Nonlinear optimization algorithm for retrieving the full complex pupil function

Gregory R. Brady and James R. Fienup

The Institute of Optics, University of Rochester, Rochester, New York 14627-0186
gbrady@optics.rochester.edu

http://www.optics.rochester.edu/workgroups/fienup

Abstract: Techniques for retrieving the phase of an optical field typically depend on assumptions about the amplitude of the field in a desired plane, usually a pupil plane. We describe an approach that makes no such assumptions and is capable of retrieving both the amplitude and phase in the desired plane. Intensity measurements in two or more planes are used by a nonlinear optimization algorithm to retrieve the phase in the measurement planes. The complex field (amplitude and phase) in the desired plane is then computed by simple propagation. We show simulation results and examine the convergence of the algorithm.

References and Links

1. J. R. Fienup, "Phase Retrieval Algorithms: a Comparison," Appl. Opt. **21**, 2758-2769 (1982).
2. J.N. Cederquist, J.R. Fienup, C.C. Wackerman, S.R. Robinson and D. Kryskowski, "Wave-Front Phase Estimation from Fourier Intensity Measurements," J. Opt. Soc. Am. A **6**, 1020-1026 (1989).
3. J. R. Fienup, "Phase-retrieval Algorithms for a Complicated Optical System," Appl. Opt. 32, 1737-1746 (1993).
4. J. R. Fienup, J. C. Marron, T. J. Schulz, and J. H. Seldin, "Hubble Space Telescope Characterized by Using Phase-retrieval Algorithms," Appl. Opt. **32**, 1747-1767 (1993).
5. D. L. Misell, "A Method for the Solution of the Phase Problem in Electron Microscopy," J. Phys. D: Appl. Phys. **6**, L6-L9 (1973).
6. C. Roddier and F. Roddier, "Combined Approach to Hubble Space Telescope Wave-Front Distortion Analysis," Appl. Opt. **32**, 2992-3008 (1993).
7. S. M. Jefferies, M. Lloyd-Hart, E. K. Hege, and J. Georges, "Sensing wave-front amplitude and phase with phase diversity," Appl. Opt. **41**, 2095-2102 (2002).
8. J. H. Seldin and R. G. Paxman, "Joint Estimation of Amplitude and Phase from Phase-Diversity Data," in "Signal Recovery and Synthesis" Topical Meeting of the Optical Society of America (June, 2005), paper JTuB4.
9. Joseph W. Goodman, *Introduction to Fourier Optics 2nd Ed.* (McGraw-Hill, 1996), Chap. 3.
10. W. H. Press, B. P. Flannery, S. A. Teukolsky, and W. T. Vetterling, *Numerical Recipes: The Art of Scientific Computing* (Cambridge University Press, 1986), Chap. 10.
11. D. Malacara and S. L. DeVore, "Interferogram Evaluation and Wavefront Fitting" in *Optical Shop Testing*, D. Malacara, Ed. (Wiley, 1992), Chap. 13.
12. G. R. Brady and J. R. Fienup, "Retrieval of Complex Field Using Nonlinear Optimization", in "Signal Recovery and Synthesis" Topical Meeting of the Optical Society of America (June, 2005), postdeadline paper JTuC3.

1. Introduction

Phase retrieval has been used in the past for determining the phase in a pupil plane from intensity measurements in one or more planes near focus and knowledge of the shape of the pupil [1-3]. However, the amplitude distribution in the pupil may not be perfectly known, and

extra means may be necessary to estimate it [4]. A second approach is to use an iterative transform (Misell) type algorithm employing two or more intensity measurements near the focus to retrieve the phase in those planes, without knowledge of the amplitude distribution in the pupil [5, 6]. Then, by propagation from one of these planes, one can estimate the complex field (amplitude and phase) in the pupil [6], or any other plane.

As an alternative to these approaches, we have developed a phase-diverse phase retrieval algorithm based on a nonlinear optimization method for estimating the phase in the measurement planes. In this approach, the pupil is illuminated by a coherent, monochromatic point object, as is the usual situation in phase retrieval. The retrieved phases are combined with the amplitude information from the intensity measurements to form a full complex field in the measurement planes, which can then be inverse propagated to the pupil plane. Thus, we arrive at an estimate of the amplitude and phase in the pupil plane with no knowledge of the amplitude distribution there. The nonlinear optimization approach is useful because of its explicit minimization of an objective function. This method permits the modeling and compensation for many non-ideal effects, such as finite detector pixel sizes.

