Projection methods for high numerical aperture phase retrieval

Thao, Nguyen Hieu; Soloviev, Oleg; Luke, Russell; Verhaegen, Michel

DOI
10.1088/1361-6420/ac3322

Publication date
2021

Document Version
Final published version

Published in
Inverse Problems

Citation (APA)
Thao, N. H., Soloviev, O., Luke, R., & Verhaegen, M. (2021). Projection methods for high numerical aperture phase retrieval. *Inverse Problems*, *37*(12), Article 125005. https://doi.org/10.1088/1361-6420/ac3322

Projection methods for high numerical aperture phase retrieval

OPEN ACCESS
IOP Publishing
Inverse Problems
Inverse Problems **37** (2021) 125005 (27pp)
https://doi.org/10.1088/1361-6420/ac3322

Projection methods for high numerical aperture phase retrieval

Nguyen Hieu Thao[1,2,*], Oleg Soloviev[1,3], Russell Luke[4] and Michel Verhaegen[1]

[1] Delft Center for Systems and Control, Delft University of Technology, 2628CD Delft, The Netherlands
[2] Department of Mathematics, School of Education, Can Tho University, Can Tho, Vietnam
[3] Flexible Optical B.V., Polakweg 10-11, 2288 GG Rijswijk, The Netherlands
[4] Institut für Numerische und Angewandte Mathematik, Universität Göttingen, 37083 Göttingen, Germany

E-mail: h.t.nguyen-3@tudelft.nl

Received 22 July 2021, revised 3 October 2021
Accepted for publication 25 October 2021
Published 9 November 2021

Abstract

We develop for the first time a mathematical framework in which the class of projection algorithms can be applied to high numerical aperture (NA) phase retrieval. Within this framework, we first analyze the basic steps of solving the high-NA phase retrieval problem by projection algorithms and establish the closed forms of all the relevant projection operators. We then study the geometry of the high-NA phase retrieval problem and the obtained results are subsequently used to establish convergence criteria of projection algorithms in the presence of noise. Making use of the vectorial point-spread-function (PSF) is, on the one hand, the key difference between this paper and the literature of phase retrieval mathematics which deals with the scalar PSF. The results of this paper, on the other hand, can be viewed as extensions of those concerning projection methods for low-NA phase retrieval. Importantly, the improved performance of projection methods over the other classes of phase retrieval algorithms in the low-NA setting now also becomes applicable to the high-NA case. This is demonstrated by the accompanying numerical results which show that available solution approaches for high-NA phase retrieval are outperformed by projection methods.

*Author to whom any correspondence should be addressed.

Keywords: phase retrieval, high numerical aperture, projection algorithm, non-convex optimization, inconsistent feasibility

(Some figures may appear in colour only in the online journal)

1. Introduction

Phase retrieval is an important inverse problem in optics which aims at recovering a complex signal at the pupil plane of an optical system given a number of intensity measurements of its Fourier transform. It appears in many scientific and engineering fields, including microscopy [2, 30], astronomy imaging [12, 25], x-ray crystallography [26, 44], adaptive optics [1, 13, 14, 46], etc. For optical systems with low numerical aperture (NA), a vast number of phase retrieval algorithms have been devised, for example, in [5, 10, 15, 18–20, 27, 35, 39, 51, 55, 57] based on the *Fresnel approximation* stating that the intensity distribution in the focal plane and the complex signal in the pupil plane are related via the Fourier transform [22]. Among solution approaches for low-NA phase retrieval, the widely used class of projection algorithms, which can be viewed as descendants of the classical Gerchberg–Saxton algorithm [20], outperforms the other classes by almost every important performance measure: computational complexity, convergence speed, accuracy and robustness [39, p 410].

In the great development of industry 4.0, high-NA lenses have played a fundamental role in high resolution imaging technology, for example, in microscopy and extreme ultraviolet (EUV) lithography. Retrieving the aberration of such high-NA lenses is critical to the performance and maintenance of the high resolution imaging machines, for example, in monitoring the performance degradation of EUV projection systems in lithography technology due to thermal effects. This challenge can be cast in the framework of the *high-NA phase retrieval* problem that will be considered in this paper. For high-NA optical systems, the vector nature of light cannot be neglected and point-spread-functions (PSFs) are formed according to a more involved imaging formulation [34, 42, 43, 49], which is called the *vectorial PSF* to be distinguished from the *scalar* one according to the Fresnel diffraction equation. In contrast to low-NA settings, only few solution algorithms have been proposed for phase retrieval in high-NA settings [9, 23, 56].

Motivated by the rapidly growing application of high-NA optical systems and the scarcity of solution methods for high-NA phase retrieval, in this paper we develop for the first time a comprehensive mathematical framework in which the class of projection algorithms can be applied to the high-NA phase retrieval problem. Within this framework, we first analyze the basic steps of solving the high-NA phase retrieval problem by projection algorithms and establish the closed forms of all the relevant projectors. We then study the geometry of the high-NA phase retrieval problem and the obtained results are subsequently used to establish convergence criteria of projection algorithms in the presence of noise. Making use of the vectorial PSF is, on the one hand, the key difference between this paper and the literature of phase retrieval mathematics which mostly deals with the scalar PSF, see, for example, [5, 18, 20, 21, 36, 37, 52, 53, 55, 57]. The results of this paper, on the other hand, can be viewed as extensions of those concerning projection methods for low-NA phase retrieval. Importantly, the improved performance of projection methods over the other classes of phase retrieval algorithms in the low-NA setting [39, p 410] now also becomes applicable to the high-NA case. This is demonstrated by the accompanying numerical results which show that all available solution approaches for high-NA phase retrieval are outperformed by projection methods.

The remainder of this paper is organized as follows. Mathematical notation is introduced in section 1.1 and the vectorial PSF model is recalled in section 1.2. In section 2, several

feasibility models of the high-NA phase retrieval problem are formulated based on the vectorial PSF model (data fidelity) and the prior knowledge of the solutions. In section 3, closed forms of the projectors on the constituent sets of the feasibility models are established. In section 4, we discuss projection algorithms for solving the feasibility problems in section 2. Section 5 is devoted to studying the geometry of the high-NA phase retrieval problem where the constituent sets of feasibility are proven to be *prox-regular* at the points relevant for the subsequent convergence analysis. Section 6 is devoted to analyzing convergence of projection algorithms for solving the high-NA phase retrieval in the presence of noise. As the first ingredient of convergence, the *pointwise almost averagedness* property of projection algorithms is established in section 6.1 based on the prox-regularity of the component sets proven in section 5. The second condition of convergence concerning the mutual arrangement of the component sets around the solution [32, 33] is beyond the analysis of this paper. Convergence criteria are formulated in section 6.2. Numerical simulations are presented in section 7.

1.1. Mathematical notation

The underlying space in this paper is a finite dimensional Hilbert space denoted by $\mathcal{H}$. The Frobenius norm denoted by $\|\cdot\|$ is used for both vector and array objects. Equality, inequalities and mathematical operations such as the multiplication, the division, the square, the square root, the amplitude $|\cdot|$, the argument $\arg(\cdot)$ and the real part $\Re(\cdot)$ acting on arrays are understood element-wise. The imaginary unit is $\mathtt{j} = \sqrt{-1}$. The distance function associated to a set $\Omega \subset \mathcal{H}$ is defined by

$$\mathtt{dist}(\cdot, \Omega) : \mathcal{H} \to \mathbb{R}_+ : x \mapsto \inf_{w \in \Omega} \|x - w\|,$$

and the set-valued mapping

$$P_\Omega : \mathcal{H} \rightrightarrows \Omega : x \mapsto \{w \in \Omega \mid \|x - w\| = \mathtt{dist}(x, \Omega)\} \tag{1}$$

is the corresponding *projector*. A selection $w \in P_\Omega(x)$ is called a *projection* of x on Ω. When the projection w is unique, we write $P_\Omega(x) = w$ instead of $P_\Omega(x) = \{w\}$ for brevity. The *reflector* associated with Ω is accordingly defined by $R_\Omega \equiv 2P_\Omega - \mathtt{Id}$, where $\mathtt{Id}$ is the *identity mapping*. Since only projections on either affine or compact sets are involved in the analysis of this paper, the existence of projections is guaranteed. The fixed point set of a self set-valued mapping $T : \mathcal{H} \rightrightarrows \mathcal{H}$ is defined by $\mathtt{Fix}\ T \equiv \{x \in \mathcal{H} | x \in T(x)\}$, see, for example, [41, definition 2.1]. An iterative sequence $x_{k+1} \in T(x_k)$ generated by T is said to *converge linearly* to a point x with rate $c \in (0, 1)$ if there exists a number $\gamma > 0$ such that

$$\|x_k - x\| \leqslant \gamma c^k \quad \forall\ k \in \mathbb{N}.$$

For $x \in \mathcal{H}$, $\Omega \subset \mathcal{H}$ and an integer $m \geqslant 2$, we make use of the following notation

$$[x]_m \equiv \underbrace{(x, x, \ldots, x)}_{m \text{ times}} \quad \text{and} \quad [\Omega]_m \equiv \{[w]_m | w \in \Omega\}. \tag{2}$$

Our other basic notation is standard; cf [16, 45, 50]. The open ball with radius $\delta > 0$ and center x is denoted by $\mathbb{B}_\delta(x)$.

