, which indicated an unexpected shift of the obscurations in PC-6 relative to those in the OTA, which indicates a misalignment of the optical axis of PC-6 relative to the OTA. Our best estimate of the spherical aberration, as specified by the coefficient of the 11th Zernike polynomial $a_{11}$ for the combined OTA and PC-6, is $-0.299$ $\mu$m rms of the wave-front error, which would correspond to a conic constant on the primary mirror of $-1.0154$. This amount of spherical aberration is larger than some of the earlier results obtained with the difference resulting from a more accurate system model, including multiple-plane propagation, and a larger, more accurate value of the plate scale. For us to determine the aberrations of the primary mirror of the OTA, an adjustment to this prescription is necessary to compensate for the known spherical aberration in PC-6, which the Jet Propulsion Laboratory reports to be equivalent to $-0.0010$ in the conic constant of the OTA. With this correction the conic constant of the primary mirror of the OTA is estimated to be $-1.0144$.

The error bars on our estimate are difficult to determine since they depend on systematic errors, such as poorly known parameters of the system, rather than random errors whose standard deviation can be derived. Cramer–Rao lower bounds show that the far out-of-focus PSF's are far more suitable for phase retrieval than the nominally in-focus PSF's. We performed retrieval experiments to show the expected errors caused by imprecise knowledge of the plate scale, a quadratic phase factor, and translation of obscurations in the PC. The jitter also limits the accuracy of our estimate by an unknown amount. We require further analyses to determine the actual error bars. The accuracy of our results to date was not sufficient to give us high confidence in the predictions of the Zernike coefficients (astigmatism, coma, etc.) other than the spherical aberration, of the point-by-point phase maps or of the extent to which there are aberrations in the secondary mirror of the OTA. We believe that the degree of confidence can be increased, and the error bars can be decreased, by several additional refinements, including (1) a more complete accounting for the effects of jitter, (2) an improved estimation of pupil functions, and (3) an automatic optimization over poorly known system parameters.

Portions of this paper were presented in Refs. 24 and 25.

Appendix A: Modified Zernike Polynomials

The polynomials used to describe the aberrations are the modified Zernike polynomials, orthonormal over an annular aperture with an inner radius of 0.330 times the outer radius. They are given in Table 17, which was adapted from Ref. 26 with corrections. Here the two spatial coordinates, $x$ and $y$, are given explicitly (whereas in the body of this paper $x$ was taken to be a two-dimensional coordinate). The radius $r = (x^2 + y^2)^{1/2}$ is normalized to unity at the outer edge of the aperture.
where parameter \( p \) in the input wave front:

\[ G^J(u') = G(u') \sum_u J(u - u') W(u) \left[ \frac{|F(u)|}{|G(u)|} \right] - 1. \]

(B7)

When Eqs. (B6) and (B7) are inserted into Eq. (B5), we have

\[ \frac{\partial E}{\partial p} = - \text{Re} \left[ \sum_{x_1} \frac{\partial U_1(x_1)}{\partial p} g^{J^*}(x_1) \right], \]

(B8)

and where \( c.c. \) denotes the complex conjugate of what precedes it. Let

\[ G^J(u') = P^*[G^J(u)] \]

(B9)

and \( P^* \) is the inverse propagation operator. For a decomposition of the phase into Zernike polynomials with coefficients \( a_j \), we have

\[ \frac{\partial E}{\partial a_j} = 2 \Im \left[ \sum_{x_1} U_1(x_1) Z_j(x_1) g^{J^*}(x_1) \right], \]

(B10)

and for a point-by-point phase map \( \theta(x_1) \) we have

\[ \frac{\partial E}{\partial \theta(x_1)} = 2 \Im[U_1(x_1)g^{J^*}(x_1)] \]

(B11)

and similarly for the other unknown parameters of \( U_1(x_1) \).

Next consider an unknown parameter \( q \) of the PC obscuration transmittance \( m_{3}(x_3) \). We assume here that there are no aberrations in plane \( x_9 \). Then we have, similar to the case in Ref. 13,

\[ \frac{\partial E}{\partial q} = - \sum_u G^{J^*}(u) \frac{\partial G(u)}{\partial q} + \text{c.c.} \]

\[ = - \sum_{x_3} \frac{m_{3}(x_3)}{q} U_3(x_3)[P_{4->3}*[G^J(u)]]^* + \text{c.c.}, \]

(B12)

where \( U_3(x_3) \) is the propagation of the wave front \( U_1(x_1) \) to the \( x_3 \) plane, and \( P_{4->3}^* \) is the inverse propagation from plane \( x_4 \) to \( x_3 \).