An example use for such an approach is the precise measurement of the form of optical surfaces or the transmitted wavefront of optical systems during manufacturing. This technique is an attractive alternative to the interferometric tests currently in use since it requires a minimum of optics, is vibration-insensitive, and is not subject to retrace errors. In this application the shape and size of the pupil of the system under test may not be known with sufficient accuracy. The clear aperture diameter of optical parts during their manufacture is generally not well controlled. If we are testing an optical system that does have a well-known aperture stop, it is possible that the exit pupil may be aberrated. If we are to test the system at an off-axis field angle, the pupil may be vignetted, which could cause a non-uniform amplitude distribution in the pupil. In these situations, it would be desirable to employ an algorithm that does not rely on precise knowledge of the pupil.

It is also possible to jointly estimate the amplitude and phase in the pupil using phase diversity approaches [7, 8]. In that case one is imaging an incoherent, extended object and estimating that as well. In those works [7, 8], estimates were made of the pupil amplitude and phase directly, rather than first estimating the phase of a coherent field in the measurement planes as is done here. Phase diversity can also be used with a point object, as this is a special case of an extended object.

Section 2 of this paper describes the basis of the algorithm and implementation details, including the propagation models used. Section 3 describes numerical experiments we have performed to demonstrate this approach and examine its accuracy and convergence.   Section 4 concludes the paper.

2. Description of algorithm

2.1. Multiple plane nonlinear optimization

A nonlinear-optimization-based phase-retrieval algorithm typically retrieves the phase in a desired plane, usually a pupil plane, where the amplitude is assumed to be well known. It does this by numerically propagating an estimated field in the desired plane to the planes where intensity measurements have been made, and comparing the computed field amplitudes with the measured field amplitudes (the square root of the measured intensities) using an objective function. The algorithm attempts to minimize this objective function by varying the phase in the desired plane, preferably making use of an analytic expression for the gradient of the objective function with respect to the phase values.

In contrast, our method requires no *a priori* knowledge of the pupil. Instead, we initially employ one of the measured amplitudes as the known amplitude distribution,

$$\left| F_j \left(r,s \right) \right| = \sqrt{I_j \left(r,s \right)} \tag{1}$$

where $I_j (r, s)$ is the measured intensity in the jth measurement plane, $F_j (r, s)$ is the complex field in this plane, and r and s are pixel indices of the sampled field and intensity distribution.

We refer to the plane where we are using the intensity measurement in this way as the "master" plane. An estimate of the phase $\theta_j(r, s)$, which is initially a guess, is combined with the known amplitudes to form a numerical field,

$$G_j(r,s) = \left|F_j(r,s)\right| \exp\left[i\theta_j(r,s)\right].\qquad(2)$$

This field is numerically propagated to each of the other measurement ("slave") planes. This propagation can be performed with different techniques. We have chosen the angular spectrum method [9] because it is exact (within the scalar assumption) and the sample spacing is invariant with propagation distance. A discrete version of a field in the kth plane, propagated a distance Δz_{kj} from the jth plane, can be written as

$$G_{kj}(r,s) = IDFT\left\{DFT\left\{G_j(r,s)\right\} \exp\left(i2\pi\Delta z_{kj}\left\{\frac{1}{\lambda^2} - \left[\left(\frac{t}{Md_r}\right)^2 + \left(\frac{u}{Nd_s}\right)^2\right]\right\}^{\frac{1}{2}}\right)\right\}.\qquad(3)$$

The operations $DFT\{\}$ and $IDFT\{\}$ are the discrete Fourier transform and its inverse, respectively, both over arrays of size $M \times N$, which are computed using a fast Fourier transform (FFT). The quantities d_r and d_s represent the spacings between these samples of the field in plane j, and t and u are indices in the Fourier domain. The nomenclature $G_{kj}(r, s)$ means the field in the kth plane obtained by propagating the field in the jth plane. We can write this propagation as a linear operator,

$$G_{kj}(r,s) = \mathbf{P_{kj}}\left[G_j(r,s)\right].\qquad(4)$$

Once we have a propagated version of the field in each of the measurement planes, we compare their amplitudes with the measured amplitudes using an objective function. There are many possible objective functions, but a simple and useful one is a weighted sum-of-squared-errors metric,

$$E_j = \sum_{k\neq j}\sum_{r,s} W_k(r,s)\left[\left|F_k(r,s)\right| - \left|G_{kj}(r,s)\right|\right]^2\qquad(5)$$

where the jth plane is the master plane, and k indexes over the slave planes. $W_k(r, s)$ is a weighting function that is used, for example, to discard the effect of bad pixels in the detector array or emphasize the contribution of certain regions of the pattern. It may be different for each plane.