1.2. Vectorial point-spread-functions

This section presents the imaging formulation considered in the paper. For high-NA optical systems, PSFs should be modeled according to the *vector diffraction theory*, see, for example,

[24, 34, 42, 43, 49]. More specifically, the x, y, z components of the electromagnetic field right after the lens should be considered separately for the x and y components of the electromagnetic field just before the lens. Here we consider collimated beams and hence the z component of the field before the lens is zero. Let the unit electromagnetic fields in the x and y directions just before the lens respectively produce the fields right after the lens with components denoted by $(E_{\mathrm{XX}}(x,y), E_{\mathrm{XY}}(x,y), E_{\mathrm{XZ}}(x,y))$ and $(E_{\mathrm{YX}}(x,y), E_{\mathrm{YY}}(x,y), E_{\mathrm{YZ}}(x,y))$, where (x,y) are the coordinates in the lens aperture denoted by $\mathcal{P}$. Let the lens aperture $\mathcal{P}$ be normalized to have radius equal the NA value. Then according to, for example, [42, table 3.1], the latter functions are given by

$$\begin{aligned}
E_{\mathrm{XX}}(x,y) &= 1 - \frac{k_{\mathrm{X}}^2(x,y)}{1+k_{\mathrm{Z}}(x,y)}, \qquad E_{\mathrm{YX}}(x,y) = -\frac{k_{\mathrm{Y}}(x,y)k_{\mathrm{X}}(x,y)}{1+k_{\mathrm{Z}}(x,y)},\\
E_{\mathrm{XY}}(x,y) &= -\frac{k_{\mathrm{X}}(x,y)k_{\mathrm{Y}}(x,y)}{1+k_{\mathrm{Z}}(x,y)}, \qquad E_{\mathrm{YY}}(x,y) = 1 - \frac{k_{\mathrm{Y}}^2(x,y)}{1+k_{\mathrm{Z}}(x,y)},\\
E_{\mathrm{XZ}}(x,y) &= -k_{\mathrm{X}}(x,y), \qquad E_{\mathrm{YZ}}(x,y) = -k_{\mathrm{Y}}(x,y),
\end{aligned} \tag{3}$$

where $(k_{\mathrm{X}}(x,y), k_{\mathrm{Y}}(x,y), k_{\mathrm{Z}}(x,y))$ is the *unit wave vector* determined for each point (x,y) of the lens aperture $\mathcal{P}$ and satisfies

$$\max_{(x,y)\in\mathcal{P}} \left(k_{\mathrm{X}}^2(x,y) + k_{\mathrm{Y}}^2(x,y)\right) = \max_{(x,y)\in\mathcal{P}} \left(x^2+y^2\right) = \mathrm{NA}^2,$$

where the maximum is attainable on the boundary of $\mathcal{P}$ and NA is the NA value. In particular, the following equality will be used frequently in our subsequent analysis:

$$\sum_{c\in\mathcal{I}} E_c^2(x,y) = 2, \quad \forall\ (x,y)\in\mathcal{P}, \tag{4}$$

where and elsewhere in the paper, the letter c stands for elements of the index set:

$$\mathcal{I} \equiv \{\mathtt{XX}, \mathtt{XY}, \mathtt{XZ}, \mathtt{YX}, \mathtt{YY}, \mathtt{YZ}\}. \tag{5}$$

In the sequel, the coordinates (x,y) of two-dimensional arrays objects will be dropped for brevity, for example, we simply write E_c instead of $E_c(x,y)$.

Each of the right-hand side terms in (3) can be treated as a corresponding amplitude modulation in the entrance pupil for calculation of a PSF according to the Fresnel diffraction equation:

$$p_c(\mathcal{A},\Phi) = \left|\mathcal{F}\left(E_c\cdot\mathcal{A}\cdot \mathrm{e}^{\mathrm{j}\Phi}\right)\right|^2, \quad (\forall\ c\in\mathcal{I}), \tag{6}$$

where $\mathcal{A}$ and Φ are respectively the amplitude and phase of the collimated beam in the pupil plane, and $\mathcal{F}$ is the two-dimensional Fourier transform. The six constituent PSFs according to (6) then can be used to calculate the vectorial PSF corresponding to any linear polarization of light in the entrance pupil. For unspecified polarization state of light, they are added incoherently as follows:

$$I(\mathcal{A},\Phi) = \sum_{c\in\mathcal{I}} p_c(\mathcal{A},\Phi).$$

Thus, the vectorial PSF with an additional *phase diversity* ϕ_d is accordingly given by

$$I(\mathcal{A},\Phi,\phi_d) = \sum_{c\in\mathcal{I}} \left|\mathcal{F}\left(E_c\cdot\mathcal{A}\cdot \mathrm{e}^{\mathrm{j}(\Phi+\phi_d)}\right)\right|^2. \tag{7}$$

Remark 1.1 (Phase-diversity phase retrieval). [56, remark 3]. There are two widely used techniques of acquiring the PSF images for the phase-diversity phase retrieval problem considered in this paper. First, a *phase modulator* can be used to introduce the phase diversity patterns in the pupil plane corresponding to which the images are measured in the focal plane. Second, the images are registered in several planes parallel and at known distances to the focal plane. It is well known that the two techniques are mathematically equivalent [22]. However, in practice each approach to data acquisition has (dis)advantages compared to the other. For example, the first one requires additional optical instruments (e.g. deformable mirrors) and suffers approximation in generating the phase diversity patterns while the second one may suffer inaccuracies in shifting the image detector (e.g. CCD arrays) and differences in signal-to-noise ratio (SNR) of the acquired images as the distance between the pupil and the image planes varies. In many practical settings of phase retrieval, the second approach is more preferable than the first one.

The computational complexity of the vectorial PSF model (7) as a sum of six constituent components is approximately six times higher than the one of the scalar PSF. There is hence a trade off between computational complexity and model accuracy in choosing the imaging model of high-NA phase retrieval. Let us briefly analyze this matter. Figure 1 reports a short comparison between the *scalar* and *vectorial* PSF models for various NA values - 0.15, 0.55 and 0.95 in order from top to bottom. The left-hand side column of figure 1 shows PSFs without phase aberration and the second one shows those with phase aberration. All the PSFs are normalized to have total intensity one. In each plot, a pair of corresponding cross-sections of the scalar (the blue curves) and vectorial (the red curves) PSFs are shown. Note that only the central parts containing the main information of the cross-sections are plotted for the sake of clarity. It is clear that for low-NA values (0.15), the use of the vectorial PSF is superfluous as the two models are almost identical while the vectorial one is much more computationally expensive. The scalar and vectorial PSF models differ more for higher-NA imaging systems and for particular application purposes their discrepancy can become substantial for NA values from 0.55. It is worth noting that the larger the NA value, the more concentrated the PSF, and hence the larger the scale of the PSF.

2. Problem formulation

2.1. High-NA phase retrieval

This paper considers the same setting of high-NA phase retrieval as in [56]. For an unknown phase aberration $\Phi \in \mathbb{R}^{n\times n}$, let $r_d \in \mathbb{R}^{n\times n}_+$ be the measurement of m PSF images $I(\mathcal{A}, \Phi, \phi_d)$ generated by (7) with phase diversities ϕ_d $(d = 1, 2, \ldots, m)$. The *high-NA phase retrieval* problem is to restore Φ given r_d and ϕ_d $(d = 1, 2, \ldots, m)$ as well as the physical parameters of the optical system. Mathematically, we consider the problem of finding $\Phi \in \mathbb{R}^{n\times n}$ such that

$$r_d = \sum_{c\in\mathcal{I}} \left| \mathcal{F}\left(E_c \cdot \mathcal{A} \cdot \mathrm{e}^{\mathrm{j}(\Phi+\phi_d)}\right)\right|^2 + w_d \quad (d = 1, 2, \ldots, m), \tag{8}$$

where $\mathcal{A}$ is the possibly unknown amplitude of the generalized pupil function (GPF) and $w_d \in \mathbb{R}^{n\times n}$ $(d = 1, 2, \ldots, m)$ represent the discrepancies between the theoretically predicted data and the actually measured one, for example, due to noise and model deviations.

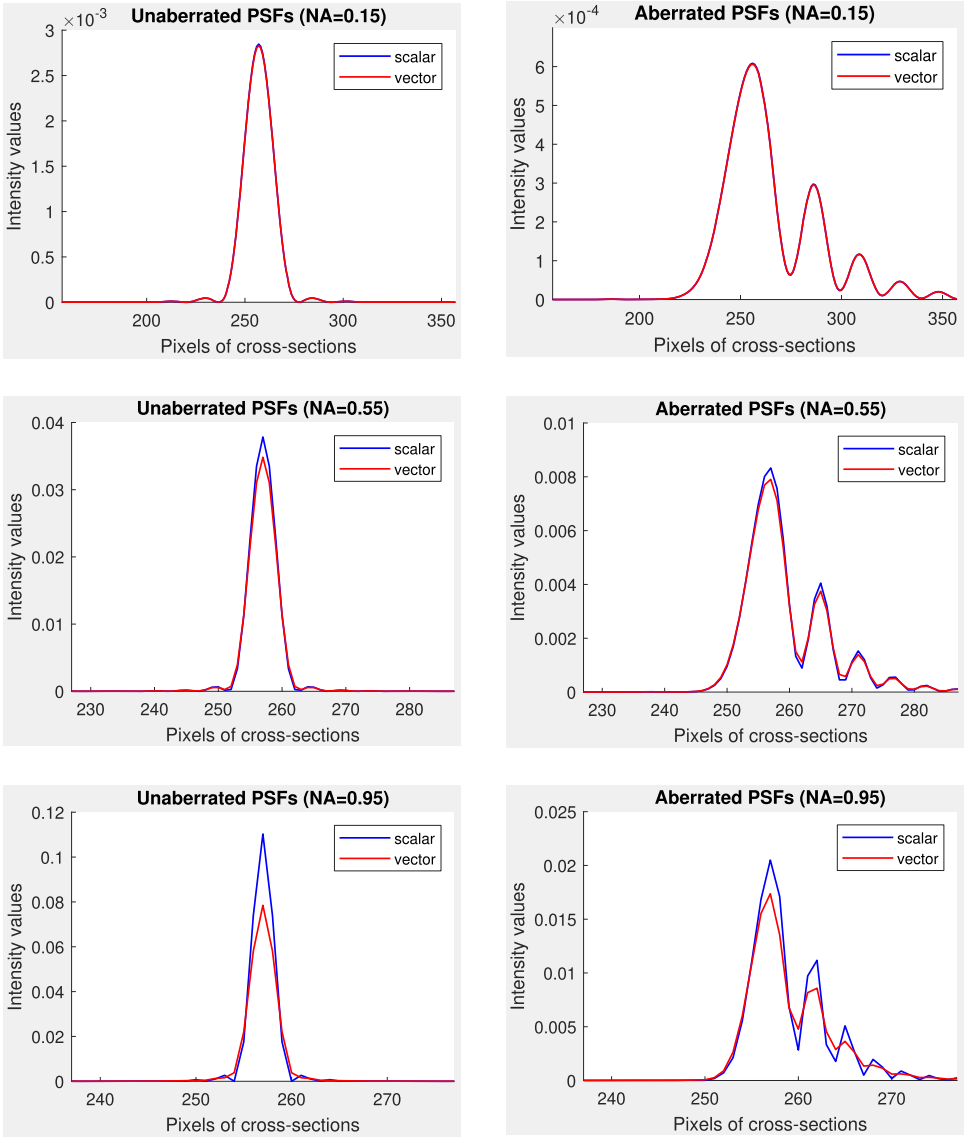

Figure 1. Comparison between the *scalar* and *vectorial* PSF models for various NA values - 0.15, 0.55 and 0.95 in order from top to bottom. In each plot, a pair of corresponding cross-sections of the scalar (in blue color) and vectorial (in red color) PSFs are shown. Only the central parts containing the main information of the cross-sections are plotted for the sake of clarity. The PSFs on the left-hand side are without phase aberration ($\Phi = 0$) and the ones on the right-hand side are with phase aberration ($\Phi \neq 0$). All the PSFs are normalized to have total intensity one. The two PSF models differ more for higher NA values and the discrepancy becomes substantial for NA from 0.55. Note that the larger the NA value, the more concentrated the PSF, and hence the larger the scale of the PSF.

