Solving large-scale general phase retrieval problems via a sequence of convex relaxations

Doelman, Reinier; Nguyen, Thao; Verhaegen, Michel

DOI
10.1364/JOSAA.35.001410

Publication date
2018

Document Version
Final published version

Published in
Journal of the Optical Society of America A: Optics and Image Science, and Vision

Journal of the Optical Society of America A
OPTICS, IMAGE SCIENCE, AND VISION

Solving large-scale general phase retrieval problems via a sequence of convex relaxations

Reinier Doelman,* Nguyen H. Thao, and Michel Verhaegen

Delft Center for Systems and Control, Delft University of Technology, Delft, The Netherlands
**Corresponding author: r.doelman@tudelft.nl*

Received 23 February 2018; revised 27 June 2018; accepted 27 June 2018; posted 27 June 2018 (Doc. ID 324747); published 24 July 2018

We present a convex relaxation-based algorithm for large-scale general phase retrieval problems. General phase retrieval problems include, e.g., the estimation of the phase of the optical field in the pupil plane based on intensity measurements of a point source recorded in the image (focal) plane. The non-convex problem of finding the complex field that generates the correct intensity is reformulated into a rank constraint problem. The nuclear norm is used to obtain the convex relaxation of the phase retrieval problem. A new iterative method referred to as convex optimization-based phase retrieval (COPR) is presented, with each iteration consisting of solving a convex problem. In the noise-free case and for a class of phase retrieval problems, the solutions of the minimization problems converge linearly or faster towards a correct solution. Since the solutions to nuclear norm minimization problems can be computed using semidefinite programming, and this tends to be an expensive optimization in terms of scalability, we provide a fast algorithm called alternating direction method of multipliers (ADMM) that exploits the problem structure. The performance of the COPR algorithm is demonstrated in a realistic numerical simulation study, demonstrating its improvements in reliability and speed with respect to state-of-the-art methods.

OCIS codes: (100.5070) Phase retrieval; (010.7350) Wave-front sensing; (100.3190) Inverse problems.

https://doi.org/10.1364/JOSAA.35.001410

1. INTRODUCTION

Recovery of a signal from several measured intensity patterns, also known as the *phase retrieval problem*, is of great interest in optics and imaging. Recently, it was shown in [1] that the problem of estimating the wavefront aberration from measurements of the point spread functions can be formulated as a phase retrieval problem.

In this paper, we consider the general phase retrieval problem [2],

$$\text{find} \quad \mathbf{a} \in \mathbb{C}^{n_a} \text{ such that } \mathbf{y}_i = |\mathbf{u}_i^H \mathbf{a}|^2 \quad \text{for } i = 1, \ldots, n_y,$$

where $\mathbf{y}_i \in \mathbb{R}_+$ and $\mathbf{u}_i \in \mathbb{C}^{n_a}$ are known and $(\cdot)^H$ denotes the Hermitian transpose of a vector (matrix). For the sake of brevity, the following compact notation will be used in this paper to denote this general noise-free phase retrieval problem:

$$\text{find } \mathbf{a} \in \mathbb{C}^{n_a} \text{ such that } \mathbf{y} = |U\mathbf{a}|^2, \tag{1}$$

where $\mathbf{y} \in \mathbb{R}_+^{n_y}$ are the measurements and $U \in \mathbb{C}^{n_y \times n_a}$ is the propagation matrix. With noise on the measurements y_i, we consider the following related optimization problem:

$$\min_{\mathbf{a} \in \mathbb{C}^{n_a}} \|\mathbf{y} - |U\mathbf{a}|^2\|, \tag{2}$$

where $\|\cdot\|$ denotes a vector norm of interest.

The sparse variant of the phase retrieval problem corresponds to the case that the unknown parameter $\mathbf{a}$ is a sparse vector. A special case of this problem is when the measurements are the magnitude of the Fourier transform of multiples of $\mathbf{a}$ with certain phase diversity patterns. A number of algorithms utilizing the Fourier transform have been proposed for solving this class of phase retrieval problems [3–5].

The fundamental nature of Eq. (1) has given rise to a wide variety of solution methods that have been developed for specific variants of this problem since the observation of Sayre in 1952 that phase information of a scattered wave may be recovered from the recorded intensity patterns at and between Bragg peaks of a diffracted wave [6]. Direct methods [7] usually use insights about the crystallographic structure and randomization to search for the missing phase information. The requirement of such *a priori* structural information and the expensive computational complexity often limit the application of these methods in practice.

A second class of methods first devised by Gerchberg and Saxton [8] and Fienup [3] can be described as variants of the method of alternating projections on certain sets defined by the constraints. For an overview of these methods and latter refinements, we refer the reader to [4,9].

In [10], Eq. (1) is relaxed to a convex optimization problem. The inclusion of the sparsity constraint in the same framework

of convex relaxations has been considered in [11]. However, as reported in [5], the combination of matrix lifting and semidefinite programming (SDP) makes this method not suitable for large-scale problems. To deal with large-scale problems, the authors of [5] have proposed an iterative solution method, called greedy sparse phase retrieval (GESPAR), which appears to yield the promising recovery of very sparse signals. However, this method consists of a heuristic search for the support of $\mathbf{a}$ in combination with a variant of the Gauss–Newton method, whose computational complexity is often expensive. These algorithmic features are potential drawbacks of GESPAR.

In this paper, we propose a sequence of convex relaxations for the phase retrieval problem in Eq. (1). Contrary to existing convex relaxation schemes such as those proposed in [10,11], matrix lifting is not required in our strategy. The obtained convex problems are affine in the unknown parameter vector $\mathbf{a}$. Contrary to [12], our strategy does not require the tuning of regularization parameters when the measurements are corrupted by noise. We then present an algorithm based on the alternating direction method of multipliers (ADMM) that can solve the resulting optimization problems effectively. This potentially addresses the restriction of current SDP-based methods to only relatively small-scale problems.

In Section 2 we formulate the estimation problem of our interest for both zonal and modal forms. In Section 3 we propose an algorithm for solving this problem. The algorithm is based on iteratively minimizing a nuclear norm. The nuclear norm of a matrix is the sum of its singular values. Its benefit in optimization is that it is used as a convex relaxation to the rank function [13]. The convexity enables direct use of standard software libraries for solving convex optimization problems. However, since it is a computationally heavy minimization problem, we suggest an ADMM-based algorithm in Section 4 that exploits the problem structure and is therefore more efficient in practical cases. This ADMM algorithm features two minimization problems whose solutions can be computed exactly and with complexity $\mathcal{O}(n_y n_a)$, where n_y is the number of measurements, and n_a is the number of unknown variables. To find these solutions, either a least-squares problem has to be solved or the singular value decompositions of 2×2 matrices have to be computed. Analytic solutions for the ADMM algorithm update steps will be presented in Subsections A and B. The convergence behavior of the algorithm proposed in Section 3 is analyzed in Section 5. Compared to the other sections, the mathematical analysis in this section is more involved, which is often the case for convergence analyses. In Section 6 we describe and discuss the results of a number of numerical experiments that demonstrate the promising performances of our algorithms. We end with concluding remarks in Section 7.

2. WAVEFRONT ESTIMATION FROM INTENSITY MEASUREMENTS

The problem of phase retrieval from the point spread function images can be approached from two directions. We take the opportunity to present them in a unified way. We first describe the problem in zonal form and then in modal form. The modal form approach used in this paper seems less popular than that of the zonal form.

A. Problem Formulation in Zonal Form

In [1] it was shown that reconstructing the wavefront from charge-coupled device (CCD) recorded images of a point source may also be formulated as a phase retrieval problem. These recorded images are called *point spread functions (PSFs)*. As such approaches avoid the requirement of extra hardware to sense the wavefront, such as a Shack–Hartmann wavefront sensor, the problem is relevant and summarized here.

The PSF is derived from the magnitude of the Fourier transform of the generalized pupil function (GPF). For an aberrated optical system, the GPF is defined as the complex-valued function [14],

$$P(\rho,\theta)=\mathbf{A}(\rho,\theta)e^{j\phi(\rho,\theta)}, \tag{3}$$

where ρ (radius) and θ (angle) specify the normalized polar coordinates in the exit pupil plane of the optical system. In Eq. (3), $\mathbf{A}(\rho,\theta)$ is the amplitude apodization function, and $\phi(\rho,\theta)$ is the phase aberration function.

The aim of the wavefront reconstruction problem is to estimate $\phi(\rho,\theta)$. Once this phase aberration of an optical system has been estimated, it can be corrected by using phase modulating devices such as deformable mirrors.