Any number of nonlinear optimization algorithms may be used to minimize E_j. It is important for the algorithm to take advantage of the gradient of the objective function with respect to each of the phase values, but not require second derivatives. This is because, for $M \times N$ phase values there will be $M^2 \times N^2$ second derivatives, which can easily form an array larger than can be straightforwardly stored in a 32-bit computer. The algorithm we use that meets these requirements is the conjugate gradient algorithm [10]. In our situation, the gradient can be calculated very efficiently using a relatively simple analytic expression,

$$\frac{\partial E_j}{\partial\theta_j(r,s)} = 2\,\mathrm{Im}\left[G_j(r,s)\sum_{k\neq j}G_{jk}^{w*}(r,s)\right],\qquad(6)$$

which follows directly from the generalization of Eq. (23) in Ref. 3. $G_{jk}{}^w(r, s)$ is given by

$$G_{jk}^w(r,s) = \mathbf{P_{kj}^\dagger}\left[G_{kj}^w(r,s)\right]\qquad(7)$$

where $\mathbf{P_{kj}}^\dagger[\ ]$ is the inverse propagation operation to the one shown in Eq. (4) and

$$G_{kj}^w(r,s) = W_k(r,s) \left[|F_k(r,s)| \frac{G_{kj}(r,s)}{|G_{kj}(r,s)|} - G_{kj}(r,s) \right]. \tag{8}$$

If we have K measurement planes, these expressions allow us to compute the derivatives of the objective function with respect to the phase value at MN points using just $2(K-1)$ propagation operations, which can be efficiently implemented using the FFT. This technique is vastly more efficient than using a finite difference technique to compute the partial derivatives individually. Optimization over the coefficients of a basis-set (such as Zernike polynomials) expansion of the phase may also be done [3]. That is not, however, best in this application because the measurement planes may be in a region of caustics where the expansion would require a large number of terms.

2.2. Algorithm iteration details

In the algorithm described above, it is unclear which of the measurement planes should be the master plane. We have addressed this by using all of the planes in sequence as the master plane, which helps to avoid algorithm stagnation and increases the resistance of the algorithm to the noise in any particular plane.

To illustrate, consider the three-plane case shown in Fig. 1. To get the algorithm started, one may make an initial guess for the phase in the master plane. However, we found that a superior approach is to use a phase in the master plane that was propagated from the pupil plane. Although the pupil is not assumed to be well known, we will usually be able to approximate it at least coarsely. To begin, we make a guess of the field in the pupil (which may be grossly in error), which is propagated to the first measurement plane (plane 1), which we will term as step A. Making our guess of the phase in the pupil plane automatically enforces phase relationships that exist in the measurement planes due to propagation so that the algorithm does not need to iterate needlessly to recreate them, during which there is a chance that it may stagnate. This propagation is performed in a two-step manner, as illustrated in Fig. 1 (a). First, our initial guess of the field in the pupil, $g_0(m,n)$, is propagated to a focal plane using standard Fresnel diffraction,

$$G_f(p,q) = \exp\left\{ \frac{i\pi}{\lambda f} \left[(p d_p)^2 + (q d_q)^2 \right] \right\} DFT\{g_0(m,n)\} \tag{9}$$

where f is the focal length of the system under consideration. The phase of $g_0(m,n)$ is the departure from a spherical wavefront. The sample spacings d_p and d_q in the focal plane are related to the sample spacings d_m and d_n in the pupil plane by the relations

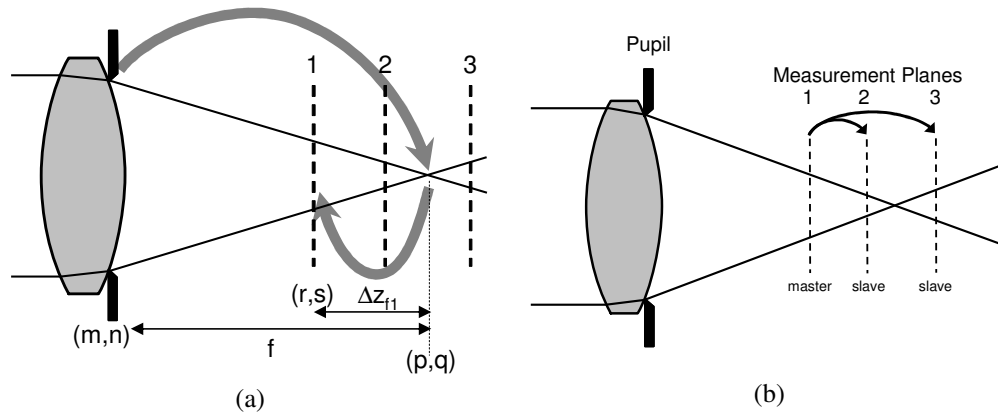$$d_p = \frac{\lambda f}{M d_m}, \quad d_q = \frac{\lambda f}{N d_n}. \tag{10}$$

Fig. 1. (a) Steps used in two-step propagation from pupil to measurement plane. (b) Example of propagations taking place when calculating the objective function, corresponding to step B of the algorithm.