2.2. Feasibility models

In this section, we formulate feasibility models of the phase retrieval problem (8) in two scenarios of application—known and unknown amplitude of the GPF. According to the vectorial PSF (7), we consider the underlying space

$$\mathcal{H} \equiv \underbrace{\mathbb{C}^{n\times n} \times \mathbb{C}^{n\times n} \times \cdots \times \mathbb{C}^{n\times n}}_{6 \text{ times}}.$$

In the sequel, for each $(x_{\mathtt{XX}}, x_{\mathtt{XY}}, x_{\mathtt{XZ}}, x_{\mathtt{YX}}, x_{\mathtt{YY}}, x_{\mathtt{YZ}}) \in \mathcal{H}$ and $z \in \mathbb{C}^{n\times n}$, we make use of the following notation in accordance with (5):

$$\begin{aligned}(x_c)_{c\in\mathcal{I}} &\equiv (x_{\mathtt{XX}}, x_{\mathtt{XY}}, x_{\mathtt{XZ}}, x_{\mathtt{YX}}, x_{\mathtt{YY}}, x_{\mathtt{YZ}}),\\ (x_c \cdot z)_{c\in\mathcal{I}} &\equiv (x_{\mathtt{XX}} \cdot z, x_{\mathtt{XY}} \cdot z, x_{\mathtt{XZ}} \cdot z, x_{\mathtt{YX}} \cdot z, x_{\mathtt{YY}} \cdot z, x_{\mathtt{YZ}} \cdot z).\end{aligned}$$

2.2.1. Unknown GPF amplitude. The following set captures the first constraint of a solution to (8) as an element of $\mathcal{H}$:

$$\Omega_0 \equiv \left\{(E_c \cdot x)_{c\in\mathcal{I}} \in \mathcal{H} \mid x \in \mathbb{C}^{n\times n}\right\}, \tag{9}$$

where the six matrices E_c ($c \in \mathcal{I}$) are defined in (3). Note that Ω_0 is linear subspace of $\mathcal{H}$ with $\dim(\Omega_0)$ being one sixth of $\dim(\mathcal{H})$. For $d = 1, 2, \ldots, m$, the intensity constraint set corresponding to phase diversity ϕ_d is given by

$$\Omega_d \equiv \left\{(x_c)_{c\in\mathcal{I}} \in \mathcal{H} \mid \sum_{c\in\mathcal{I}} \left|\mathcal{F}\left(x_c \cdot \mathrm{e}^{\mathrm{j}\phi_d}\right)\right|^2 = r_d\right\}. \tag{10}$$

The high-NA phase retrieval problem (8) then can be addressed via the following $(m+1)$-set feasibility:

$$\text{find} \quad x \in \bigcap_{d=0}^{m} \Omega_d. \tag{11}$$

The following two-set feasibility models formulated in the product spaces, which are equivalent to (11) in the case of consistent feasibility (i.e. the intersection is nonempty) [47], are widely used in practice:

$$\text{find} \quad u \in A \cap B \subset \mathcal{H}^m, \tag{12}$$

$$\text{find} \quad u \in D \cap B^+ \subset \mathcal{H}^{m+1}, \tag{13}$$

where

$$\begin{aligned} A &\equiv \left\{(x, x, \ldots, x) \in \mathcal{H}^m | x \in \Omega_0\right\}, & B &\equiv \Omega_1 \times \Omega_2 \times \cdots \times \Omega_m,\\ D &\equiv \left\{(x, x, \ldots, x) \in \mathcal{H}^{m+1} | x \in \mathcal{H}\right\}, & B^+ &\equiv \Omega_0 \times \Omega_1 \times \cdots \times \Omega_m. \end{aligned} \tag{14}$$

2.2.2. Known GPF amplitude. When the amplitude $\mathcal{A}$ of the GPF is known, it brings stronger constraint on the solutions of (8) than (9):

$$\chi \equiv \left\{ \left(E_c \cdot \mathcal{A} \cdot \mathrm{e}^{\mathrm{j}\Phi} \right)_{c\in\mathcal{I}} \in \mathcal{H} | \Phi \in \mathbb{R}^{n\times n} \right\}. \tag{15}$$

Similar to the case of unknown GPF amplitude, the phase retrieval problem (8) then can be addressed via one of the following feasibility models:

$$\text{find} \quad x \in \chi \cap \Omega_1 \cap \Omega_2 \cap \cdots \cap \Omega_m, \tag{16}$$

$$\text{find} \quad u \in A_\chi \cap B, \tag{17}$$

$$\text{find} \quad u \in D \cap B_\chi, \tag{18}$$

where

$$A_\chi \equiv \left\{ (x, x, \ldots, x) \in \mathcal{H}^m | x \in \chi \right\}, \qquad B_\chi \equiv \chi \times \Omega_1 \times \Omega_2 \times \cdots \times \Omega_m. \tag{19}$$

Remark 2.1 (Inconsistent feasibility). Due to noise and model deviations, the intersections in (11)–(13) and (16)–(18) are empty for all practical purposes. Keeping in mind, however, that projection algorithms as fixed point operators are not limited to finding points in the intersection of the sets. The convergence of their iterations to fixed points is desirable and sufficient in all scenarios of feasibility. Such fixed points should admit interpretation in terms of meaningful (approximate) solutions to the practical problem captured by the feasibility model. We refer the reader to [39, p 414] and [57, remark 5] for more details on inconsistency of feasibility formulations of (low-NA) phase retrieval.

Remark 2.2 (Effectiveness of the feasibility approach). It was observed in the recent benchmark paper [39, p 410] concerning low-NA phase retrieval that algorithms built on feasibility models outperform all other classes of solution methods by almost every important performance measure. This observation has strongly encouraged the current work which extends this class of algorithms for high-NA phase retrieval.

Remark 2.3 (Choice of feasibility models). Depending on the specific setting of phase retrieval, one feasibility model can result in better approximate solutions than others.

3. Calculation of projectors

The decisive step of solving feasibility problems is to calculate the projectors associated to the relevant sets. The results of this section, which can be viewed as the high-NA extensions of the ones concerning projection operators for low-NA phase retrieval [5, 18, 21, 36, 37, 52], enable us to address the feasibility models formulated in section 2.2 using projection algorithms.

For convenience let us first introduce further notation and preliminary results. For each $d = 1, 2, \ldots, m$ we define the operator $M_d : \mathcal{H} \to \mathcal{H}$ by

$$x = (x_c)_{c\in\mathcal{I}} \quad \mapsto \quad M_d(x) \equiv \left(\mathcal{F}\left(x_c \cdot \mathrm{e}^{\mathrm{j}\phi_d} \right) \right)_{c\in\mathcal{I}}, \tag{20}$$

which is a unitary transform and its inverse is given by

$$M_d^{-1} : x = (x_c)_{c\in\mathcal{I}} \quad \mapsto \quad \left(\mathcal{F}^{-1}(x_c) \cdot \mathrm{e}^{-\mathrm{j}\phi_d} \right)_{c\in\mathcal{I}}. \tag{21}$$

We then define the matrix-valued function $\mathcal{G}_d : \mathcal{H} \to \mathbb{R}_+^{n\times n}$ by

$$x = (x_c)_{c\in\mathcal{I}} \in \mathcal{H} \mapsto \mathcal{G}_d(x) \equiv \sqrt{\sum_{c\in\mathcal{I}} |M_d(x)_c|^2} \in \mathbb{R}_+^{n\times n}. \tag{22}$$

Fact 3.1 (Continuity of $\mathcal{G}_d$). The matrix-valued function $\mathcal{G}_d$ is continuous on $\mathcal{H}$.

Proof. Since compositions of continuous mappings are continuous, the statement follows from the continuity of M_d and the elementwise amplitude and summation operations. □

We define the set $S_d \subset \mathcal{H}$ by

$$S_d \equiv \left\{ x = (x_c)_{c\in\mathcal{I}} \in \mathcal{H} \mid \sum_{c\in\mathcal{I}} |x_c|^2 = r_d \right\}. \tag{23}$$

In the sequel, we also make use of the following set of indices:

$$\mathcal{J} \equiv \{\xi = (\xi_1, \xi_2) | 1 \leqslant \xi_1, \xi_2 \leqslant n\},$$

and for any $x = (x_c)_{c\in\mathcal{I}} \in \mathcal{H}$, we denote $x[\xi] \equiv (x_c[\xi])_{c\in\mathcal{I}}$ and $S_d[\xi] \equiv \{x[\xi] | x \in S_d\}$. In other words, the index of discretized two-dimensional signals (for example, x_c for each $c \in \mathcal{I}$) is specified by ξ in square brackets while the index of higher-dimensional arrays such as $x = (x_c)_{c\in\mathcal{I}} \in \mathcal{H}$ or S_d is defined inductively.

Fact 3.2 (Projection on S_d). It holds that

$$P_{S_d}(z) = \prod_{\xi\in\mathcal{J}} P_{S_d[\xi]}(z[\xi]) \quad (\forall\, z \in \mathcal{H}), \tag{24}$$

where

$$P_{S_d[\xi]}(z[\xi]) = \begin{cases} \dfrac{\sqrt{r_d[\xi]}}{\|z[\xi]\|} \cdot z[\xi] & \text{if } \|z[\xi]\| \neq 0, \\ \left\{ s \in \mathbb{C}^6 \mid \|s\|^2 = r_d[\xi] \right\} & \text{if } \|z[\xi]\| = 0. \end{cases} \tag{25}$$

Proof. The product structure (24) is inherent from the product structure of the set S_d, that is,

$$S_d = \prod_{\xi\in\mathcal{J}} S_d[\xi] \text{ and hence } P_{S_d}(z) = \prod_{\xi\in\mathcal{J}} P_{S_d[\xi]}(z[\xi]) \quad (\forall\, z \in \mathcal{H}).$$

Let us compute $P_{S_d[\xi]}(z[\xi])$ for each index $\xi \in \mathcal{J}$. Note that by its definition the set $S_d[\xi]$ is the sphere in $\mathbb{C}^6$ centered at the origin with radius $\sqrt{r_d[\xi]}$, that is,

$$S_d[\xi] = \left\{ s \in \mathbb{C}^6 \mid \|s\|^2 = r_d[\xi] \right\}. \tag{26}$$

Hence its associated projector $P_{S_d[\xi]}$ admits the closed form (25) as claimed. □

The next two results are widely known in the literature of feasibility analysis [47]. Recall the notation in (2).

Fact 3.3 (Projection on diagonals, P_D). For any $w = (w_0, w_1, \ldots, w_m) \in \mathcal{H}^{m+1}$ it holds that

$$P_D(w) = [\overline{w}]_{m+1} \quad \text{with } \overline{w} \equiv \frac{1}{m+1}\sum_{d=0}^{m} w_d.$$

Fact 3.4 (Projection on product sets, P_B). For any $w = (w_1, w_2, \ldots, w_m) \in \mathcal{H}^m$ it holds that

$$P_B(w) = \prod_{d=1}^{m} P_{\Omega_d}(w_d).$$

We can now calculate the projectors associated with the sets defined in section 2.2.