In order to estimate $\phi(\rho,\theta)$, a known phase diversity pattern $\phi_d(\rho,\theta)$ can be introduced (e.g., by using a deformable mirror) to transform the GPF in a controlled manner into the aberrated GPF,

$$P_d(\rho,\theta)=\mathbf{A}(\rho,\theta)e^{j\phi(\rho,\theta)}e^{j\phi_d(\rho,\theta)}. \tag{4}$$

The noise-free intensity pattern of $P_d(\rho,\theta)$ measured at the image plane is denoted as

$$\mathbf{y}_d=|\mathcal{F}\{\mathbf{A}(\rho,\theta)e^{j\phi(\rho,\theta)}e^{j\phi_d(\rho,\theta)}\}|^2. \tag{5}$$

If we sample the function $P_d(\rho,\theta)$ at points corresponding to a square grid of size $m\times m$ on the pupil plane, then $\mathbf{A}(\rho,\theta)$, $\phi_d(\rho,\theta)$, and $\phi(\rho,\theta)$ are square matrices of that size.

Let us define vect$(\cdot)$, the vectorization operator, such that vect(Z) yields the vector obtained by stacking the columns of matrix Z into a column vector. The inverse operator vect$^{-1}(\cdot)$, which maps a column vector of size m^2 to a square matrix of size $m\times m$, is also well defined. Let in particular the matrix Z and the vector $\mathbf{a}$ be defined as

$$Z=\mathbf{A}(\rho,\theta)e^{j\phi(\rho,\theta)}\in\mathbb{C}^{m\times m},\qquad \mathbf{a}=\text{vect}(Z)\in\mathbb{C}^{m^2}.$$

With the definition of the vector $\mathbf{p}_d$,

$$\mathbf{p}_d=\text{vect}(e^{j\phi_d(\rho,\theta)})\in\mathbb{C}^{m^2},$$

and with $D_d=d(\mathbf{p}_d)\in\mathbb{C}^{m^2\times m^2}$ as the diagonal matrix with diagonal entries taken from the vector $\mathbf{p}_d$, we can write the noise-free intensity measurements in Eq. (5) as

$$\mathbf{y}_d=|\mathcal{F}\{e^{j\phi_d(\rho,\theta)}Z\}|^2=|\mathcal{F}\{\text{vect}^{-1}(D_d\mathbf{a})\}|^2.$$

As the Fourier transform is a linear operator, we can write our noise-free intensity measurements in the form

$$\mathbf{y}_d=|U_d\mathbf{a}|^2, \tag{6}$$

where in this case U_d is a unitary matrix.

By stacking the vectors $\mathbf{y}_d$ and the matrices U_d obtained from the n_d images with n_d different phase diversities,

correspondingly, into the vector $\mathbf{y}$ and the matrix U (of size $n_d m^2 \times m^2$), the problem of finding $\mathbf{a}$ from noise-free intensity measurements can be formulated as in Eq. (1), and that from noisy measurements can be formulated as in Eq. (2) for $n_a = m^2$ and $n_y = n_d m^2$.

It is worth noting that the dimension of the unknown $\mathbf{a}$ with m in the range of a couple of hundreds turns this problem into a non-convex large-scale optimization problem. For such a problem, the implementation of PhaseLift [12] using standard SDP using libraries like MOSEK [15] will not be tractable because of the large matrix dimensions of the unknown quantity. If we assume that the computational complexity of semidefinite programming with matrix constraints of size $n \times n$ increases with $\mathcal{O}(n^6)$ [16], then a naive implementation of the PhaseLift method applied to Eq. (2) involving a single image has a worst-case computational complexity of $\mathcal{O}(m^{12})$.

B. Problem Formulation in Modal Form

In general, only approximate solutions can be expected for a phase retrieval problem. In the modal form of the phase retrieval problem, also considered in [1] for extended Nijboer–Zernike (ENZ) basis functions, the GPF is assumed to be well approximated by a weighted sum of basis functions. We make use of real-valued radial basis functions [17] with complex coefficients to approximate the GPF. These are studied in the scope of wavefront estimation in [18], and an illustration of these basis function on a 4×4 grid in the pupil plane is given in Fig. 1.

Switching from the polar coordinates (ρ, θ) to the Cartesian coordinates (x, y) in the pupil plane, let us consider the radial basis functions and the approximate GPF given by

$$G_i(x,y) = \chi(x,y) e^{-\lambda_i((x-x_i)^2+(y-y_i)^2)},$$
$$P(x,y) \approx \tilde{P}(x,y,\mathbf{a}) = \sum_{i=1}^{n_a} a_i G_i(x,y), \tag{7}$$

where (x_i, y_i) are the centers of basis functions $G_i(x,y)$, $a_i \in \mathbb{C}$, $\lambda_i \in \mathbb{R}_+$ determines the spread of that function, $\chi(x,y)$ denotes the support of the aperture, and $\mathbf{a}$ is the coefficient parameter vector to be estimated. The parameters λ_i are usually taken equal for all basis functions, and for their tuning we refer to [18].

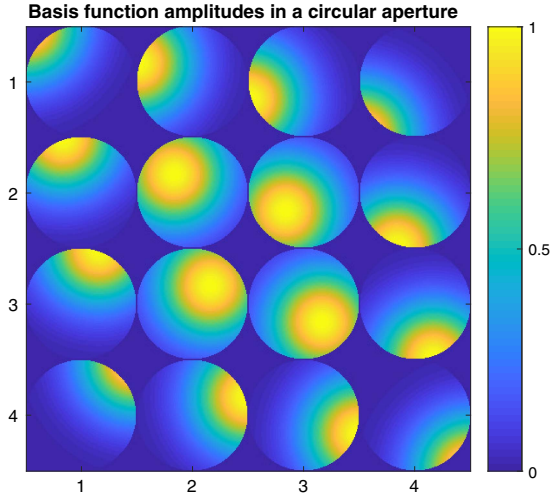

Fig. 1. 16 radial basis functions with centers in a 4×4 grid, with circular aperture support.

The aberrated GPF corresponding to the introduction of phase diversity ϕ_d is

$$\tilde{P}_d(x,y,\mathbf{a},\phi_d) = \sum_{i=1}^{n_a} a_i G_i(x,y) e^{j\phi_d(x,y)}. \tag{8}$$

The normalized complex PSF is the two-dimensional Fourier transform of the GPF [19,20]. The aberrated PSF corresponding to the aberrated GPF in Eq. (8) is given as

$$p_d(u,v) = \sum_{i=1}^{n_a} a_i \mathcal{F}\{G_i(x,y) e^{j\phi_d(x,y)}\} = \sum_{i=1}^{n_a} a_i U_{d,i}(u,v), \tag{9}$$

where (u, v) are the Cartesian coordinates in the image plane of the optical system.

We now drop the dependency on the coordinates and vectorize expression Eq. (9) for all n_d diversities that have been applied to obtain the following compact form of a single matrix-vector multiplication,

$$\mathbf{p} = U\mathbf{a}. \tag{10}$$

The vector $\mathbf{p}$ is the obtained vectorization and combination over all the aberrated PSFs, and the matrix U is the vectorized and concatenated version of the functions $U_{d,i}$ sampled on a grid of size $m \times m$.

Let the intensity of the PSFs be recorded on the corresponding grid of pixels of size $m \times m$, and let the vectorization of this intensity pattern for different phase diversities be concatenated into the vector $\mathbf{y}$. We can again formulate the problem of finding $\mathbf{a}$ from noise-free intensity measurements as in Eq. (1) and from noisy measurements as in Eq. (2) for $n_y = m^2 n_d$.

It is worth noting that the dimension of $\mathbf{a}$ is not dependent on the size of the sample grid (the size of the problem). This is the fundamental advantage of the modal form formulation over the zonal form one, for which the size of $\mathbf{a}$ directly depends on the size of the problem, i.e., $n_a = m^2$.

In this paper two steps are combined to deal with the large-scale nature of optimization Eq. (2):

1. The unknown pupil function $P(\rho, \theta)$ can be represented as a linear combination of a number of basis functions. In [1], the ENZ basis functions were used, whereas in [18] radial basis functions were used instead of ENZ ones. The radial basis functions are used here, since [18] demonstrated their advantages over the ENZ type.
2. A new strategy is proposed for solving optimization Eq. (1) via a sequence of convex optimization problems. Each of the subproblems can be solved effectively by an iterative ADMM algorithm that exploits the problem structure.

In the following we assume that the problem is normalized such that all entries of $\mathbf{y}$ have values between 0 and 1.