The second step is to propagate the field in the focal plane to the master plane using Eq. (3). We will refer to this two step operation as $\mathbf{P_{0j}}$,

$$G_j(r,s) = \mathbf{P_{0j}}\big[g_0(m,n)\big]. \tag{11}$$

Because of the invariance of the sample spacing when using the angular spectrum method, $d_r = d_p$ and $d_s = d_q$.

In step B, plane 1 is the master plane. The initial field guess uses the phase of the propagated field from step A and the measured amplitudes in plane 1. The nonlinear optimization described in Sect. 2.1 is performed using plane 1 as the master and planes 2 and 3 as the slaves, but only for a small number of iterations. In step C, the resulting field is propagated to plane 2, where the resulting phase is used with the measured amplitudes in plane 2 to form the initial guess for another set of optimization iterations, step D, for which plane 2 is the master and planes 1 and 3 are the slaves. This process is repeated in steps E and F using plane 3 as the master and the planes 1 and 2 as the slaves. In step G the result of the optimization iterations with plane 3 as the master is then propagated back to plane 1. We then loop through the entire process, steps B through G, employing each of the planes as the master repeatedly, until we have reduced the value of the objective function to an acceptably small value. The resulting field is propagated back to the pupil plane in step H,

$$g(m,n) = \mathbf{P_{j0}^{\dagger}}\big[G_j(r,s)\big], \tag{12}$$

where we now have an estimate of the pupil amplitude and phase.

Any of the measurement planes can be used for the final propagation to the pupil plane. Which plane yields an estimate, $g(m,n)$, in the pupil that is closest to the true field, $f(m, n)$, is not yet known. Another approach is to average in the pupil plane the inverse propagated results (either the fields or the amplitudes and phases separately) from all of the measurement planes. We then must choose the appropriate method to average these complex numbers. We examine these results in Sect. 3.3 below.

3. Simulation Results

3.1. Example results

We demonstrate this new approach with computer simulations. We generated a complex pupil function using a uniformly illuminated circular pupil of diameter 10 mm paired with a smoothly varying phase function, shown in Fig. 2. The phase function contains mostly spherical aberration and some small asymmetric aberrations. The magnitude of the aberration was approximately 0.30 waves RMS. The pupil function was applied to an f/50 lens with a

focal length of 500 mm. This system is slow enough that the resulting intensity patterns could easily be Nyquist sampled by a detector with a relatively large pixel pitch. The assumed illumination wavelength was 632.8 nm.

Amplitude   Phase
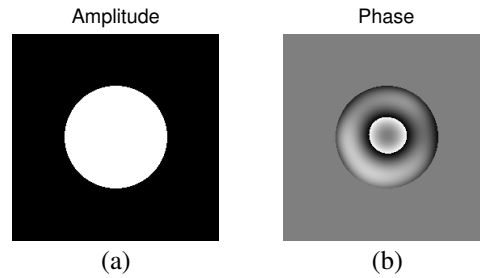

(a)   (b)

Fig. 2. (a) Amplitude and (b) phase functions used to generate intensity patterns and that will be retrieved by our algorithm. The phase is shown modulo 2π.
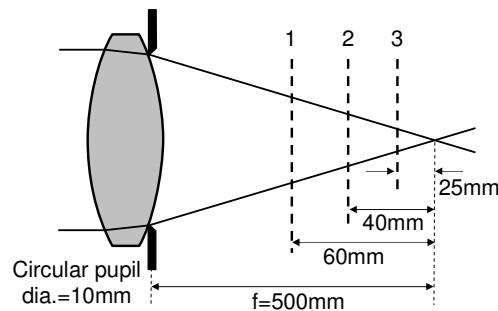

Fig. 3. Schematic of system modeled in simulations.

Three simulated intensity patterns were computed using Eq. (11). These three planes were located 60 mm, 40 mm and 25 mm inside the paraxial focal plane of the system (given a negative sign in our convention), as illustrated in Fig. 3. The resulting patterns are shown in Fig. 4. We added realistic detector noise to the simulated intensity patterns. Poisson noise was included such that the brightest pixel in each image contained 32,000 photoelectrons. Gaussian read noise of 16 photoelectrons was also included, and the images were quantized to 12 bits. These values are consistent with a moderate-quality CCD camera.