Lemma 3.5 (Projection on Ω_0). *For any $x = (x_c)_{c\in\mathcal{I}} \in \mathcal{H}$ it holds that*

$$P_{\Omega_0}(x) = (E_c \cdot z)_{c\in\mathcal{I}} \quad \textit{with } z \equiv \frac{1}{2}\sum_{c\in\mathcal{I}} (E_c \cdot x_c)\,.$$

Proof. By definition (9) of Ω_0 and the definition of projector in (1), $\left(E_c \cdot a \cdot \mathrm{e}^{\mathrm{j}\Psi}\right)_{c\in\mathcal{I}}$ is a projection of x on Ω_0 if and only if (a, Ψ) is a solution to the following minimization problem:

$$\min_{a\in\mathbb{R}_+^{n\times n},\Psi\in\mathbb{R}^{n\times n}} \left\|x - \left(E_c \cdot a \cdot \mathrm{e}^{\mathrm{j}\Psi}\right)_{c\in\mathcal{I}}\right\|^2. \tag{27}$$

The objective function of (27) can be rewritten as

$$\left\|x - \left(E_c \cdot a \cdot \mathrm{e}^{\mathrm{j}\Psi}\right)_{c\in\mathcal{I}}\right\|^2 = \|x\|^2 + \|(E_c \cdot a)_{c\in\mathcal{I}}\|^2 - 2\Re\left(\left(\sum_{c\in\mathcal{I}} (E_c \cdot a \cdot x_c)\right) \cdot \mathrm{e}^{-\mathrm{j}\Psi}\right).$$

The problem (27) is hence equivalent to the following one:

$$\min_{a\in\mathbb{R}_+^{n\times n},\Psi\in\mathbb{R}^{n\times n}} \|(E_c \cdot a)_{c\in\mathcal{I}}\|^2 - 2\Re\left(\left(\sum_{c\in\mathcal{I}} (E_c \cdot a \cdot x_c)\right) \cdot \mathrm{e}^{-\mathrm{j}\Psi}\right). \tag{28}$$

The structure of (28) allows us to solve for Ψ and a successively though its objective function is not completely separable in a and Ψ. Indeed, since a has no influence on the argument of $\sum_{c\in\mathcal{I}} (E_c \cdot a \cdot x_c)$, the set of optimal Ψ is given by

$$\left\{\Psi \in \mathbb{R}^{n\times n} \middle| \Psi \in \arg\left(\sum_{c\in\mathcal{I}} (E_c \cdot x_c)\right)\right\}. \tag{29}$$

Plugging the optimal Ψ above into (28), we arrive at minimizing a quadratic function of variable a. Taking into account that $\sum_{c\in\mathcal{I}} |E_c|^2 = 2J_n$ by (4) where J_n is the all-ones matrix of size $n \times n$, we obtain by direct calculation that the unique optimal a is given by

$$a = \frac{\left|\sum_{c\in\mathcal{I}} (E_c \cdot x_c)\right|}{\sum_{c\in\mathcal{I}} |E_c|^2} = \frac{1}{2}\left|\sum_{c\in\mathcal{I}} (E_c \cdot x_c)\right|.$$

Note that for any index $\xi \in \mathcal{J}$, if $a[\xi] = 0$, then $\Psi[\xi]$ does not play any role in the product $a[\xi]\mathrm{e}^{\mathrm{j}\,\Psi[\xi]}$. Otherwise, $\Psi[\xi]$ is uniquely determined in view of (29). Hence, the unique optimal solution to (27) is given by

$$z = a \cdot \mathrm{e}^{\mathrm{j}\,\Psi} = \frac{1}{2}\sum_{c\in\mathcal{I}} (E_c \cdot x_c)\,.$$

The proof is complete. □

Lemma 3.6 (Projection on Ω_d). *For each $d = 1, 2, \ldots, m$ and any $x \in \mathcal{H}$ it holds that*

$$P_{\Omega_d}(x) = M_d^{-1}(y),$$

where $y \in \mathcal{H}$ is characterized as follows.

(a) If $\mathcal{G}_d(x)[\xi] \neq 0$, then $y[\xi] = \frac{\sqrt{r_d[\xi]}}{\mathcal{G}_d(x)[\xi]} \cdot M_d(x)[\xi]$.
(b) If $\mathcal{G}_d(x)[\xi] = 0$, then $y[\xi]$ varies on the set $S_d[\xi]$ defined in (26).

Proof. By definitions (10), (20) and (23) of Ω_d, M_d and S_d respectively, it holds that

$$S_d = M_d(\Omega_d).$$

Then by the unitarity property of M_d, we have that

$$P_{\Omega_d}(x) = M_d^{-1}\left(P_{M_d(\Omega_d)}(M_d(x))\right) = M_d^{-1}\left(P_{S_d}(M_d(x))\right) \quad (\forall\ x \in \mathcal{H})\,.$$

Plugging the formulas of M_d, M_d^{-1} and P_{S_d} respectively given by (20), (21) and fact 3.2 into the above identity, we obtain the characterization of P_{Ω_d} as claimed. □

Lemma 3.7 (Projection on χ). *For any $x = (x_c)_{c\in\mathcal{I}} \in \mathcal{H}$, it holds that*

$$P_\chi(x) = \left\{ \left(E_c \cdot \mathcal{A} \cdot \mathrm{e}^{\mathrm{j}\Psi}\right)_{c\in\mathcal{I}} | \Psi \in \arg\left(\sum_{c\in\mathcal{I}} (E_c \cdot \mathcal{A} \cdot x_c)\right)\right\}. \tag{30}$$

Proof. By definition (15) of χ and the definition of projector in (1), $\left(E_c \cdot \mathcal{A} \cdot \mathrm{e}^{\mathrm{j}\,\Psi}\right)_{c\in\mathcal{I}}$ is a projection of x on χ if and only if Ψ is a solution to the following minimization problem:

$$\min_{\Psi\in\mathbb{R}^{n\times n}} \left\| x - \left(E_c \cdot \mathcal{A} \cdot \mathrm{e}^{\mathrm{j}\,\Psi}\right)_{c\in\mathcal{I}}\right\|^2. \tag{31}$$

The objective function of (31) can be rewritten as

$$\left\| x - \left(E_c \cdot \mathcal{A} \cdot \mathrm{e}^{\mathrm{j}\,\Psi}\right)_{c\in\mathcal{I}}\right\|^2 = -2\Re\left(\mathrm{e}^{-\mathrm{j}\,\Psi} \cdot \sum_{c\in\mathcal{I}} (E_c \cdot \mathcal{A} \cdot x_c)\right) + C,$$

where $C \equiv \|x\|^2 + \|(E_c \cdot \mathcal{A})_{c\in\mathcal{I}}\|^2$ is independent of Ψ. The problem (31) is hence equivalent to the following one:

$$\max_{\Psi\in\mathbb{R}^{n\times n}} \Re\left(\mathrm{e}^{-\mathrm{j}\,\Psi} \cdot \sum_{c\in\mathcal{I}} (E_c \cdot \mathcal{A} \cdot x_c)\right). \tag{32}$$

It is clear that the solution set of the problem (32) is given by

$$\left\{\Psi \in \mathbb{R}^{n\times n} | \Psi \in \arg\left(\sum_{c\in\mathcal{I}} (E_c \cdot \mathcal{A} \cdot x_c)\right)\right\}.$$

The proof is complete. □

Lemma 3.8 (Projection on *A*). *For any $w = (w_1, w_2, \ldots, w_m) \in \mathcal{H}^m$, it holds that*

$$P_A(w) = \left[(E_c \cdot \overline{z})_{c\in\mathcal{I}}\right]_m \quad \textit{with } \overline{z} \equiv \frac{1}{2}\sum_{c\in\mathcal{I}} (E_c \cdot \overline{w}_c),$$

where $\overline{w} = (\overline{w}_c)_{c\in\mathcal{I}} \equiv (1/m)\sum_{d=1}^m w_d$.

Proof. We first note that the set

$$L \equiv \{[x]_m \in \mathcal{H}^m | x \in \mathcal{H}\} \tag{33}$$

is a linear subspace of $\mathcal{H}^m$ and contains the set A. Then, for any $w \in \mathcal{H}^m$ and $a \in A$, it holds that $\langle w - P_L(w), a - P_L(w)\rangle = 0$, and thus

$$\|w - a\|^2 = \|P_L(w) - a\|^2 + \|P_L(w) - w\|^2.$$

Since $\|P_L(w) - w\| = \mathtt{dist}(w, L)$ is independent of the points of A, we have that

$$\arg\min_{a\in A} \|w - a\| = \arg\min_{a\in A} \|P_L(w) - a\| \quad \forall\ w \in \mathcal{H}^m.$$

In other words,

$$P_A(w) = P_A\left(P_L(w)\right) \quad \forall\ w = (w_1, w_2, \ldots, w_m) \in \mathcal{H}^m.$$

By fact 3.3, the projector P_L admits the following form:

$$P_L(w) = [\overline{w}]_m \quad \text{with } \overline{w} \equiv \frac{1}{m}\sum_{d=1}^m w_d. \tag{34}$$

This together with the definition of A in (14) yields that

$$P_A(w) = P_A\left([\overline{w}]_m\right) = \left[P_{\Omega_0}(\overline{w})\right]_m. \tag{35}$$

The claimed characterization of P_A then follows from (35) and lemma 3.5. □

Lemma 3.9 (Projection on A_χ). *For any $w = (w_1, w_2, \ldots, w_m) \in \mathcal{H}^m$, it holds that*

$$P_{A_\chi}(w) = \left\{\left[\left(E_c \cdot \mathcal{A} \cdot \mathrm{e}^{\mathrm{j}\Psi}\right)_{c\in\mathcal{I}}\right]_m | \Psi \in \arg\left(\sum_{c\in\mathcal{I}} (E_c \cdot \mathcal{A} \cdot \overline{w}_c)\right)\right\},$$

where $\overline{w} = (\overline{w}_c)_{c\in\mathcal{I}} \equiv (1/m)\sum_{d=1}^m w_d$.

Proof. The proof is similar to that of lemma 3.8. We first observe that the linear subspace L defined in (33) contains the set A_χ. As a consequence, it holds that

$$P_{A_\chi}(w) = P_{A_\chi}(P_L(w)) \quad \forall\, w \in \mathcal{H}^m.$$

In view of fact 3.3, the projector P_L admits the explicit form (34). This together with the definition of A_χ in (19) yields that

$$P_{A_\chi}(w) = P_{A_\chi}([\overline{w}]_m) = \left[P_\chi(\overline{w})\right]_m. \tag{36}$$

The claimed characterization of P_{A_χ} then follows from (36) and lemma 3.7. □

Remark 3.10 (Projections on B^+ and B_χ). The projectors P_{B^+} and P_{B_χ} are analogous to P_B in view of fact 3.4.