3. CONVEX OPTIMIZATION-BASED PHASE RETRIEVAL ALGORITHM

Equation (1) is equivalent to a rank constraint. Define the matrix-valued function

$$L(A,B,C,X,Y)=\begin{pmatrix} C+AY+XB+XY & A+X \\ B+Y & I \end{pmatrix}, \tag{11}$$

where I is the identity matrix of appropriate size. Let $\mathbf{b}\in\mathbb{C}^{n_a}$ be a coefficient vector. For notational convenience, we will denote

$$M(U,\mathbf{a},\mathbf{b},\mathbf{y}) := L(d(\mathbf{a}^H U^H), d(U\mathbf{a}), d(\mathbf{y}), d(\mathbf{b}^H U^H), d(U\mathbf{b})).$$

Our proposed algorithm in this paper relies on the following fundamental result.

Lemma 1. [21] *For any* $\mathbf{b}\in\mathbb{C}^{n_a}$, *the constraint* $\mathbf{y}=|U\mathbf{a}|^2$ *is equivalent to the constraint*

$$\operatorname{rank}(M(U,\mathbf{a},\mathbf{b},\mathbf{y}))=n_y.$$

For addressing problem Eq. (2), Lemma 1 suggests a consideration of the following approximate problem for a user-selected parameter vector $\mathbf{b}$:

$$\min_{\mathbf{a}\in\mathbb{C}^{n_a}} \operatorname{rank}(M(U,\mathbf{a},\mathbf{b},\mathbf{y})). \tag{12}$$

Since Eq. (12) is a non-convex problem, and to anticipate the presence of measurement noise, we propose to solve the following convex optimization problem:

$$\min_{\mathbf{a}\in\mathbb{C}^{n_a}} f(\mathbf{a}) := \|M(U,\mathbf{a},\mathbf{b},\mathbf{y})\|_*, \tag{13}$$

where $\|\cdot\|_*$ denotes the nuclear norm of a matrix, the sum of its singular values [13,22]. The motivation to choose M (and L) in the structure of Eq. (11) is that it is affine in the unknown $\mathbf{a}$. By relaxing the rank constraint into Eq. (13), we obtain a convex relaxation without "lifting" (substituting) the variables, as is the case with PhaseLift. One advantage is that the solution for $\mathbf{a}$ can be easily influenced if we have prior knowledge. For example, in the case that prior knowledge on the problem indicates that $\mathbf{a}$ is a sparse vector, the objective function in Eq. (13) can easily be extended with an ℓ_1-regularization to stimulate sparse solutions, since the vector $\mathbf{a}$ appears affinely in $M(U,\mathbf{a},\mathbf{b},\mathbf{y})$,

$$\min_{\mathbf{a}\in\mathbb{C}^{n_a}} f(\mathbf{a})+\lambda\|\mathbf{a}\|_1, \tag{14}$$

for a regularization parameter λ.

Note that for $\mathbf{b}=-\mathbf{a}$,

$$\|M(U,\mathbf{a},-\mathbf{a},\mathbf{y})\|_* = \|\mathbf{y}-|U\mathbf{a}|^2\|_1 + n_y. \tag{15}$$

Since the result of the optimization in Eq. (13) might not produce a desired solution sufficiently fitting the measurements, we propose the iterative convex optimization-based phase retrieval (COPR) algorithm, outlined in Algorithm 1.

Algorithm 1: Convex Optimization-Based Phase Retrieval (COPR)

```
procedure COPR(b, τ)                          ▷ Some guess for b
    while ‖y − |Ua|²‖₁ > τ do                 ▷ Termination criterion
        a₊ ∈ arg min_a ‖M(U, a, b, y)‖*
        b₊ ← −a₊
```

The nuclear norm is a convex function, and standard software like YALMIP [23] or CVX [24] can be used to concisely implement Algorithm 1. However, the nuclear norm minimization in Algorithm 1 is the main computational burden for an implementation. Usual implementations of the nuclear norm involve semidefinite constraints and require a semidefinite optimization solver. If we assume that their computational complexity increases with $\mathcal{O}(n^6)$ [16] with constraint on matrices of size $n\times n$, then minimizing the nuclear norm of the matrix $M(U,\mathbf{a},\mathbf{b},\mathbf{y})$ of size $2n_y\times 2n_y$ is computationally infeasible, even for relatively small-scale problems. Therefore, we propose a tailored ADMM algorithm of which the computational complexity of the iterations scales $\mathcal{O}(n_y n_a)$ and which requires the inverse of a matrix of size $2n_a\times 2n_a$ for every iteration of Algorithm 1.

4. EFFICIENT COMPUTATION OF THE SOLUTION TO EQ. (13)

The minimization problem Eq. (13) can be reformulated as

$$\min_{X,\mathbf{a}} \|X\|_* \quad \text{subject to } X=M(U,\mathbf{a},\mathbf{b},\mathbf{y}). \tag{16}$$

Applying the ADMM optimization technique [25,26] to the constraint optimization problem Eq. (16), we obtain the steps in Algorithm 2.

Algorithm 2: ADMM Algorithm for Solving Eq. (16)

```
procedure NN-ADMM(b, y, ρ, τ)
    a ← −b
    X ← M(U, a, b, y)
    Y ← 0
    while |‖M(U, a₊, b, y)‖* − ‖M(U, a, b, y)‖*| > τ do
        a₊ ∈
```

$$\arg\min_{\mathbf{a}} \left\| X - M(U,\mathbf{a},\mathbf{b},\mathbf{y}) + \frac{1}{\rho}Y \right\|_F^2 \tag{17}$$

```
        X₊ ∈
```

$$\arg\min_{X} \|X\|_* + \frac{\rho}{2}\left\| X - M(U,\mathbf{a}_+,\mathbf{b},\mathbf{y}) + \frac{1}{\rho}Y \right\|_F^2 \tag{18}$$

```
        Y₊ ← Y + ρ(X₊ − M(U, a₊, b, y))
        update ρ according to the rules in [25]
```

The advantage of using this ADMM formulation is that both of the update steps in Eqs. (17) and (18) have solutions that can be computed analytically. The efficient computation of the solutions is described in the following two subsections.

A. Efficient Computation of the Solution to Eq. (17)

Upon inspection of Eq. (17), we see that this is a complex-valued standard least-squares problem, since $M(U,\mathbf{a},\mathbf{b},\mathbf{y})$ is parameterized affinely in $\mathbf{a}$. Let $\mathcal{R}(\cdot)$ and $\mathcal{I}(\cdot)$ denote the real and the imaginary parts of a complex object, respectively. Let the subscripts $(\cdot)_1$, $(\cdot)_2$, and $(\cdot)_3$ denote the top-left, top-right, and bottom-left submatrices, respectively, according to Eq. (11). Define

$$Z = X+\frac{1}{\rho}Y, \qquad X = d(b^H U^H).$$

In the sequel, let $\widehat{d}(P)$ denote the vector with the diagonal entries of a square matrix P.

Reordering the elements in Eq. (17), separating the real and the imaginary parts, removing all matrix elements in the

argument of the Frobenius norm that do not depend on $\mathbf{a}$, and vectorizing the result gives the following least-squares problem:

$$\min_{\mathbf{x}} \|\mathbf{u}_{\mathrm{ADMM}} - \mathbf{u}_{\mathrm{COPR}} - AB\mathbf{x}\|_2^2. \tag{19}$$

The variables $\mathbf{u}_{\mathrm{ADMM}}, \mathbf{u}_{\mathrm{COPR}}, A, B$, and $\mathbf{x}$ are given by

$$\mathbf{u}_{\mathrm{ADMM}} = \begin{pmatrix} \widehat{d}(\mathcal{R}(\boldsymbol{Z}_1)) \\ \widehat{d}(\mathcal{R}(\boldsymbol{Z}_2)) \\ \widehat{d}(\mathcal{R}(\boldsymbol{Z}_3)) \\ \widehat{d}(\mathcal{I}(\boldsymbol{Z}_2)) \\ \widehat{d}(\mathcal{I}(\boldsymbol{Z}_3)) \end{pmatrix}, \quad \mathbf{u}_{\mathrm{COPR}} = \begin{pmatrix} \mathbf{y} + \widehat{d}(|X|^2) \\ \widehat{d}(\mathcal{R}(X)) \\ \widehat{d}(\mathcal{R}(X)) \\ \widehat{d}(\mathcal{I}(X)) \\ -\widehat{d}(\mathcal{I}(X)) \end{pmatrix},$$

$$A = \begin{pmatrix} 2\mathcal{R}(X) & 2\mathcal{I}(X) \\ I & 0 \\ I & 0 \\ 0 & I \\ 0 & -I \end{pmatrix}, \quad B = \begin{pmatrix} \mathcal{R}(U) & -\mathcal{I}(U) \\ -\mathcal{I}(U) & -\mathcal{R}(U) \end{pmatrix}, \tag{20}$$

and $\mathbf{x} = (\mathcal{R}(\mathbf{a})^T \quad \mathcal{I}(\mathbf{a})^T)^T$. This means that the optimal solution to Eq. (19) is given by

$$\mathbf{x}^* = (B^T A^T A B)^{-1} B^T A^T (\mathbf{u}_{\mathrm{ADMM}} - \mathbf{u}_{\mathrm{COPR}}).$$

During the ADMM iterations, only $\mathbf{u}_{\mathrm{ADMM}}$ changes. The inverse $(B^T A^T A B)^{-1}$ has to be computed once for every iteration of Algorithm 1 (i.e., it remains constant throughout the ADMM iterations). Since the complexity of computing an inverse is $\mathcal{O}(n^3)$ for matrices of size $n \times n$, the computational complexity of this inverse process scales cubically with the number of basis functions.