–60 mm   –40 mm   –25 mm
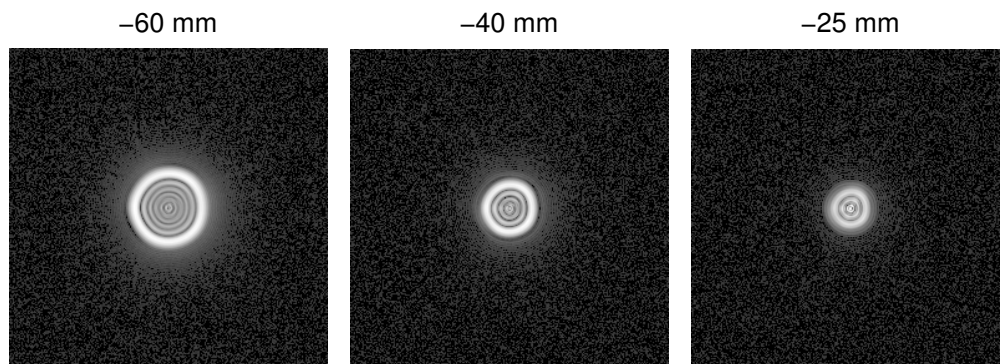

Fig. 4. Simulated measured intensity patterns in the three measurement planes. The dynamic range was compressed to bring out the noise in the background.

We performed four reconstructions with this simulated data using all three possible combinations of two planes and all three planes together. Fig. 5 shows some representative reconstructed intensity patterns in the case where the optimization was run using all three planes. The algorithm converged to a propagating field in good agreement with all three intensity measurements. It is interesting to note that, while the core of the reconstructed intensity patterns agrees nearly perfectly (visually) with that in the measurement plane, the noise in the outskirts is markedly different. The noise present in these patterns is noise propagated from master plane.

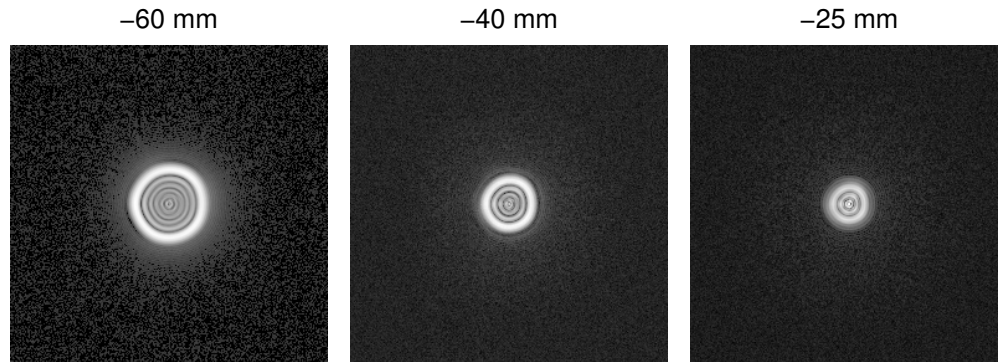| −60 mm	−40 mm	−25 mm

Fig. 5. Master plane (-60 mm plane) and reconstructed intensities (-40 mm and -25 mm planes) formed by propagating field with magnitudes given by master plane and our retrieved phases.

The initial guess and reconstructed phases in the pupil plane for the case where only two measurement planes, the -60 mm and -25 mm planes, were used are shown in Fig. 6. A similar case where all three planes were used is shown in Fig. 7. The initial guess for the pupil amplitude was a circular aperture of diameter 8 mm for Fig. 6 and 13 mm for Fig. 7, markedly different from the actual. This demonstrates that an accurate pupil estimate is not necessary for this case. The initial guess for the phase array was zero-mean Gaussian-distributed random phases with a standard deviation of 0.1 waves. The examples shown here were results after 25 cycles (a cycle being one loop of steps B to G in Sect. 2.2) of the algorithm were performed, with 10 optimization iterations performed with each plane as the master for each cycle. In the results shown, the field in the pupil was computed by inverse propagating the complex field from a single plane, the plane furthest from focus, to the pupil plane.