Remark 3.11 (Nonconvexity). Lemmas 3.6, 3.7 and 3.9 show that the projectors P_{Ω_d}, P_χ and P_{A_χ} are not single-valued in general. This particularly implies that the feasibility models of high-NA phase retrieval formulated in section 2.2 are nonconvex.

4. Projection algorithms

Projection methods for phase retrieval can be viewed as descendants of the famous Gerchberg–Saxton algorithm [20]. Its expansions to become the most widely used class of algorithms has been motivated by the rapidly widening scope of phase retrieval applications. Having calculated the relevant projectors in section 3, we can implement every projection algorithm for solving the feasibility models formulated in section 2.2. This section will briefly recall widely known projection methods for solving both two and more-set feasibility problems, typical examples of which are (12) and (11), respectively.

Widely known projection methods for solving two-set feasibility are recalled next.

(a) The alternating projection (AP) algorithm

$$T_{\mathrm{AP}}[A,B] \equiv P_A P_B.$$

(b) The Douglas–Rachford (DR) algorithm

$$T_{\mathrm{DR}}[A,B] \equiv \frac{1}{2}(R_A R_B + \mathtt{Id}) = P_A R_B - P_B + \mathtt{Id},$$

and its Krasnoselski–Mann relaxation (KM–DR algorithm)

$$T_{\mathrm{KM-DR}} \equiv \beta T_{\mathrm{DR}}[A,B] + (1-\beta)\mathtt{Id},$$

where $\beta \in (0,1]$ is the tuning parameter.

(c) The *hybrid projection-reflection* (HPR) algorithm [6, equation (19)]:

$$T_{\mathrm{HPR}} \equiv P_A((1+\beta)P_B - \mathtt{Id}) - \beta P_B + \mathtt{Id},$$

where $\beta \in (0,1]$ is the tuning parameter. As shown in [6, proposition 1], the HPR algorithm is equivalent to the Fienup's *hybrid input–output* method [18] when A is a linear subspace.

(d) The *relaxed-averaged-alternating-reflections* (RAAR) algorithm [35]:

$$T_{\mathtt{RAAR}}[A,B] \equiv \frac{\beta}{2}\left(R_A R_B + \mathtt{Id}\right) + (1-\beta)P_B$$
$$= \beta T_{\mathtt{DR}}[A,B] + (1-\beta)P_B,$$

where $\beta \in (0,1]$ is the tuning parameter.

(e) The *relaxed-reflect–reflect* (RRR) algorithm [17, algorithm 1]:

$$T_{\mathtt{RRR}}[A,B] \equiv \beta P_A\left(2P_B - \mathtt{Id}\right) - \beta P_B + \mathtt{Id},$$

where $\beta \in (0,1]$ is the tuning parameter.

(f) The DRAP algorithm [54]:

$$T_{\mathtt{DRAP}} \equiv P_A\left((1+\beta)P_B - \beta\mathtt{Id}\right) - \beta\left(P_B - \mathtt{Id}\right),$$

where $\beta \in [0,1]$ is the tuning parameter. This algorithm covers both $T_{\mathtt{DR}}$ (by setting $\beta = 1$) and $T_{\mathtt{AP}}$ (by setting $\beta = 0$). When A is affine, $T_{\mathtt{DRAP}}$ is convex combination of these two operators [57]. The latter also explains its name DRAP which stands for Doughlas–Rachford and AP.

Solution algorithms for solving the $(m+1)$-set feasibility are the cyclic projection and the cyclic versions of two-set feasibility based algorithms.

(g) The cyclic projection algorithm

$$T_{\mathtt{CP}}[\Omega_0,\Omega_1,\ldots,\Omega_m] \equiv P_{\Omega_0}P_{\Omega_1}\cdots P_{\Omega_m}.$$

(h) The *cyclic DR algorithm* proposed and analyzed in the context of convex feasibility [8]:

$$T_{\mathtt{CDR}}[\Omega_0,\Omega_1,\ldots,\Omega_m] \equiv T_{\mathtt{DR}}[\Omega_0,\Omega_1]T_{\mathtt{DR}}[\Omega_1,\Omega_2]\cdots T_{\mathtt{DR}}[\Omega_m,\Omega_0].$$

(i) The *cyclic RAAR algorithm* proposed in the context of low-NA phase retrieval [39]:

$$T_{\mathtt{CRAAR}}[\Omega_0,\Omega_1,\ldots,\Omega_m] \equiv T_{\mathtt{RAAR}}[\Omega_0,\Omega_1]T_{\mathtt{RAAR}}[\Omega_1,\Omega_2]\cdots T_{\mathtt{RAAR}}[\Omega_m,\Omega_0].$$

(j) The *cyclically anchored Douglas–Rachford algorithm* proposed in the context of convex set feasibility [7]:

$$T_{\mathtt{CADR}}[\Omega_0,\Omega_1,\ldots,\Omega_m] \equiv T_{\mathtt{DR}}[\Omega_0,\Omega_1]T_{\mathtt{DR}}[\Omega_0,\Omega_2]\cdots T_{\mathtt{DR}}[\Omega_0,\Omega_m].$$

The cyclic versions of $T_{\mathtt{KM-DR}}$, $T_{\mathtt{RRR}}$ and $T_{\mathtt{DRAP}}$ as well as the cyclically anchored versions of $T_{\mathtt{KM-DR}}$, $T_{\mathtt{RAAR}}$, $T_{\mathtt{RRR}}$ and $T_{\mathtt{DRAP}}$ can also be designed similarly. Note that all the sets Ω_d $(d = 0,1,\ldots,m)$ are treated equally in the cyclic-type algorithms while the set Ω_0 distinctively plays the role of an *anchor* for the others in the cyclically-anchored-type algorithms. The latter can be supported by the fact that Ω_0 represents a mathematical constraint while the other sets capture the physical measurements. In fact, the distinctive role of Ω_0 has been taken into account in (12), the problem formulation we mainly work with in sections 6 and 7.

Remark 4.1 (Multi-valuedness of projection algorithms). Since the projectors presented in section 3 are potentially multi-valued, the above algorithms built on them are in general not single-valued.

Remark 4.2 (choice of algorithms). Depending on the specific setting of phase retrieval, one algorithm can result in better approximate solutions than others, see also remark 2.3. It is worth mentioning that AP is eventually needed for suppressing noise and model deviations regardless of the chosen algorithm.

5. Geometry of high-NA phase retrieval

In this section we analyze the geometry of the high-NA phase retrieval problem. The sets constituting the feasibility models in section 2 will be shown to be *prox-regular* at the points relevant to our subsequent convergence analysis in section 6. We mention that the prox-regularity property in the context of phase retrieval was first analyzed by Luke [36, section 3.1].

Definition 5.1 (Prox-regularity) [48]. A set Ω is *prox-regular at a point* $\hat{x} \in \Omega$ if the associated projector P_Ω is single-valued around $\hat{x}$. Ω is *prox-regular* if it is prox-regular at every of its points.

Example 5.2 (Prox-regularity of Ω_0, A and D). Any closed and convex set is prox-regular [50]. In particular, the linear subspaces Ω_0, A and D defined in (9) and (14) are prox-regular.

The next two assertions follow from the definition of prox-regularity. Recall the notation $[\cdot]_p$ in (2).

Fact 5.3. Let $\Omega \subset \mathcal{H}$ be prox-regular at a point $\hat{x} \in \Omega$ and $p \geqslant 2$ be an integer. Then the set $\mathbf{\Omega} \equiv \{[x]_p \in \mathcal{H}^p | x \in \Omega\}$ is prox-regular at $[\hat{x}]_p$.

Proof. By definition 5.1, there is a neighborhood U of $\hat{x}$ on which P_Ω is single-valued. Let us define the set $\mathcal{U} \subset \mathcal{H}^p$ by

$$\mathcal{U} \equiv \left\{ [\hat{x}]_p + (r_1, r_2, \ldots, r_p) \in \mathcal{H}^p \mid \hat{x} + \frac{1}{p}\sum_{k=1}^{p} r_k \in U \right\}. \tag{37}$$

Note that $\mathcal{U}$ is a neighborhood of $[\hat{x}]_p$ since U is a neighborhood of $\hat{x}$. It suffices to check that $P_{\mathbf{\Omega}}$ is single-valued on $\mathcal{U}$. Indeed, take an arbitrary point

$$(x_1, x_2, \ldots, x_p) = [\hat{x}]_p + (r_1, r_2, \ldots, r_p) \in \mathcal{U}.$$

Then in view of (37), it holds that

$$\bar{x} \equiv \frac{1}{p}\sum_{k=1}^{p} x_k = \hat{x} + \frac{1}{p}\sum_{k=1}^{p} r_k \in U,$$

and hence $P_\Omega(\bar{x})$ is singleton since P_Ω is single-valued on U. Using the reasoning in the proof of lemma 3.8, we have $P_{\mathbf{\Omega}}(x_1, x_2, \ldots, x_p) = \left\{[P_\Omega(\bar{x})]_p\right\}$ which is singleton. Hence $P_{\mathbf{\Omega}}$ is singled-valued on $\mathcal{U}$ and the proof is complete. □

Fact 5.4 (Prox-regularity of products). For each $k = 1, 2, \ldots, p$ let Ω_k be prox-regular at $\hat{x}_k$. Then the product set $\prod_{k=1}^{p} \Omega_k$ is prox-regular at $(\hat{x}_1, \hat{x}_2, \ldots, \hat{x}_p)$.

Proof. The proof follows from the definition of prox-regularity and the separation property of projection on product sets. □

We can now analyze the prox-regularity of the other sets defined in section 2.2.