Once this inverse matrix is obtained, the optimal solution to the least-squares problem in Eq. (19) can be computed by a simple matrix-vector multiplication, whose complexity scales with $\mathcal{O}(n_y n_a)$.

Note that in the case that the objective term includes regularization as in Eq. (14), the optimization in Eq. (19) should be modified appropriately to include the additive regularization term $\lambda\|\mathbf{a}\|_1$.

B. Efficient Computation of the Solution to Eq. (18)

The optimization in Eq. (18) is of the form

$$\arg\min_{\boldsymbol{X}} \|X\|_* + \lambda\|X - C\|_F^2. \tag{21}$$

Let $C = U_C \Sigma_C V_C^T$ be the singular value decomposition (SVD) of $C \in \mathbb{C}^{2n_y \times 2n_a}$.

Lemma 2. *The solution $\boldsymbol{X}$ to* Eq. (21) *has singular vectors U_C and V_C.*

proof. Let $X = U_X \Sigma_X V_X^T$ be a singular value decomposition of X. Then

$$\|X\|_* + \lambda\|X - C\|_F^2 = \mathrm{trace}(\Sigma_X) + \lambda(\langle X, X\rangle + \langle C, C\rangle - 2\langle X, C\rangle).$$

Using Von Neumann's trace inequality, we get

$$\min_X(\mathrm{trace}(\Sigma_X) + \lambda(\langle X, X\rangle + \langle C, C\rangle - 2\langle X, C\rangle)) \geq \min_X(\mathrm{trace}(\Sigma_X) + \lambda(\langle X, X\rangle + \langle C, C\rangle - 2\mathrm{trace}(\Sigma_X \Sigma_C))),$$

which with equality holds true when C and X are simultaneously unitarily diagonalizable. The optimal solution $\boldsymbol{X}$ to Eq. (21) therefore has the same singular vectors as C, i.e., $U_{\boldsymbol{X}} = U_C$, $V_{\boldsymbol{X}} = V_C$. □

Denote the singular values of C in descending order as $\sigma_{C,1}, \ldots, \sigma_{C,2n_y}$, and those of X similarly. Thanks to Lemma 2, Eq. (21) can be simplified to

$$\arg\min_{\sigma_{X,i}} \sum_{i=1}^{2n_y} (\sigma_{X,i} + \lambda(\sigma_{X,i} - \sigma_{C,i})^2). \tag{22}$$

This problem is completely decoupled in $\sigma_{X,i}$, and the optimal solution to Eq. (22) is computed with

$$\sigma_{\boldsymbol{X},i} = \max\left(0, \sigma_{C,i} - \frac{1}{2\lambda}\right), \qquad i = 1, \ldots, 2n_y.$$

By row and column permutations, matrix C is block-diagonal with blocks of size 2×2. The SVD of this permuted matrix therefore involves block-diagonal matrices U_C, Σ_C, and V_C, and these blocks can be obtained separately and in parallel. Since the blocks are of size 2×2, the SVD can be obtained analytically.

This shows that a valid SVD can be computed very efficiently in $\mathcal{O}(1)$, that is, in theory, in a computation time independent of the number of pixels in the image, the number of images taken, or of the number of basis functions.

5. CONVERGENCE ANALYSIS OF ALGORITHM 1

Algorithm 1 can be reformulated as a Picard iteration $\mathbf{a}_{k+1} \in T(\mathbf{a}_k)$, where the fixed point operator $T: \mathbb{C}^{n_a} \to \mathbb{C}^{n_a}$ is given by

$$T(\mathbf{a}) = \arg\min_{\mathbf{x} \in \mathbb{C}^{n_a}} \|M(U, \mathbf{x}, -\mathbf{a}, \mathbf{y})\|_*. \tag{23}$$

Our subsequent analysis will show that the set of fixed points, Fix T, the set of $\mathbf{a}$s for which $\mathbf{a} = T(\mathbf{a})$, of T is in general non-convex and as a result, iterations generated by T cannot be *Fejér monotone* (Definition 5.1 of [27]) with respect to Fix T. That is, each new iterate is not guaranteed to be closer to *all* fixed points in Fix T. Therefore, the widely known convergence theory based on the properties of *Fejér monotone operators* and *averaging operators* is not applicable to the operator T given in Eq. (23).

In this section, we make an attempt to prove the convergence of Algorithm 1, which has been observed from our numerical experiments, via a relatively newly developed convergence theory based on the theory of *pointwise almost averaging operators* [28]. It is worth mentioning that we are not aware of any other analysis schemes addressing the convergence of Picard iterations generated by general *non-averaging* fixed point operators. Our discussion consists of two stages. Based on the convergence theory developed in [28], we first formulate a convergence criterion for Algorithm 1 (Proposition 5.1) under rather abstract assumptions on the operator T. Due to the

highly complicated structure of the nuclear norm of a general complex matrix, we are unable to verify these mathematical conditions for general matrices U. However, we will verify that they are well satisfied in the case that U is a unitary matrix (Theorem 5.2). From the latter result, we heuristically hope that Algorithm 1 still enjoys the convergence result when the matrix U is close to being unitary in a certain sense. In Section 6 we demonstrate that convergence is obtained in practice for the imaging case.

It is a common prerequisite for analyzing the local convergence of a fixed point algorithm that the set of solutions to the original problem is non-empty. That is, there exists $\mathbf{a} \in \mathbb{C}^{n_a}$ such that $\mathbf{y} = |U\mathbf{a}|^2$. Before stating the convergence result, we need to verify that the fixed point set of T is non-empty.

Lemma 3. *The fixed point operator T defined at* Eq. (23) *holds*

$$\{\mathbf{a} | \mathbf{y} = |U\mathbf{a}|^2\} \subseteq \operatorname{Fix} T := \{\mathbf{a} \in \mathbb{C}^{n_a} | \mathbf{a} \in T(\mathbf{a})\}.$$

proof. See Appendix A. □

The next proposition provides an abstract convergence result for Algorithm 1. Fix T is supposed to be closed. In the sequel, the metric projection associated with a set Ω is denoted P_Ω,

$$P_\Omega(x) := \{\omega \in \Omega | \|x - \omega\| = \operatorname{dist}(x, \Omega)\}, \quad \forall x.$$

Proposition 5.1. (*simplified version of Theorem 2.2 of* [28]) *Let $S \subset \operatorname{Fix} T$ be closed with $T(\mathbf{a}^*) \subset \operatorname{Fix} T$ for all $\mathbf{a}^* \in S$, and let W be a neighborhood of S. Suppose that T satisfies the following conditions.*

(i) *T is pointwise averaging at every point of S with constant $\alpha \in (0, 1)$ on W. That is, for all $\mathbf{a} \in W$, $\mathbf{a}_+ \in T(\mathbf{a})$, $\mathbf{a}^* \in P_S(\mathbf{a})$, and $\mathbf{a}_+^* \in T(\mathbf{a}^*)$,*

$$\|\mathbf{a}_+ - \mathbf{a}_+^*\|^2 \leq \|\mathbf{a} - \mathbf{a}^*\|^2 - \frac{1-\alpha}{\alpha}\|(\mathbf{a}_+ - \mathbf{a}) - (\mathbf{a}_+^* - \mathbf{a}^*)\|^2. \tag{24}$$

(ii) *The set-valued mapping $\psi := T - \mathrm{Id}$ is metrically* subregular *on W for 0 with constant $\gamma > 0$, where* Id *is the identity mapping. That is,*

$$\gamma \operatorname{dist}(\mathbf{a}, \psi^{-1}(0)) \leq \operatorname{dist}(0, \psi(\mathbf{a})), \quad \forall \mathbf{a} \in W. \tag{25}$$

(iii) *It holds $\operatorname{dist}(\mathbf{a}, S) \leq \operatorname{dist}(\mathbf{a}, \operatorname{Fix} T)$ for all $\mathbf{a} \in W$.*

Then, all Picard iterations $\mathbf{a}_{k+1} \in T(\mathbf{a}_k)$ starting in W satisfy $\operatorname{dist}(\mathbf{a}_k, S) \to 0$ as $k \to \infty$, at least linearly.