The agreement between the actual and retrieved pupil phase is excellent: 0.029 waves RMS in the two-plane case shown and 0.017 waves RMS in the three-plane case shown. The reconstructed pupil amplitude distributions do not agree as well, being more sensitive than the retrieved phase to measurement noise. Increasing the SNR of the measurement or the number of measurement planes improves these pupil amplitude reconstructions. However, the pupil amplitude reconstructions are clear enough to determine the shape of the aperture to within a pixel.
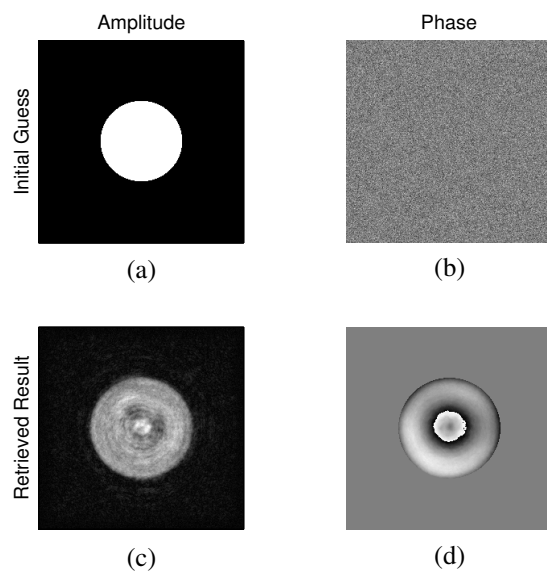
Fig. 6. Initial guess for (a) amplitude and (b) phase in pupil and resulting reconstructed (c) amplitude and (d) phase for the case where two intensity measurements in the -60 mm and -25 mm planes were used. The error in the phase from the actual, shown in Fig. 2 (b), is 0.029 waves RMS.
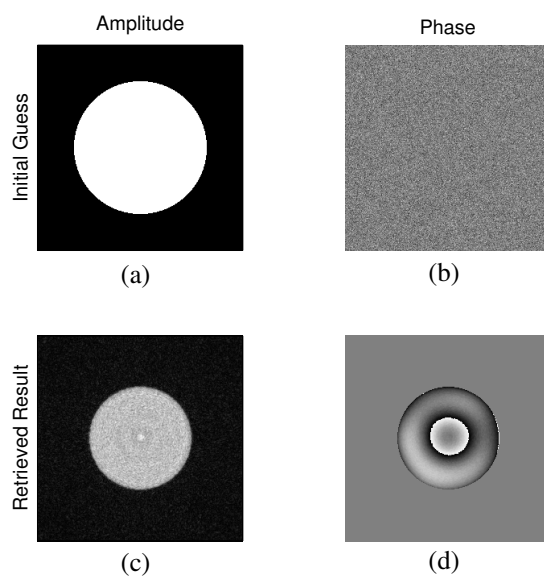

Fig. 7. Initial guess for (a) amplitude and (b) phase in pupil and resulting reconstructed (c) amplitude and (d) phase for the case where all three intensity measurements were used. The error in the phase from the actual, shown in Fig. 2 (b), is 0.017 waves RMS.

   In Fig. 8 we show a result obtained with a more complicated pupil amplitude: a collection of three circular subapertures. The initial pupil guess was a large circle, quite different from the actual pupil, but the algorithm was able to retrieve an amplitude and phase in good agreement.
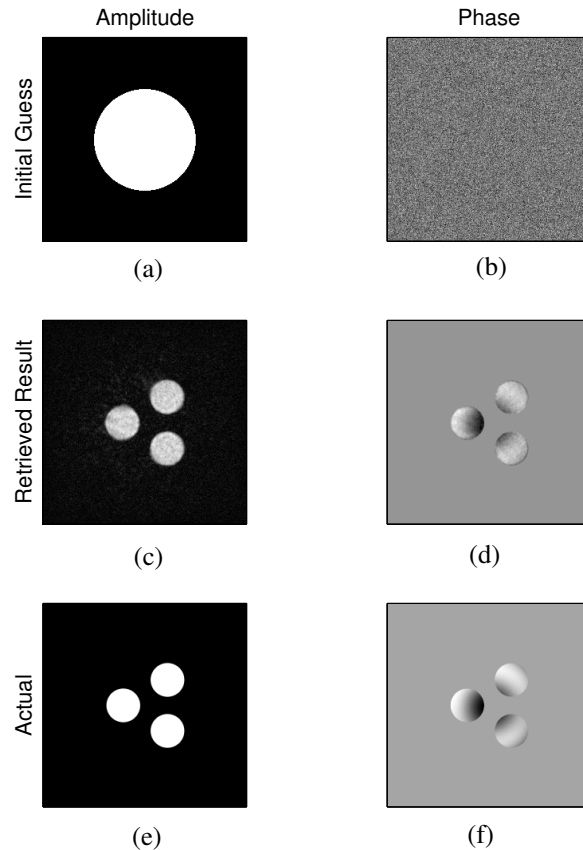
Fig. 8. Initial guess for (a) amplitude and (b) phase in pupil, the resulting reconstructed (c) amplitude and (d) phase, and actual (e) amplitude and (f) phase for a more complicated aperture. The error in the phase from the actual is 0.020 waves RMS.