Proposition 5.5 (Prox-regularity of Ω_d). *For each $d = 1, 2, \ldots, m$ the set Ω_d defined in (10) is prox-regular at every point $\hat{x} \in \Omega_d$ with $\mathcal{G}_d(\hat{x})$ nonzero everywhere.*

Proof. Consider a point $\hat{x} \in \Omega_d$ with $\mathcal{G}_d(\hat{x})$ nonzero everywhere. By definition 5.1, it suffices to find a neighborhood of $\hat{x}$ on which P_{Ω_d} is single-valued. Let us define the set U_d by

$$U_d \equiv \{\hat{x} + r | r \in \mathcal{H},\ \mathcal{G}_d(r) < \mathcal{G}_d(\hat{x})\}\,. \tag{38}$$

Since $\mathcal{G}_d$ is continuous by fact 3.1, it holds that

$$\mathcal{G}_d(r) \to 0 \text{ in } \mathbb{R}^{n\times n}_{+} \quad \text{as } r \to 0 \text{ in } \mathcal{H}.$$

This together with $\mathcal{G}_d(\hat{x})$ being nonzero everywhere implies that U_d is a neighborhood of $\hat{x}$. We will show that P_{Ω_d} is single-valued on U_d. Indeed, let us take an arbitrary point $x = \hat{x} + r \in U_d$ and first check that $\mathcal{G}_d(x) \neq 0$ for all entries. Using (22), the linearity of M_d and the triangle inequality successively, we have that

$$\begin{aligned}
\mathcal{G}_d(x) = \mathcal{G}_d(\hat{x} + r) &= \sqrt{\sum_{c\in\mathcal{I}} |M_d(\hat{x} + r)_c|^2} \\
&= \sqrt{\sum_{c\in\mathcal{I}} |M_d(\hat{x})_c + M_d(r)_c|^2} \\
&\geqslant \sqrt{\sum_{c\in\mathcal{I}} \big(|M_d(\hat{x})_c| - |M_d(r)_c|\big)^2}.
\end{aligned} \tag{39}$$

Suppose on the contrary that $\mathcal{G}_d(x)[\xi] = 0$ for some index $\xi \in \mathcal{J}$. Then (39) implies that

$$|M_d(\hat{x})_c|\,[\xi] = |M_d(r)_c|\,[\xi] \quad \forall\ c \in \mathcal{I}.$$

This in particular yields $\mathcal{G}_d(\hat{x})[\xi] = \mathcal{G}_d(r)[\xi]$ which is a contradiction to (38) as $\hat{x} + r \in U_d$. Hence we have $\mathcal{G}_d(x) \neq 0$ for all entries as claimed. Now by lemma 3.6, $P_{\Omega_d}(x)$ is the singleton $\{M_d^{-1}(y)\}$, where y is uniquely determined. The proof is complete. □

Proposition 5.6 (Prox-regularity of χ). *Suppose that the amplitude $\mathcal{A}$ is nonzero everywhere. Then the set χ defined in (15) is prox-regular.*

Proof. Let us consider an arbitrary point $\hat{x} = (\hat{x}_c)_{c\in\mathcal{I}} \in \chi$. By definition 5.1, it suffices to find a neighborhood of $\hat{x}$ on which P_χ is single-valued. Let us define the set U_χ by

$$U_\chi \equiv \left\{ \hat{x} + (r_c)_{c\in\mathcal{I}} \in \mathcal{H} \mid \sqrt{\sum_{c\in\mathcal{I}} |r_c|^2} < \sqrt{2}\,\mathcal{A} \right\}. \tag{40}$$

Since $\mathcal{A}$ is nonzero everywhere, the set U_χ defined in (40) is a neighborhood of $\hat{x}$. We will show that P_χ is single-valued on U_χ. Take an arbitrary point $x = (x_c)_{c\in\mathcal{I}} = \hat{x} + (r_c)_{c\in\mathcal{I}} \in U_\chi$. Then using the triangle inequality, the Cauchy–Schwarz inequality, (15), (4) and (40) successively, we get that

$$\begin{aligned}
\left|\sum_{c\in\mathcal{I}} (E_c \cdot \mathcal{A} \cdot x_c)\right| &= \left|\sum_{c\in\mathcal{I}} (E_c \cdot \mathcal{A} \cdot (\hat{x}_c + r_c))\right| \\
&\geqslant \left|\sum_{c\in\mathcal{I}} (E_c \cdot \mathcal{A} \cdot \hat{x}_c)\right| - \left|\sum_{c\in\mathcal{I}} (E_c \cdot \mathcal{A} \cdot r_c)\right| \\
&\geqslant \left|\sum_{c\in\mathcal{I}} (E_c \cdot \mathcal{A} \cdot \hat{x}_c)\right| - \sqrt{\sum_{c\in\mathcal{I}} (E_c \cdot \mathcal{A})^2} \cdot \sqrt{\sum_{c\in\mathcal{I}} |r_c|^2} \\
&= \sum_{c\in\mathcal{I}} (E_c \cdot \mathcal{A})^2 - \sqrt{\sum_{c\in\mathcal{I}} (E_c \cdot \mathcal{A})^2} \cdot \sqrt{\sum_{c\in\mathcal{I}} |r_c|^2} \\
&= 2\mathcal{A}^2 - \sqrt{2}\,\mathcal{A} \cdot \sqrt{\sum_{c\in\mathcal{I}} |r_c|^2} > 0.
\end{aligned}$$

This implies that $\sum_{c\in\mathcal{I}} (E_c \cdot \mathcal{A} \cdot x_c)$ is nonzero everywhere. Hence, by lemma 3.7, $P_\chi(x)$ is the singleton $\left\{\left(E_c \cdot \mathcal{A} \cdot \mathrm{e}^{\mathrm{j}\Psi}\right)_{c\in\mathcal{I}}\right\}$, where $\Psi \in \mathbb{R}^{n\times n}$ is uniquely given by (30). The proof is complete. □

Proposition 5.7 (Prox-regularity of A_χ). *Suppose that the amplitude $\mathcal{A}$ is nonzero everywhere. Then the set A_χ defined in (19) is prox-regular at every point $[\hat{x}]_m$ with $\hat{x} \in \chi$.*

Proof. The proof follows from proposition 5.6 and fact 5.3. □

Proposition 5.8 (Prox-regularity of B, B^+ and B_χ). *The following statements hold true.*

(*a*) *The set B defined in (14) is prox-regular at every point $(\hat{x}_1, \hat{x}_2, \ldots, \hat{x}_m) \in B$ with $\mathcal{G}_d(\hat{x}_d)$ nonzero everywhere ($\forall\, d = 1, 2, \ldots, m$).*

(*b*) *The set B^+ defined in (14) is prox-regular at every point $(\hat{x}, \hat{x}_1, \hat{x}_2, \ldots, \hat{x}_m) \in B^+$ with $\mathcal{G}_d(\hat{x}_d)$ nonzero everywhere ($\forall\, d = 1, 2, \ldots, m$).*

(*c*) *Suppose that the amplitude $\mathcal{A}$ is nonzero everywhere. Then the set B_χ defined in (19) is prox-regular at every point $(\hat{x}, \hat{x}_1, \hat{x}_2, \ldots, \hat{x}_m) \in B_\chi$ with $\mathcal{G}_d(\hat{x}_d)$ nonzero everywhere ($\forall\, d = 1, 2, \ldots, m$).*

Proof. (a) By proposition 5.5, for each $d = 1, 2, \ldots, m$ there exists a neighborhood U_d of $\hat{x}_d$ on which P_{Ω_d} is single-valued. This combined with fact 3.4 yields that $P_B = \prod_{d=1}^{m} P_{\Omega_d}$ is single-valued in the neighborhood $\prod_{d=1}^{m} U_d$ of $(\hat{x}_1, \hat{x}_2, \ldots, \hat{x}_m)$. This yields the prox-regularity of B at this point as claimed. (b) This part is also encompassed by part (a) since Ω_0 is prox-regular in view of example 5.2. (c) Thanks to proposition 5.6, there exists a neighborhood U_χ of $\hat{x}$ on which P_χ is single-valued. By proposition 5.5, for each $d = 1, 2, \ldots, m$ there exists a neighborhood U_d of $\hat{x}_d$ on which P_{Ω_d} is single-valued. We thus have in view of remark 3.10 that $P_{B_\chi} = P_\chi \times \prod_{d=1}^{m} P_{\Omega_d}$ is single-valued on the neighborhood $U_\chi \times \prod_{d=1}^{m} U_d$ of $(\hat{x}, \hat{x}_1, \hat{x}_2, \ldots, \hat{x}_m)$. This yields the prox-regularity of B_χ at this point as claimed. The proof is complete. □

Remark 5.9. The condition that $\mathcal{A}$ are nonzero everywhere imposed in propositions 5.6, 5.7 and 5.8(c) physically means that the entire aperture of the imaging system is illuminated.

6. Convergence analysis

Since the feasibility models of high-NA phase retrieval formulated in section 2.2 are nonconvex (remark 3.11), the projectors upon which our algorithms are based are not nonexpansive (Lipschitz continuous with constant 1). As a result, the tools of convex analysis and monotone/averaged operator theory (for example, [3, 4]) are not directly applicable to the problem under consideration. Recently, however, a framework has been established that accommodates fixed point iterations built from compositions and averages of set-valued, *expansive* mappings [41]. The extended analysis scheme is based on the theory of *pointwise almost averaged mappings* and has been applied to prove, for the first time, local (linear) convergence of fundamental algorithms like cyclic projections and relaxed DR for solving inconsistent, nonconvex feasibility problems; see, for example, [15, 38, 54, 55, 57]. It is worth mentioning that (low-NA) phase retrieval has been a main motivation of the mathematical development in [41]. In this paper, we also follow the analysis scheme of the aforementioned paper, according to which convergence of the iterative sequences generated by a fixed point operator T is guaranteed by the *pointwise almost averagedness* of T and the *metric subregularity* of the mapping $\mathtt{Id} - T$ on the relevant regions. The contribution of this section concerns the first condition of convergence. The almost averagedness property of projection algorithms will be derived from the geometry of the high-NA phase retrieval problem analyzed in section 5.

Although being derived from the general scheme of [41], convergence analysis is different for each projection method, depending on its fixed point set and its complexity, especially for solving nonconvex and inconsistent feasibility problems. In this section, we analyze the AP algorithm for high-NA phase retrieval in the presence of noise. We consider the two-set feasibility model (12) in the *inconsistent* setting, i.e. the sets do not intersect. It is worth emphasizing that the class of projection methods for high-NA phase retrieval is first considered in this paper, and thus the obtained results are new from the application point of view, even in the consistent case.

6.1. Pointwise almost averagedness

The following property is taken from definition 2.2 and proposition 2.1 of [41].

Definition 6.1 (Pointwise almost averaged mappings). A fixed point mapping $T : \mathcal{H} \rightrightarrows \mathcal{H}$ is *pointwise almost averaged* at a point y on a set $\Omega \subset \mathcal{H}$ with violation $\varepsilon \geqslant 0$ and averaging constant $\alpha \in (0, 1)$ if for all $y^+ \in T(y)$, $z \in \Omega$ and $z^+ \in T(z)$, it holds that

$$\left\|z^+ - y^+\right\|^2 \leqslant (1+\varepsilon)\left\|z - y\right\|^2 - \frac{1-\alpha}{\alpha}\left\|(z^+ - z) - (y^+ - y)\right\|^2.$$

When the violation $\varepsilon = 0$, the quantifiers 'almost' and 'violation' in definition 6.1 are dropped and the property goes back to the conventional averagedness property, see, for example, [4]. When the property holds for every point $y \in \Omega$ with the same violation and averaging constant, the quantifiers 'pointwise' and 'at a point' in definition 6.1 are dropped. The property is well defined for any averaging constant $\alpha > 0$, not necessarily limited to $\alpha \in (0, 1)$ though the latter is often of the main interest.