The pointwise property, instead of the standard averaged property, imposed in (i) of Propositon 5.1 allows us to deal with the intrinsic non-convexity of the fixed point set Fix T. The metric subregularity assumption imposed in (ii) technically ensures adequate progression of the iterates relative to the distance from the current iterate to the fixed point set. This is not only a technical assumption but also a necessary condition for local linear convergence of a fixed point algorithm, Theorem 3.12 of [29]. Condition (iii) is, on one hand, a technical assumption and becomes redundant when $S = \operatorname{Fix} T$. On the other hand, the set S allows one to exclude from the analysis possible *inhomogeneous* fixed points of T, at which the algorithm often exposes weird convergence behavior (see Example 2.1 of [28]).

The size of neighborhood W appearing in Proposition 5.1 indicates the robustness of the algorithm in terms of erroneous input (the distance from the starting point to the nearest solution).

We now apply the abstract result of Proposition 5.1 to the following special but important case.

Theorem 5.2. *Let $U \in \mathbb{C}^{n_a \times n_a}$ be unitary and $\mathbf{a}^* \in \mathbb{C}^{n_a}$ be such that $|U\mathbf{a}^*|^2 = \mathbf{y}$. Then, every Picard iteration generated by Algorithm 1 $\mathbf{a}_{k+1} \in T(\mathbf{a}_k)$ starting sufficiently close to $\mathbf{a}^*$ converges linearly to a point $\tilde{\mathbf{a}} \in \operatorname{Fix} T$, satisfying $|U\tilde{\mathbf{a}}|^2 = \mathbf{y}$.*

proof. See Appendix B. □

6. NUMERICAL EXPERIMENTS

Four important numerical aspects of the COPR algorithm, including convergence, flexibility, complexity, and robustness, are tested on relevant problems. First we discuss convergence and the number of iterations of the COPR and the ADMM algorithms. Second, we demonstrate the flexibility of the convex relaxation by comparing the COPR algorithm with an added ℓ_1-regularization to the PhaseLift method [12] and to the compressive sensing phase retrieval (CPRL) method in [11] on an under-determined sparse estimation problem. Then we compare the practically observed computational complexity of COPR and a naive implementation of PhaseLift [12]. Finally, we investigate the robustness of COPR relative to noise in a Monte Carlo simulation for 25 and 100 basis functions. We compare four algorithms: COPR, PhaseLift [12], a basic alternating projections method (Section 4.3 in [12]), and an averaged projections method based on [30]. We note that the latter method fundamentally employs the Fourier transform at every iteration and hence is in general not applicable for phase retrieval in the modal form.

A. Convergence

The while loops in Algorithms 1 and 2 can be run for a fixed number of iterations. Figure 2 shows four such combinations for a typical problem with five images of size 256×256, of which a subset of 25×25 pixels per image is used, and 64 basis functions. All cases are identically initialized with coefficients that best approximate a flat wavefront. As can be seen from the figure and the line with a square marker, only one COPR iteration is necessary here, as the ADMM algorithm slowly converges towards 0. However, stopping the ADMM algorithm after a limited number of iterations and having more

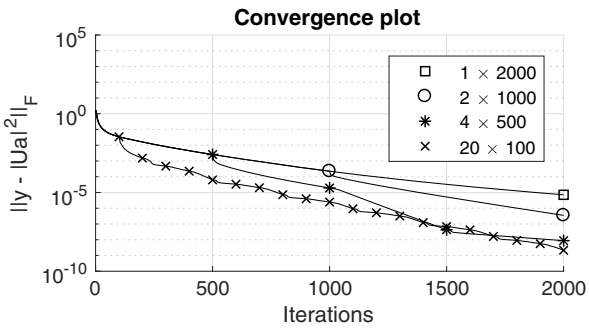

Fig. 2. Convergence plot for four different combinations of COPR iterations and ADMM iterations. Denoted in the legend are first the number of COPR iterations and then the number of ADMM iterations used to solve Eq. (16) in each COPR iteration. Markers denote each new COPR iteration.

than one COPR iteration can have a clear benefit, since faster convergence is achieved this way.

B. Application of COPR to Compressive Sensing Problems

The first problem is to estimate 16 coefficients from 8 measurements, where the optimal vector is known to be sparse.

We generate a sparse coefficient vector $\mathbf{a}$ with two randomly generated non-zero complex elements. We generate two images ($n_d = 2$, $m = 128$) by applying two different amounts of defocus with Zernike coefficients $-\frac{\pi}{8}$ and $\frac{\pi}{8}$, respectively. From each image we use the center 2×2 pixels, resulting in a total of $n_y = 8$ measurements.

The applied algorithms are the COPR algorithm, the COPR algorithm with an additional ℓ_1-regularization, the PhaseLift algorithm [12], and the CPRL algorithm of [11]. The results are displayed in Fig. 3. As can be seen from the figure, COPR and PhaseLift fail to retrieve the correct solution. The CPRL method and the regularized COPR algorithm compute the correct solution.

C. Computational Complexity

The second problem demonstrates the trends of the required computation time when the number of estimated coefficients increases. The underlying estimation problem consists of seven images with different amounts of defocus applied as phase diversity, where each image is of size 64 by 64 pixels. A subset of 20 by 20 pixels of each image is used in the estimation. We compare in Fig. 4 the COPR algorithm to the PhaseLift algorithm, which is implemented according to the optimization problem (2.5) in [12]. For PhaseLift, the reported time is the time it takes the MOSEK solver [15] to solve the optimization problem. This does not include the time taken by YALMIP [23] to convert the problem as given to the solver-specific form. For COPR, the initial guesses for the coefficients are drawn randomly from a Gaussian distribution, the number of iterations is set beforehand according to the convergence to the correct solution, and the total time is recorded. The implementation of COPR does not exploit the parallelism referred to in Section 4.B. By convergence, we mean that the estimated vector $\hat{\mathbf{a}}$ satisfies the tolerance criterion

$$\min_{c\in\mathbb{C},\,|c|=1} \|c\hat{\mathbf{a}} - \mathbf{a}^*\|_2^2 \le 10^{-5}, \tag{26}$$

where $\mathbf{a}^*$ is the exact solution.

The minimization over parameter c ensures that the (unobservable) piston mode in the phase is canceled (Let $(\hat{\mathbf{a}} \quad \mathbf{a}^*) = \mathrm{QR}$ be the QR decomposition. Then $\angle c^* = \angle\frac{R_{12}}{R_{11}}$). The computational complexity of PhaseLift is, as implemented, approximately $\mathcal{O}(n^4)$. The MOSEK solver ran into numerical issues for more than 25 estimated parameters. The COPR algorithm's computational complexity is approximately $\mathcal{O}(n)$. The better complexity is offset by a longer computation time for very small problems.

D. Robustness to Noise

When estimating an unknown phase aberration, it is more logical to evaluate the performance of the algorithm on its ability to estimate the phase and not the coefficients of the basis functions.

We assume the phase is randomly generated with a deformable mirror. Let $H \in \mathbb{R}^{m^2\times n_u}$ be the mirror's influence matrix and $\mathbf{u} \in \mathbb{R}^{n_u}$ be the input to the mirror's actuators, such that

$$\phi_{\mathrm{DM}} = H\mathbf{u}. \tag{27}$$

The input values u_i are drawn from the uniform distribution between 0 and 1. The mirror has $n_u = 44$ actuators, and the images have sides $m = 128$. The aperture radius is 0.4.