3.2. Algorithm convergence

The convergence of the algorithm is shown in Fig. 9, a plot of the error in the retrieved wavefront versus number of iterations. The case where we have used three measurement planes converges most quickly and to the lowest error value, 0.019 waves RMS, after only 60 optimization iterations (2 cycles, with 10 iterations per plane per cycle). All three combinations of two planes were tried. The best two-plane case had about twice the error of the three-plane case, and the second-best two-plane case had twice that error, or about 0.06 waves RMS. The third two-plane case did not converge and stagnated at a result worse than the initial guess. In this case the objective function value was small and the reconstructed intensity pattern in the measurement plane compared well with the measurement. For comparison, the final value of the objective function for the -60 mm, -40 mm planes that did not converge was 0.008, while all of the other three cases resulted in an objective function of about 0.014. Note that in the three-plane case, there is an additional plane so the objective function for a single plane is less than in the two plane case. These results indicate that, to within the measurement noise, there is more than one solution consistent with the noisy measurements when only these two planes are employed. This illustrates the desirability of using additional measurement planes and the importance of carefully selecting the locations of the planes, from the standpoint of convergence, uniqueness and absolute accuracy of the results. A theoretical or more involved numerical study of the convergence of this algorithm is a topic for further research, since convergence is not guaranteed.
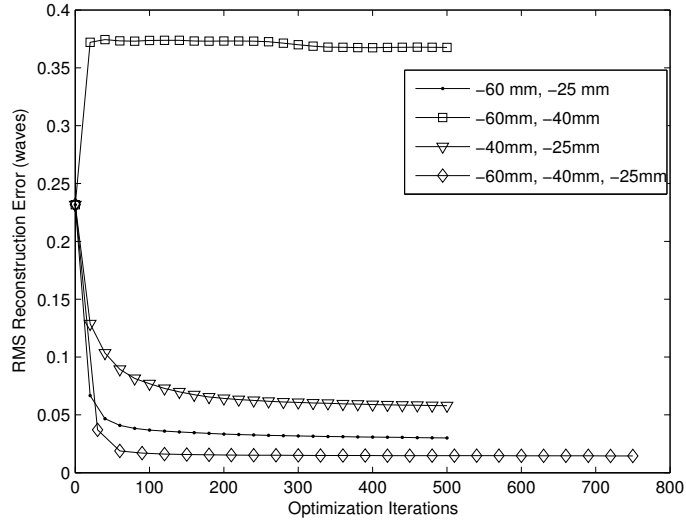
Fig. 9. Reconstruction error versus number of iterations for cases where two planes and threes planes have been used.

3.3. Effects of pupil plane field averaging

As was mentioned in Sect. 2.2, it is unclear which plane is optimal for performing the inverse propagation back to the pupil plane, or whether an average should be used. To learn what works best, we performed retrievals with eight different phase distributions, each with the same circular amplitude distribution. For this experiment, the measurements were those at the -60 mm, -40 mm and -25 mm described above. The results, shown in Table 1, seem to indicate that when inverse propagating using the field from a single plane, the best plane to use, on average, would be the -60 mm plane. This is most likely true because the dynamic range is lower (and hence the average signal-to-noise ratio is higher). Averaging the fields (the real and imaginary parts) gave, on average, a worse RMS error value than the best single plane, on average. It may nevertheless be preferable to average the fields, even though it may not always produce an error as low as from the best single plane, because it is usually comparable to the best plane value and does not require us to determine which plane is optimal. Averaging the phase and amplitude separately requires unwrapping of the phase and usually produced a slightly worse result than averaging the field, making it a less attractive method.

Table 1. RMS error of pupil plane phase formed by inverse propagating from each of the measurement planes and by averaging the pupil plane field with two different methods.

	RMS Reconstruction Error (waves)				
Trial Number	-60 mm plane	-40 mm plane	-25 mm plane	Average fields	Average amplitude and phase
1	0.01198	0.01208	0.01454	0.01192	0.01217
2	0.05254	0.05234	0.05220	0.05229	0.05224
3	0.01698	0.02192	0.02691	0.01933	0.02037
4	0.01536	0.01482	0.02687	0.01540	0.01737
5	0.01156	0.01118	0.01251	0.01123	0.01131
6	0.02550	0.05058	0.03679	0.03007	0.03969
7	0.01268	0.01332	0.01483	0.01284	0.01307
8	0.01203	0.01483	0.02669	0.01351	0.01611
Mean	0.01983	0.02388	0.02642	0.02083	0.02279

3.4. Using results as input to phase-only retrieval algorithm

Fig. 10 shows an additional method for improving the phase estimate that we have explored. The results from the algorithm described above are used as the input to a phase-only retrieval algorithm that directly estimates the Zernike coefficients of the wavefront, as described in Ref. 3, using an assumed amplitude for the pupil. The amplitude constraint fed into the phase-only retrieval algorithm is produced by thresholding the pupil amplitude estimate from our algorithm. The initial guess of the Zernike coefficients is formed by fitting Zernikes to the point-by-point phase estimate from our algorithm using the method described in Ref. 11. This procedure incorporates two assumptions that our more general algorithm does not make: that the amplitude distribution in the pupil is nominally uniform and that the phase in the pupil is smooth and can be well described by a modest number of Zernike polynomials.
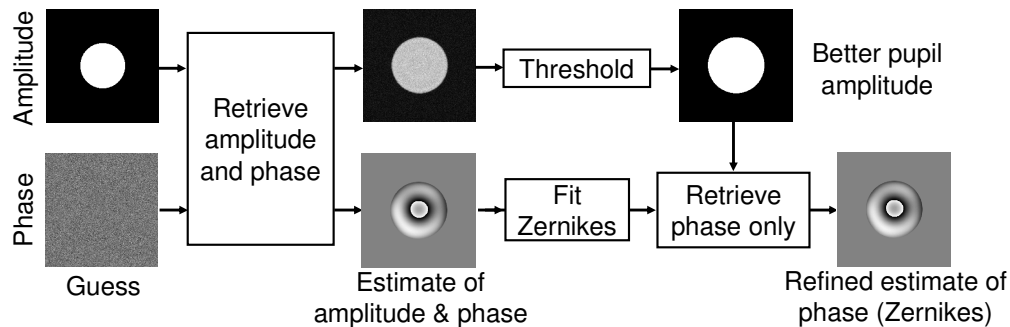