Example 6.2 (Projection on convex sets). The projectors associated with closed and convex sets are globally averaged with averaging constant $\alpha = 1/2$ (i.e. firmly nonexpansive), see, for example, [11, theorem 2.2.21].

The following statement is a consequence of widely known results concerning projections on nonconvex sets, see, for example, [28, theorem 2.14].

Proposition 6.3 (Projection on prox-regular sets). *Let Ω be closed and prox-regular at $\hat{x} \in \Omega$. Then given an arbitrarily small number $\varepsilon > 0$, there exists a neighborhood of $\hat{x}$ (depending on ε) on which P_Ω is almost averaged with violation ε and averaging constant $\alpha = 1/2$.*

The next property of pointwise almost averaged mappings is needed [41, proposition 2.4(ii)]. The version specialized to the problem (12) is presented here for brevity.

Proposition 6.4 (Pointwise almost averagedness of composite mappings). *Let $T_k : \mathcal{H} \rightrightarrows \mathcal{H}$ for $k = 1, 2$ be pointwise almost averaged on U_k at all $y_k \in S_k$ with violation $\varepsilon_k \geqslant 0$ and averaging constant $\alpha_k \in (0, 1)$. If $T_2(U_2) \subseteq U_1$ and $T_2(S_2) \subseteq S_1$, then the composite mapping $T \equiv T_1 \circ T_2$ is pointwise almost averaged on U_2 at all $y \in S_2$ with violation ε and averaging constant α given by*

$$\varepsilon = \varepsilon_1 + \varepsilon_2 + \varepsilon_1\varepsilon_2; \qquad \alpha = \frac{2\max\{\alpha_1, \alpha_2\}}{1 + \max\{\alpha_1, \alpha_2\}}.$$

The next result links the prox-regularity of the sets in (12) with the almost averagedness of the AP operator.

Proposition 6.5 (Almost averagedness of $T_{\mathtt{AP}}$). *Let $\widehat{\mathbf{b}} \equiv (\hat{x}_1, \hat{x}_2, \ldots, \hat{x}_m) \in B$, where $\hat{x}_d \in \Omega_d$ with $\mathcal{G}_d(\hat{x}_d)$ nonzero everywhere ($\forall\, d = 1, 2, \ldots, m$). Then given any number $\varepsilon > 0$, there is a neighborhood of $\widehat{\mathbf{b}}$, denoted by $U_\varepsilon(\widehat{\mathbf{b}})$, on which the AP operator $T_{\mathtt{AP}} \equiv P_A P_B$ associated with (12) is almost averaged with violation ε and averaging constant $\alpha = 2/3$.*

Proof. By proposition 5.5, the sets Ω_d are prox-regular at $\hat{x}_d$ as $\mathcal{G}_d(\hat{x}_d)$ are nonzero everywhere ($\forall\, d = 1, 2, \ldots, m$). Thanks to fact 5.4, the set B is prox-regular at $\widehat{\mathbf{b}}$. Then by proposition 6.3, there exists a neighborhood $U_\varepsilon(\widehat{\mathbf{b}})$ of $\widehat{\mathbf{b}}$ on which the projector P_B is almost averaged with violation ε and averaging constant $1/2$. On the other hand, since A is convex in view of example 5.2, the projector P_A is globally averaged with averaging constant $1/2$ (i.e. firmly nonexpansive) in view of example 6.2. Thus by proposition 6.4, the composite mapping $T_{\mathtt{AP}} \equiv P_A P_B$ is almost averaged on $U_\varepsilon(\widehat{\mathbf{b}})$ with violation ε and averaging constant $\alpha = 2/3$ as claimed. □

6.2. Convergence statements

The goal of this section is to combine the results of section 6.1 with the analysis scheme of [41, section 2.2] to obtain convergence criteria for the AP algorithm for solving (12) in the inconsistent setting. The following notion of metric subregularity is a cornerstone of variational analysis and optimization theory with many important applications, such as in establishing calculus rules for subdifferentials and coderivatives [29, 45, 50] and in analyzing stability and convergence of numerical algorithms, see, for example, [16, 31].

Definition 6.6 (Metric subregularity on a set). A set-valued mapping $\Theta : \mathcal{H} \rightrightarrows \mathcal{H}$ is *metrically subregular* on $U \subset \mathcal{H}$ for $\hat{y} \in \mathcal{H}$ relative to $\Lambda \subset \mathcal{H}$ with *modulus* $\kappa > 0$ if

$$\kappa\, \mathtt{dist}(x, \Theta^{-1}(\hat{y}) \cap \Lambda) \leqslant \mathtt{dist}(\hat{y}, \Theta(x)), \quad \forall\, x \in U \cap \Lambda.$$

When U is some neighborhood of a point $\hat{x} \in \Theta^{-1}(\hat{y})$, the property is called *metric subregularity of* Θ *at* $\hat{x}$ *for* $\hat{y}$ *relative to* Λ.

The next lemma is a specification of [41, corollary 2.3] to our target application.

Lemma 6.7 (Linear convergence with metric subregularity). *Let* $T : \mathcal{H} \rightrightarrows \mathcal{H}$ *be a fixed point operator with* $\mathtt{Fix}\ T$ *closed,* $\Lambda \subset \mathcal{H}$ *with* $T(\Lambda) \subset \Lambda$, $\hat{x} \in \Lambda \cap \mathtt{Fix}\ T$ *and* U *a neighborhood of* $\hat{x}$ *with* $T(U) \subset U$. *Suppose that*

(*a*) T *is pointwise almost averaged at* $\hat{x}$ *on* $\Lambda \cap U$ *with violation* $\varepsilon \geqslant 0$ *and averaging constant* $\alpha \in (0, 1)$;

(*b*) *the mapping* $\mathtt{Id} - T$ *is metrically subregular on* U *for 0 relative to* Λ *with modulus* $\kappa > \sqrt{\varepsilon\alpha/(1-\alpha)}$.

Then every iterative sequence generated by T *with the initial point in* $\Lambda \cap U$ *converges linearly to a point in* $\mathtt{Fix}\ T$ *with rate at most (worst)* $c \equiv \sqrt{1 + \varepsilon - \kappa^2(1-\alpha)/\alpha} < 1$.

We are now ready to formulate the main convergence results.

Theorem 6.8 (Linear convergence of $T_{\mathtt{AP}}$ for (12)). *Let* $\widehat{\mathbf{a}} \in A$ *be a fixed point of* $T_{\mathtt{AP}} \equiv P_A P_B$ *and suppose that* $P_B(\widehat{\mathbf{a}}) = \left\{\widehat{\mathbf{b}} \equiv (\hat{x}_1, \hat{x}_2, \ldots, \hat{x}_m)\right\}$ *is singleton with* $\mathcal{G}_d(\hat{x}_d)$ *nonzero everywhere* ($\forall\, d = 1, 2, \ldots, m$). *Given a number* $\varepsilon > 0$, *let* $U_\varepsilon(\widehat{\mathbf{b}})$ *be the neighborhood of* $\widehat{\mathbf{b}}$ *on which* $T_{\mathtt{AP}}$ *is almost averaged with violation* ε *and averaging constant* $\alpha = 2/3$ *as determined by proposition 6.5. Suppose further that* $\widehat{\mathbf{a}} \in U_\varepsilon(\widehat{\mathbf{b}})$, $T_{\mathtt{AP}}(A \cap U_\varepsilon(\widehat{\mathbf{b}})) \subset U_\varepsilon(\widehat{\mathbf{b}})$ *and the mapping* $\Theta \equiv \mathtt{Id} - T_{\mathtt{AP}}$ *is metrically subregular on* $U_\varepsilon(\widehat{\mathbf{b}})$ *for 0 relative to* A *with modulus* $\kappa > \sqrt{2\varepsilon}$. *Then every iterative sequence generated by* $T_{\mathtt{AP}}$ *with the initial point in* $A \cap U_\varepsilon(\widehat{\mathbf{b}})$ *converges linearly to a point in* $\mathtt{Fix}\ T_{\mathtt{AP}}$ *with rate at most* $c \equiv \sqrt{1 + \varepsilon - \kappa^2/2} < 1$.

Proof. The assumption $\widehat{\mathbf{a}} \in U_\varepsilon(\widehat{\mathbf{b}})$ ensures that $T_{\mathtt{AP}}$ is pointwise almost averaged at $\widehat{\mathbf{a}}$ on $U_\varepsilon(\widehat{\mathbf{b}})$ with violation ε and averaging constant $\alpha = 2/3$ in view of proposition 6.5. Hence all the assumptions of lemma 6.7 are satisfied with $\Lambda = A$ and $U = U_\varepsilon(\widehat{\mathbf{b}})$ and the convergence statement follows as claimed. □

We next explain and remark on the assumptions imposed in theorem 6.8.

Remark 6.9. It is important to keep in mind that $U_\varepsilon(\widehat{\mathbf{b}})$ is not limited to some ball centered at $\widehat{\mathbf{b}}$. It can be an unbounded set, see, for example, the typical intuitive example of phase retrieval in [38, figure 3 and example 3.9(ii)]. This in particular makes the assumption $\widehat{\mathbf{a}} \in U_\varepsilon(\widehat{\mathbf{b}})$ not restrictive. A more general notion than prox-regularity called *regularity at a distance* was proposed in [38] for analyzing the RAAR algorithm for nonconvex and inconsistent feasibility. However, we are unable to verify that property for the high-NA phase retrieval problem and hence do not apply it to the analysis in this application paper to avoid further unverifiable assumptions.

Remark 6.10. Since the set B in (12) is compact, every iterative sequence generated by the AP methods has a subsequence converging to a point in $\mathtt{Fix}\ T_{\mathtt{AP}}$, a local best approximation point to B. Theorem 6.8 provides sufficient conditions for local linear convergence of the algorithm around a single fixed point. Its assumptions can be strengthened for all fixed points of $T_{\mathtt{AP}}$ to yield global convergence of the algorithm, but the quality of the fixed point it converges to and the convergence rate in general depend on where it starts as the problem is nonconvex. However, such additional assumptions would be unverifiable for the high-NA phase retrieval problem, we chose not to include them in this application paper.

Remark 6.11 (Necessity of metric subregularity). As mentioned early this section there are two groups of properties often required to prove convergence of nonconvex optimization algorithms. The geometry of the high-NA phase retrieval problem analyzed in section 5 yields the first one—*pointwise almost averagedness*. It has been known that the second one—*metric subregularity* is difficult to verify, but as been shown in [40] this condition is not only *sufficient* but also *necessary* for local linear convergence.