Five different defocus diversities are applied with Zernike coefficients uniformly spaced between $-\frac{\pi}{2}$ and $\frac{\pi}{2}$. Gaussian noise is added to the obtained images such that

$$\mathbf{y} = \max(0, |\mathcal{F}\{P_d(\rho,\theta)\}|^2 + \varepsilon), \qquad \varepsilon \in N(0,\sigma I), \tag{28}$$

and σ is the noise variance. No denoising methods were applied. The signal-to-noise ratio (SNR) is computed according to

$$10\log_{10}\frac{\|\mathbf{y} - |\mathcal{F}\{P_d(\rho,\theta)\}|^2\|_2^2}{\||\mathcal{F}\{P_d(\rho,\theta)\}|^2\|_2^2}. \tag{29}$$

The phase is estimated from $\mathbf{y}$ using four different algorithms. The first is the COPR algorithm. The second is the averaged projections (AvP) algorithm [30]. The AvP algorithm is an extension of the well-known Gerchberg–Saxton algorithm [31] for solving problems with multiple images and is in the same class of algorithms as the hybrid-input-output algorithm and the difference map algorithm [32,33]. This makes this algorithm relevant for comparison. The third is the alternating

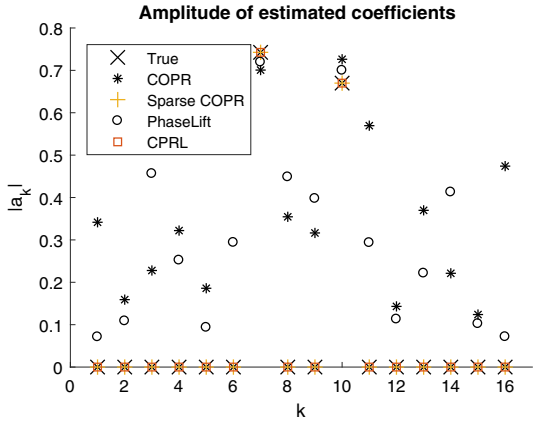

Fig. 3. Absolute values of 16 estimated coefficients according to four different algorithms.

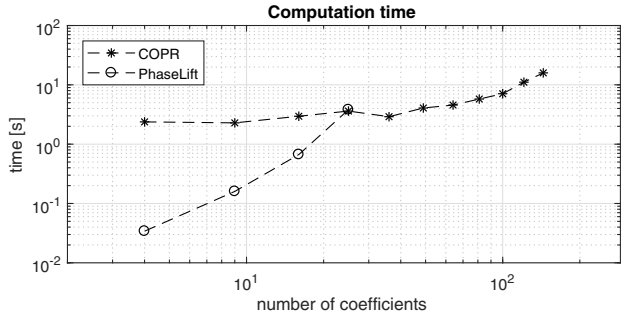

Fig. 4. Computation time comparison between PhaseLift and COPR for different numbers of coefficients.

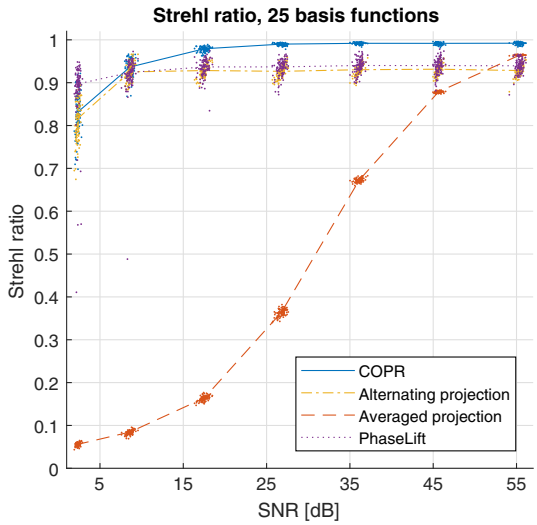

Fig. 5. Strehl ratio of the estimated phase aberration as a function of SNR.

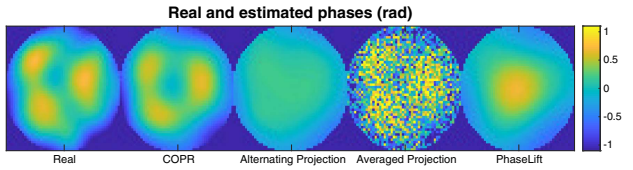

Fig. 6. Example PSF and phase estimates of the COPR, Alternating Projection [12], Averaged Projection [30], and PhaseLift [12] algorithms for three PSF measurements with an SNR of approximately 36 dB.

projections (AlP) method ([12], Section 4.3), and the fourth algorithm is the PhaseLift method [12].

The COPR, PhaseLift, and AlP methods are applied to estimate the phase using 25 basis functions, where the initial guesses for the coefficients are those coefficients that best approximate a flat wavefront. The AvP method is not based on the use of basis functions but on projection and the Fourier transform.

We make use of the Strehl ratio as a measure of optical quality. The Strehl ratio S is the ratio of the maximum intensity of the aberrated PSF and that of the unaberrated one and can be approximated with the expression of Mahajan,

$$S \approx e^{-\delta^2},$$

where $\delta = \|\phi_{\mathrm{DM}} - \hat{\phi}\|_2$ and the mean residual phase has been removed [34].

For every noise level, 100 different phases were generated with the deformable mirror model Eq. (27). The results are presented in Fig. 5. The resulting Strehl ratios are plotted with a trend line connecting the 50% quantiles. Figure 6 gives a qualitative comparison of the estimates for a single case.

In the case of PhaseLift, the tuning parameter that trades off measurement fit and the rank of the "lifted" matrix is tuned once and applied to all problems. This has the effect that the reported performance is not as high as it could be with optimal tuning for individual problems. This points to another advantage of COPR: the absence of tuning parameters aside from the choice of basis functions.

The two figures show that COPR is robust to noise and gives accurate phase estimates for a wide range of noise levels.

7. CONCLUDING REMARKS

The convex relaxations in solving the phase retrieval problem as proposed in Eq. (13) have the advantage over current convex relaxation methods, such as PhaseLift, that our strategy is affine in the coefficients that are to be estimated. This allows for easy extension of the proposed method to phase retrieval problems that incorporates prior knowledge on the coefficients by regularization of the objective function. One such successful extension is the regularization with the ℓ_1-norm to find sparse solutions, as demonstrated in Fig. 3.

In Section 4, an ADMM algorithm was proposed for efficient computation of the solution to Eq. (13). The result is that for the COPR algorithm a better computational complexity is observed compared to that of PhaseLift; see Fig. 4. COPR is also able to solve phase estimation problems with larger numbers of parameters.

The required computations are favorable both in computation time and accuracy (they have simple analytic solutions) and in worst-case scaling behavior $\mathcal{O}(n_y n_a)$ for every ADMM iteration, where n_y is the number of pixels, and n_a is the number of basis functions.

We discussed the convergence properties of the COPR algorithm in Section 5 and showed that for selected problems this convergence is linear or faster.

Finally, COPR has been shown to be robust against measurement noise and outperform the two projection-based methods whose naive forms are often sensitive to noise, as expected.

We are aware that in practice the performance of projection methods can be substantially better than what we have observed in this study, provided that appropriate denoising techniques are also applied. Keeping aside from the matter of using denoising techniques, we have chosen to compare the algorithms in their very definition forms.

APPENDIX A: PROOF OF LEMMA 3

proof. Let $\mathbf{a}$ satisfy $\mathbf{y} = |U\mathbf{a}|^2$. It suffices to check that $\mathbf{a} \in T(\mathbf{a})$. We first observe that

$$\operatorname{rank}(M(U, \mathbf{a}, -\mathbf{a}, \mathbf{y})) = \operatorname{rank}\begin{pmatrix} 0 & 0 \\ 0 & I_{n_y} \end{pmatrix} = n_y.$$

This means that $\mathbf{a}$ is a global minimizer of $\operatorname{rank} M(U, \mathbf{x}, -\mathbf{a}, \mathbf{y})$ as a function of $\mathbf{x} \in \mathbb{C}^{n_a}$. Since the nuclear norm $\|M(U, \mathbf{x}, -\mathbf{a}, \mathbf{y})\|_*$ is the *convex envelope* of the $\operatorname{rank} M(U, \mathbf{x}, -\mathbf{a}, \mathbf{y})$, they have the same global minimizers. Hence, $\mathbf{a}$ is also a global minimizer of $\|M(U, \mathbf{x}, -\mathbf{a}, \mathbf{y})\|_*$ as a function of $\mathbf{x}$, that is,

$$\mathbf{a} \in \arg\min_{\mathbf{x} \in \mathbb{C}^{n_a}} \|M(U, \mathbf{x}, -\mathbf{a}, \mathbf{y})\|_*.$$

In other words, $\mathbf{a} \in T(\mathbf{a})$, and the proof is complete. □

APPENDIX B: PROOF OF THEOREM 5.2

Lemma 4 will serve as the basic step for proving Theorem 5.2.