Fig. 10. Schematic diagram of a method to improve retrieved phase results.

The result of applying this method to the amplitude and phase retrieved from two intensity measurements is shown in Fig. 11. The amplitude and phase that we wish to retrieve are those shown in Fig. 2. The two planes employed are the -60 mm plane and the -25 mm planes, the same as used for the result shown in Fig. 6; however we ran only 10 cycles of the algorithm as opposed to the 25 above. This causes our retrieved result to be somewhat poorer than above and allows our refinement to do more improvement.

We see that the simple act of fitting the wavefront to the first 36 Zernikes polynomials reduced the error from 0.053 waves RMS to 0.018 waves RMS, and that the phase-only retrieval algorithm reduced it further to 0.011 waves RMS. We also found that if the

wavefront retrieved by our algorithm is already quite smooth, such as it is for the three-plane example shown in Sect. 3.1, there is no benefit from performing this additional refinement.
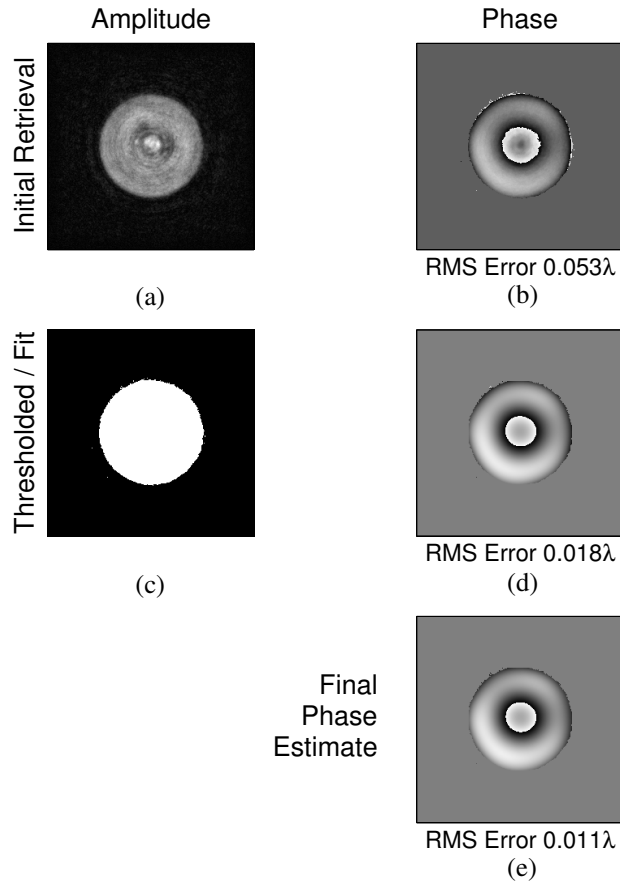

Fig. 11. (a) Amplitude and (b) phase result from our algorithm. (c) Thresholded amplitude and (d) Zernike fit phase. (e) Final phase estimate after running through a phase-only retrieval algorithm.

4. Conclusion

We have shown that it is possible to get good reconstructions of the full field (amplitude and phase) in the pupil plane of an optical system using a nonlinear-optimization algorithm and intensity measurements from two or more planes. The algorithm converges to a result that allows for an excellent estimate of the phase and a good estimate of the amplitude in the pupil after a modest number of optimization iterations. Simulation results making use of two and three measurement planes show that reconstruction errors on the order of 0.01 waves RMS are achievable with realistic SNR. We also showed that the algorithm converges rapidly to a low error value, but that this is sensitive to the location of and number of measurement planes. We examined methods for determining the final field in the pupil plane and determined that averaging the inverse-propagated fields from each of the measurement planes reliably results in accurate retrieval phases. We also examined using a phase-only retrieval algorithm to refine the estimate from our algorithm. Further work is needed to more fully understand the algorithm's uniqueness and convergence properties, and the relative performance of using a master plane as one of the measurement planes, as opposed to estimating both amplitude and phase in a master plane such as the pupil plane.

Portions of this paper were presented in [12].

Acknowledgements