First, let the parameter \( q \) be the transmittance magnitude \( m_{3}(x_3) \) at point \( x_3 \). Then we have

\[ \frac{\partial E}{\partial m_{3}(x_3)} = - 2 \Re[U_3(x_3)[P_{4->3}*[G^J(u)]]^*]. \]

(B13)

Next let the parameter \( q \) be the unknown shift \( x_6 \) of the PC mask \( m_{3}(x_3 - x_6) \). Letting

\[ m_{3}(x_3) = \sum_u M_{5}(u') \exp(i2\pi u'x_3/N), \]

(B14)

we have

\[ \frac{\partial m_{3}(x_3 - x_6)}{\partial x_6} = \sum_u (-i2\pi u'/N) M_{5}(u') \times \exp[i2\pi u'(x_3 - x_6)/N], \]

(B15)
and, about \( x_0 = 0 \), we have

\[
\frac{\partial E}{\partial x_0} = -\sum_{x_3} \left( -i2\pi u'/N \right) M_3(u') \times \exp(i2\pi u'x_3/N)U_3(x_3)\left[ P_{4\rightarrow3}^* [G'(u)] \right] + \text{c.c.} \nonumber \\
= -2 \text{Im} \left( \sum_{x_3} F^{-1}[(2\pi u'/N)M_3(u')] \times U_3(x_3)\left[ P_{4\rightarrow3}^* [G'(u)] \right] \right) 
\]  

(B16)

Note that this requires three FFT’s for \( G(u) \) to be computed, four more FFT’s (or two convolutions) for \( G'(u) \) to be computed, and two more FFT’s (or one convolution) for each of the two components of the derivative of \( M_3(x_3) \) to be computed.

These analytic expressions for the derivatives of the error metric allow for an efficient gradient search algorithm that minimizes the error metric as a function of the unknown parameters, when we assume that the jitter function is known.

### Appendix C: Verification of Dust Artifacts

The dust artifacts sharpened during the jitter deconvolution, as can be seen in Fig. 17. If the dust is on the field flattener and moves along with the detector array, it should not be blurred by the jitter. Therefore it would seem that, by deblurring the entire image for the jitter, we would cause the dust artifacts to become more blurred. The fact that the dust artifacts sharpened during jitter deconvolution was contrary to our expectations and caused us to question the hypothesis that the dust was located on the first surface of the field lens. However, we confirmed the hypothesis that the dust is on the field flattener, as described below. A possible explanation of why the dust looks sharper after deblurring for jitter is that the deblurring operation is a high-pass filter that enhances all the edges in the image.

The axial location of the dust causing the artifacts was determined as follows: Assuming an essentially plane wave front at the plane of the dust (which is accurate for the far out-of-focus PSF’s) and assuming that the dust acts as a pointlike scatterer, the expected intensity pattern caused by the dust is

\[
I(x) = \left| \exp(i2\pi d/\lambda) + a \exp(i2\pi(r/\lambda + c)) \right|^2 
= 2 + 2a \cos[2\pi(r-d)/\lambda + 2\pi c], 
\]  

(C1)

where the dust is assumed to be at lateral location \( x = 0, d \) is the distance (which we wish to determine) from the plane of the dust to the plane of the CCD, \( a \) is the amplitude of the wave front scattered from the dust, \( c \) is an unknown phase constant associated with scattering from the dust, and \( r = (d^2 + x^2)^{1/2} \) is the distance from the dust to a given point \( x \) on the CCD. The radii of successive peaks and nulls of the resultant concentric-circular fringe patterns are given by

\[
x_m = \left[ 2d\lambda[(m + n_0)/2 - c] \right]^{1/2}, 
\]  

(C2)

where \( n_0 \), an integer, is a reference fringe number.

We obtained this expression by setting the argument of the cosine equal to \( 2\pi n_0 \) and using a Taylor series expansion of \( r \) in terms of \( d \) and \( x \). For two of the dust artifacts in image PC-6P89N.P2, we measured three successive radii at a null, a peak, and a null to be 28.5, 55.4, and 75.9 \( \mu \text{m} \). There could be a 20% error in these numbers, because the data are given at a spacing of 15.24 \( \mu \text{m} \). Fitting the model for \( x_m \) above to these three values, we determined the three unknowns to be \( n_0 = 0, c = 0.34 \) wave, and \( d = 2.7 \) mm. For comparison, for diffractive propagation purposes, the effective distance from the front surface of the field flattener, at its thinnest point, to the CCD is 1.63 mm + 1.27 mm/1.378 = 2.5 mm, which is in good agreement with the value of \( d \) determined from the dust artifacts, considering the uncertainties in estimating the locations of the peaks and nulls of the fringe.

The authors gratefully acknowledge the assistance of A. M. Kowalczyk who wrote much of the software that was necessary for us to obtain the results in this paper and C. A. Bobrowski who exercised some of the software. This work was funded primarily by Jet Propulsion Laboratory under contract JPL 968892 and partially by the Environmental Research Institute of Michigan with internal research and development funds.

### References