Lemma 4. *Let* $U = I_{n_a}$ *and* $\mathbf{a}^* \in \mathbb{C}^{n_a}$ *be such that* $|U\mathbf{a}^*|^2 = \mathbf{y}$. *Then, every Picard iteration* $\mathbf{a}_{k+1} \in T(\mathbf{a}_k)$ *starting sufficiently close to* $\mathbf{a}^*$ *converges linearly to a point* $\tilde{\mathbf{a}} \in \operatorname{Fix} T$, *satisfying* $|U\tilde{\mathbf{a}}|^2 = \mathbf{y}$.

proof. Since $U = I_{n_a}$, the nuclear norm of $M(I_{n_a}, \mathbf{x}, -\mathbf{a}, \mathbf{y})$ can be calculated from the nuclear norms of n_a matrices $M(1, x_i, -a_i, y_i) \in \mathbb{C}^{2\times 2}$ $(1 \le i \le n_a)$. Let us do the calculation for an arbitrary $\mathbf{a} \in \mathbb{C}^{n_a}$. We first calculate the nuclear norm of each 2×2 matrix,

$$M(1, x_i, -a_i, y_i) = \begin{pmatrix} y_i - 2\mathcal{R}(x_i\overline{a_i}) + |a_i|^2 & x_i - a_i \\ \overline{x_i} - \overline{a_i} & 1 \end{pmatrix}.$$

Indeed, we have by direct calculation that

$$f_i(x_i) := \|M(1, x_i, -a_i, y_i)\|_*^2 = \left\| \begin{pmatrix} r & s \\ s & 1 \end{pmatrix} \right\|_*^2 = r^2 + 2s^2 + 1 + 2|r - s^2|, \tag{B1}$$

where

$$r := y_i - 2\mathcal{R}(x_i\overline{a_i}) + |a_i|^2, \qquad s := |x_i - a_i|.$$

Let us denote

$$T_i(a_i) := \arg\min_{x_i \in \mathbb{C}} f_i(x_i). \tag{B2}$$

Analytically solving the minimization problem on the right-hand side of Eq. (B2), we obtain the explicit form of T_i as follows:

$$T_i(a_i) = \begin{cases} \{z \in \mathbb{C} \,||z| \le \sqrt{y_i}\}, & \text{if } a_i = 0, \\ \left\{\frac{\sqrt{y_i}}{|a_i|} a_i\right\}, & \text{if } 0 < |a_i| \le \sqrt{\lambda_i}, \\ \left\{\frac{y_i + |a_i|^2 + 1}{2(|a_i|^2 + 1)} a_i\right\}, & \text{if } |a_i| \ge \sqrt{\lambda_i}, \end{cases} \tag{B3}$$

where λ_i is the unique real positive root of the real polynomial $g_i(t) := t^3 + 2(1 - y_i)t^2 + (y_i^2 - 6y_i + 1)t - 4y_i$.

We need to take care of the two possible cases of y_i.

Case 1. $y_i \in (0, 1]$. Then we have $\frac{3}{2}\sqrt{y_i} < \sqrt{\lambda_i} < 2\sqrt{y_i}$ since $g_i(\frac{9}{4}y_i) < 0$ and $g_i(4y_i) > 0$. The following properties of T_i can be verified.

- Fix $T_i = \{z \in \mathbb{C} \,||z| = \sqrt{y_i}\} \cup \{0\}$, where 0 is an inhomogeneous fixed point of T_i, that is, $T_i(0) \nsubseteq \operatorname{Fix} T_i$.
- The set of homogeneous fixed points of T_i is $S_i := \{z \in \mathbb{C} \,||z| = \sqrt{y_i}\}$.
- T_i is pointwise averaging at every point of S_i on $W_i := \{z \in \mathbb{C} \,||z| \ge \sqrt{y_i}/2\}$ with constant 3/4.
- The set-valued mapping $\psi_i := T_i - \mathrm{Id}$ is metrically subregular on W_i for 0 with constant 1/2.
- The technical assumption $\operatorname{dist}(z, S_i) \le \operatorname{dist}(z, \operatorname{Fix} T_i)$ holds for all $z \in W_i$.

Case 2. $y_i = 0$. Then $\lambda_i = 0$. Note also that $a_i^* = 0$ and the formula Eq. (B3) becomes $T_i(a_i) = \frac{1}{2}a_i$. The following properties of T_i can be verified.

- Fix $T_i = \{0\}$, where 0 is a homogeneous fixed point of T_i.
- T_i is pointwise averaging at every point of S_i on $\mathbb{C}$ with constant 1/4.
- The set-valued mapping $\psi_i := T_i - \mathrm{Id}$ is metrically subregular on $\mathbb{C}$ for 0 with constant 1/2.
- The technical assumption $\operatorname{dist}(z, S_i) \le \operatorname{dist}(z, \operatorname{Fix} T_i)$ holds for all $z \in \mathbb{C}$.

In this case, we denote $S_i := \{0\}$ and $W_i := \mathbb{C}$.

The operator T can be calculated explicitly,

$$T(\mathbf{a}) = \arg\min_{\mathbf{x} \in \mathbb{C}^{n_a}} \sum_{i=1}^{n_a} \sqrt{f_i(x_i)}, \quad \forall \mathbf{a} \in \mathbb{C}^{n_a}, \tag{B4}$$

where the constituent functions $f_i(x_i)$ are given by Eq. (B1).

Minimizing f_i $(i = 1, 2\ldots, n_a)$ separately yields the explicit form of T as a Cartesian product,

$$T(\mathbf{a}) = T_1(a_1) \times T_2(a_2) \cdots \times T_{n_a}(a_{n_a}), \tag{B5}$$

where the component operators T_i are given by Eq. (B3).

Thanks to the separability structure of T as a Cartesian product at Eq. (B5), the following properties of T in relation to Proposition 5.1 can be deduced from the corresponding ones of the component operators T_i.

- Fix $T = \prod_{i=1}^{n_a} \operatorname{Fix} T_i$ and the set of homogeneous fixed points of T is $S := \prod_{i=1}^{n_a} S_i$. It is clear that $|U\mathbf{a}|^2 = \mathbf{y}$ for $U = I_{n_a}$ and all $\mathbf{a} \in S$.
- T is pointwise averaging at every point of S on $W := \prod_{i=1}^{n_a} W_i$ with constant $\alpha = 3/4$.
- The set-valued mapping $\psi := T - \mathrm{Id}$ is *metrically subregular* on W for 0 with constant $\kappa = 1/2$.
- The technical assumption (iii) of Proposition 5.1 is satisfied on W. That is,

$$\operatorname{dist}(\mathbf{w}, S) \le \operatorname{dist}(\mathbf{w}, \operatorname{Fix} T), \quad \forall \mathbf{w} \in W. \tag{B6}$$

Now we can apply Proposition 5.1 to conclude that every Picard iteration $\mathbf{a}_{k+1} \in T(\mathbf{a}_k)$ starting in W converges linearly to a point in S as claimed. □

Remark B.1. Under the assumption that $y_i > 0$ for all $1 \le i \le n_a$, then the linear convergence result established in Lemma 4 can be sharpened to finite convergence.

In order to distinguish the fixed point operator Eq. (23) corresponding to a general unitary matrix U from the one analyzed in Lemma 4 corresponding to the identity matrix I_{n_a}, in the following proof we will use the notation $\widehat{T}$ for one specified in Theorem 5.2.

proof. Let T be the fixed point operator Eq. (23) that corresponds to the identity matrix and has been analyzed in Lemma 4. We start the proof by proving that

$$\widehat{T}(\mathbf{a}) = U^{-1}T(U\mathbf{a}), \quad \forall \mathbf{a} \in \mathbb{C}^{n_a}. \tag{B7}$$

Indeed, let us take an arbitrary $\mathbf{a} \in \mathbb{C}^{n_a}$ and denote $\mathbf{a}' = U\mathbf{a}$. Then we have

$$\begin{aligned} \widehat{T}(\mathbf{a}) &= \arg\min_{\mathbf{x} \in \mathbb{C}^{n_a}} \|M(U, \mathbf{x}, -\mathbf{a}, \mathbf{y})\|_* \\ &= \arg\min_{\mathbf{x} \in \mathbb{C}^{n_a}} \|M(I_{n_a}, U\mathbf{x}, -\mathbf{a}', \mathbf{y})\|_* \\ &= U^{-1}(\arg\min_{\mathbf{x} \in \mathbb{C}^{n_a}} \|M(I_{n_a}, \mathbf{x}, -\mathbf{a}', \mathbf{y})\|_*) \\ &= U^{-1}(T(\mathbf{a}')) = U^{-1}(T(U\mathbf{a})). \end{aligned} \tag{B8}$$

We have proved Eq. (B7). As a consequence,

$$\begin{aligned}\operatorname{Fix}\widehat{T} &= \{\mathbf{a}\in\mathbb{C}^{n_a}|\mathbf{a}\in\widehat{T}(\mathbf{a})\}\\ &= \{\mathbf{a}\in\mathbb{C}^{n_a}|\mathbf{a}\in U^{-1}T(U\mathbf{a})\}\\ &= \{\mathbf{a}\in\mathbb{C}^{n_a}|U\mathbf{a}\in T(U\mathbf{a})\}\\ &= \{\mathbf{a}\in\mathbb{C}^{n_a}|U\mathbf{a}\in\operatorname{Fix}T\} = U^{-1}(\operatorname{Fix}T).\end{aligned}\tag{B9}$$

For the sets S and W determined in the proof of Lemma 4, we denote $\widehat{S}\coloneqq U^{-1}(S)$ and $\widehat{W}\coloneqq U^{-1}(W)$. Since U is a unitary matrix, the set of homogeneous fixed points of $\widehat{T}$ is $\widehat{S}\coloneqq U^{-1}(S)$. It also holds by the definition of projection and Eq. (B9) that, for all $\mathbf{w}\in W$,

$$P_{U^{-1}(S)}(U^{-1}\mathbf{w}) = U^{-1}(P_S(\mathbf{w})),\tag{B10}$$

$$\operatorname{dist}(U^{-1}\mathbf{w},U^{-1}(S)) = \operatorname{dist}(U^{-1}\mathbf{w},U^{-1}(\operatorname{Fix}T)).\tag{B11}$$

By direct calculation one can verify the three assumptions on $\widehat{T}$ imposed in Proposition 5.1.

- $\widehat{T}$ is pointwise averaging at every point of $\widehat{S}$ on $\widehat{W}$ with constant $\alpha = 3/4$.
- The set-valued mapping $\widehat{\psi}\coloneqq\widehat{T}-\mathrm{Id}$ is metrically subregular on $\widehat{W}$ for 0 with constant $\gamma = 1/2$.
- The technical assumption (iii) of Proposition 5.1 is satisfied on $\widehat{W}$.

Therefore, we can apply Proposition 5.1 to conclude that every Picard iteration $\mathbf{a}_{k+1}\in\widehat{T}(\mathbf{a}_k)$ generated by the COPR algorithm starting in $\widehat{W}$ converges linearly to a point $\tilde{\mathbf{a}}\in\widehat{S}$. Finally, let $\tilde{\mathbf{w}}\in S$ such that $\tilde{\mathbf{a}} = U^{-1}\tilde{\mathbf{w}}$. It holds that $|U\tilde{\mathbf{a}}|^2 = |\tilde{\mathbf{w}}|^2 = \mathbf{y}$ by the structure of S.

The proof is complete. □

Funding. FP7 Ideas: European Research Council (IDEAS-ERC) (FP7/2007-2013, 339681).

REFERENCES

1. J. Antonello and M. Verhaegen, "Modal-based phase retrieval for adaptive optics," J. Opt. Soc. Am. A **32**, 1160–1170 (2015).
2. Y. Shechtman, Y. C. Eldar, O. Cohen, H. N. Chapman, J. Miao, and M. Segev, "Phase retrieval with application to optical imaging: a contemporary overview," IEEE Signal Process. Mag. **32**(3), 87–109 (2015).
3. J. R. Fienup, "Phase retrieval algorithms: a comparison," Appl. Opt. **21**, 2758–2769 (1982).
4. D. R. Luke, J. V. Burke, and R. G. Lyon, "Optical wavefront reconstruction: theory and numerical methods," SIAM Rev. **44**, 169–224 (2002).
5. Y. Shechtman, A. Beck, and Y. C. Eldar, "GESPAR: efficient phase retrieval of sparse signals," IEEE Trans. Signal Process. **62**, 928–938 (2014).
6. D. Sayre, "Some implications of a theorem due to Shannon," Acta Crystallogr. **5**, 843 (1952).
7. H. Hauptman, "The direct methods of X-ray crystallography," Science **233**, 178–183 (1986).
8. R. Gerchberg and W. Saxton, "A practical algorithm for the determination of phase from image and diffraction plane pictures," Optik **35**, 237–246 (1972).
9. H. Bauschke, P. Combetters, and D. Luke, "Phase retrieval, error reduction algorithm and Fienup variants: a view from convex optimization," J. Opt. Soc. Am. A **19**, 1334–1345 (2002).
10. E. J. Candes, X. Li, and M. Soltanolkotabi, "Phase retrieval via Wirtinger flow: theory and algorithms," IEEE Trans. Inf. Theory **61**, 1985–2007 (2015).
11. H. Ohlsson, A. Y. Yang, R. Dong, and S. S. Sastry, "Compressive phase retrieval from squared output measurements via semidefinite programming," arXiv: 1111.6323 (2011).
12. E. J. Candes, T. Strohmer, and V. Voroninski, "Phaselift: exact and stable signal recovery from magnitude measurements via convex programming," Commun. Pure Appl. Math. **66**, 1241–1274 (2013).
13. M. Fazel, H. Hindi, and S. P. Boyd, "A rank minimization heuristic with application to minimum order system approximation," in *Proceedings of the 2001 American Control Conference* (IEEE, 2001), Vol. **6**, pp. 4734–4739.
14. J. Goodman, *Introduction to Fourier Optics* (McGraw-Hill, 2008).
15. MOSEK ApS, The MOSEK Optimization Toolbox for MATLAB Manual. Version 7.1 (Revision 28) (2015).
16. L. Vandenberghe, V. R. Balakrishnan, R. Wallin, A. Hansson, and T. Roh, "Interior-point algorithms for semidefinite programming problems derived from the KYP lemma," in *Positive Polynomials in Control* (2005), p. 579.
17. A. Martinez-Finkelshtein, D. Ramos-Lopez, and D. Iskander, "Computation of 2D Fourier transforms and diffraction integrals using Gaussian radial basis functions," Appl. Comput. Harmon. Anal. **43**, 424–448 (2017).
18. P. J. Piscaer, A. Gupta, O. Soloviev, and M. Verhaegen, "Modal-based phase retrieval using Gaussian radial basis functions," J. Opt. Soc. Am. A **35**, 1233–1242 (2018).
19. A. J. Janssen, "Extended Nijboer-Zernike approach for the computation of optical point-spread functions," J. Opt. Soc. Am. A **19**, 849–857 (2002).
20. J. Braat, P. Dirksen, and A. J. Janssen, "Assessment of an extended Nijboer-Zernike approach for the computation of optical point-spread functions," J. Opt. Soc. Am. A **19**, 858–870 (2002).
21. R. Doelman and M. Verhaegen, "Sequential convex relaxation for convex optimization with bilinear matrix equalities," in *Proceedings of the European Control Conference* (2016).
22. B. Recht, M. Fazel, and P. A. Parrilo, "Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization," SIAM Rev. **52**, 471–501 (2010).
23. J. Löfberg, "YALMIP: a toolbox for modeling and optimization in MATLAB," in *Proceedings of the CACSD Conference*, Taipei, Taiwan (2004).
24. M. Grant and S. Boyd, "CVX: Matlab software for disciplined convex programming, version 2.1," 2014, http://cvxr.com/cvx.
25. S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," Found. Trends Mach. Learn. **3**, 1–122 (2011).
26. J.-F. Cai, E. J. Candès, and Z. Shen, "A singular value thresholding algorithm for matrix completion," SIAM J. Optim. **20**, 1956–1982 (2010).
27. H. H. Bauschke and P. L. Combettes, *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*, CMS Books Math./Ouvrages Math. SMC (Springer, 2011).
28. D. R. Luke, N. H. Thao, and M. K. Tam, "Quantitative convergence analysis of iterated expansive, set-valued mappings," arXiv: 1605.05725 (2016).
29. D. R. Luke, M. Teboulle, and N. H. Thao, "Necessary conditions for linear convergence of iterated expansive, set-valued mappings with application to alternating projections," arXiv: 1704.08926v2.
30. D. R. Luke, "Matlab proxtoolbox," 2018, http://num.math.uni-goettingen.de/proxtoolbox/.
31. R. W. Gerchberg and W. O. Saxton, "A practical algorithm for the determination of phase from image and diffraction plane pictures," Optik **35**, 237–246 (1972).
32. C.-C. Chen, J. Miao, C. Wang, and T. Lee, "Application of optimization technique to noncrystalline x-ray diffraction microscopy: guided hybrid input-output method," Phys. Rev. B **76**, 064113 (2007).
33. V. Elser, "Solution of the crystallographic phase problem by iterated projections," Acta Crystallogr. A **59**, 201–209 (2003).
34. F. Roddier, *Adaptive Optics in Astronomy* (Cambridge University, 1999).