The mathematical complication of theorem 6.8 is mainly due to the *inconsistency* of the problem under study. In the consistent setting, it reduces to the following much simpler form, where the metric subregularity of $\mathtt{Id} - T_{\mathtt{AP}}$ also reduces to the more intuitive notion called *subtransversality* of the collection of sets $\{A, B\}$ at the intersection point. For cartoon model of phase retrieval consisting of two (products of) spheres, the subtransversality property is satisfied except when they are tangent. The proof of the next statement follows from the one of theorem 6.8 and is left out for brevity.

Corollary 6.12 (Linear convergence of $T_{\mathtt{AP}}$ for consistent (12)). *Consider the problem (12) with $A \cap B \neq \emptyset$. Let $\widehat{\mathbf{a}} \equiv [\hat{x}]_m \in A \cap B$ with $\mathcal{G}_d(\hat{x})$ nonzero everywhere $(\forall\, d = 1, 2, \ldots, m)$. Given a number $\varepsilon > 0$, let $\mathbb{B}_\varepsilon(\widehat{\mathbf{a}})$ be the ball on which $T_{\mathtt{AP}}$ is almost averaged with violation ε and averaging constant $\alpha = 2/3$ as determined by proposition 6.5. Suppose that the mapping $\Theta \equiv \mathtt{Id} - T_{\mathtt{AP}}$ is metrically subregular at $\widehat{\mathbf{a}}$ for 0 relative to A with modulus $\kappa > \sqrt{2\varepsilon}$. Then every iterative sequence generated by $T_{\mathtt{AP}}$ with the initial point in $A \cap \mathbb{B}_\varepsilon(\widehat{\mathbf{a}})$ converges linearly to a point in $\mathtt{Fix}\ T_{\mathtt{AP}}$ with rate at most $c \equiv \sqrt{1 + \varepsilon - \kappa^2/2} < 1$.*

Table 1. Parameters used in numerical simulations: NA-numerical aperture, λ-wavelength of illumination light (μm), s-pixel size (μm), m-number of images, $n \times n$-image size (pixels), w-noise model, and SNR-signal-to-noise ratio (decibels).

Parameter	NA	λ	s	m	$n \times n$	w	SNR
Value	0.95	0.3	0.06	7	128×128	Gaussian	30 dB

It is worth mentioning that the technical assumptions imposed in theorem 6.8 and corollary 6.12 concerning high-NA phase retrieval also remain unverifiable for the low-NA problem.

7. Numerical simulations

The goal of this section is to demonstrate that the new mathematical analysis obtained in this paper enables us to apply the class of projection algorithms for solving the high-NA phase retrieval problem. In contrast to a vast number of existing solution methods for low-NA phase retrieval, very few algorithms have been proposed for the high-NA case. The *vectorial PSF model-based alternating minimization* (VAM) algorithm was proposed in [56]. It outperforms several available high-NA phase retrieval approaches, including the *scalar PSF model-based alternating minimization* (SAM) algorithm of Hanser *et al* [23] which is limited in accuracy due to model deviations, and the modal-based approach through the use of extended Nijboer–Zernike expansion of Braat *et al* [9] which is of high computational complexity and excludes applications with discontinuous phase. The VAM algorithm is nothing else, but the AP method applied to the feasibility model (12). The projectors computed in section 3 enable the implementation of every projection method (not only those mentioned in section 4) for solving every corresponding feasibility model formulated in section 2. This section aims at demonstrating the improved performance of more delicate projection algorithms over available solution methods for high-NA phase retrieval. As projection methods have not been applied to high-NA phase retrieval before, their comparison is not a goal of this paper, which instead establishes groundwork enabling the implementation and analysis of this efficient class of solution methods for high-NA phase retrieval.

We consider the practically relevant simulation setting of high-NA phase retrieval as in [56, section 5] where the vectorial PSF (7) is taken as the forward imaging model for generating the images. The simulated imaging system has circular aperture with the amplitude $\mathcal{A}$ being the two-dimensional Gaussian distribution truncated at 0.5 on the boundary. We do 75 experiments for different phase realizations with values in $[-\pi, \pi]$. Each data set consists of seven out-of-focus PSF images which are uniformly separated by one depth of focus along the optical axis. A schematic diagram of this phase retrieval setup can be seen, for example, in [56, figure 1]. The generated PSF images after being normalized to unity energy are corrupted by additive white Gaussian noise with SNR 30 decibels (dB). Recall that $\text{SNR} = 10 \ln\left(P/P_0\right)$, where P and P_0 are the powers of the signal and the noise, respectively. The parameters used in the simulation experiments are summarized in table 1. The quality of phase retrieval is measured by the relative root mean square (RMS) error $\|\widehat{\Phi} - \Phi\| / \|\Phi\|$, where Φ and $\widehat{\Phi}$ are the simulation and the retrieved phase aberrations, respectively. As phase retrieval is ambiguous up to at least a global phase shift (a piston term or the first Zernike mode), the norms of the phases are computed with the piston terms removed.

We first analyze the performance of the SAM, VAM (equivalently, AP), DR, KM-DR, HPR, RAAR, RRR and DRAP algorithms for solving the feasibility problem (12), for which recall that the amplitude $\mathcal{A}$ is assumed unknown to the algorithms. As the DR, KM–DR, HPR and RRR algorithms are clearly outperformed by the RAAR and DRAP methods, we chose to

Table 2. The number of iterations (the second row) and the parameter β (the third row) of the algorithms used in numerical simulations. The averaged RMS errors of phase retrieval over 75 phase realizations are presented in the last row.

Algorithm	SAM	VAM	DRAP	RAAR	VAM_+	DRAP_+	RAAR_+
#Iterations	100	100	$30 + 20$	$30 + 20$	100	$30 + 20$	$30 + 20$
Parameter β			0.95	0.95		0.95	0.95
Error (%)	8.47	7.69	6.14	5.98	6.82	4.68	4.69

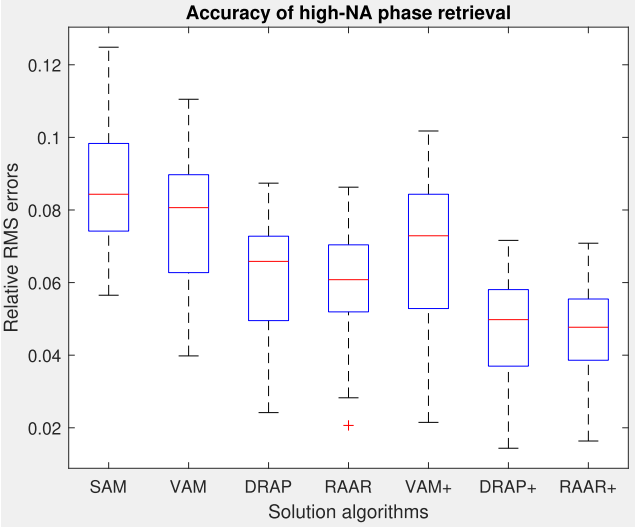

Figure 2. The box-plots show the improved performance of the RAAR and DRAP algorithms over available high-NA phase retrieval methods, including SAM [23] and VAM [56]. Each box-plot summarizes the numerical results in relative RMS errors of seventy-five examples with different phase realizations taking values in $[-\pi, \pi]$. The RAAR algorithm yields phase retrieval with the smallest RMS error on average, 5.98% compared to 8.47% of SAM, 7.69% of VAM and 6.14% of DRAP. RAAR also has smaller error variance than the others as indicated by its shorter box-plot. The additional '$_+$' sign in the algorithm names (for example, RAAR_+) indicates that the algorithms in addition know the amplitude $\mathcal{A}$, i.e. they are applied to the more informative feasibility model (17) instead of (12). The additional information of the amplitude $\mathcal{A}$ improves the performance of every solution method. In this case, we also observe the improved performance of DRAP_+ and RAAR_+ over VAM_+, with average relative RMS errors 4.68%, 4.69% and 6.82%, respectively. The RAAR_+ algorithm also has the smallest error variance.

skip their results for brevity. Table 2 shows the number of iterations (the second row), the tuning parameter β (the third row) of the algorithms, and the averaged RMS errors over the 75 experiments (the last row). Due to the extrapolation feature of RAAR and DRAP, each experiment with them is also followed by an averaging process of 20 iterations of AP, indicated by the term '+20' in the second row of table 2. Figure 2 shows the improved performance in terms of accuracy of RAAR and also DRAP over SAM and VAM. The RAAR algorithm yields phase retrieval with the smallest RMS error on average, 5.98% compared to 8.47% of SAM, 7.69% of VAM and 6.14% of DRAP as shown in the last row of table 2. RAAR also

has smaller error variance than the others as indicated by its shorter box-plot in figure 2. In terms of computational complexity, RAAR and DRAP (50 iterations) are much more efficient than VAM (100 iterations) as shown in table 2 (the second row). Note that SAM making use of the scalar PSF model has about six times lower complexity per iteration than the other methods; however, this advantage is often dominated by the disadvantage of model deviations for high-NA phase retrieval.

We consider the same 75 high-NA phase retrieval examples as above, but the amplitude $\mathcal{A}$ is now assumed known. The tighter feasibility model (17) then comes into play in place of (12). In this section, the algorithms applied to (17) will be indicated by the additional '$_+$' sign in their names (for example, RAAR$_+$) to distinguish with themselves for solving (12). We analyze the performance of the VAM$_+$ [56] (equivalently, AP$_+$), DR$_+$, KM–DR$_+$, HPR$_+$, RAAR$_+$, RRR$_+$ and DRAP$_+$ algorithms for solving (17). For the same reason as for solving (12), we chose to skip the phase retrieval results of DR$_+$, KM–DR$_+$, HPR$_+$ and RRR$_+$ for brevity. The additional information of $\mathcal{A}$ clearly improves the performance of every solution method as shown by figure 2, which also demonstrates the improved performance of DRAP$_+$ and RAAR$_+$ over VAM$_+$, with average relative RMS errors 4.68%, 4.69% and 6.82%, respectively. The RAAR$_+$ algorithm has the smallest error variance.

Remark 7.1. As discussed in remark 2.3, the choice of a feasibility model of high-NA phase retrieval depends on the specific conditions of the problem under consideration, e.g. the number of images, noise level and model deviations. In particular, for problems given sufficiently many noisy images like those simulated in this section, two-set feasibility models like (12) are a better choice than multiple-set models like (11). This is because the algorithms built on the two-set models implicitly include an averaging process on the data and thus can better suppress noise. For this reason, the cyclic-type algorithms are outperformed by those reported in table 2 and thus their performance was left out for brevity.

Acknowledgments

The authors thank the two anonymous referees for the careful reading of the manuscript and constructive comments and suggestions.

Data availability statement

No new data were created or analyzed in this study.

Funding

This project has received funding from the ECSEL Joint Undertaking (JU) under grant agreement No. 826589. The JU receives support from the European Union's Horizon 2020 research and innovation programme and Netherlands, Belgium, Germany, France, Italy, Austria, Hungary, Romania, Sweden and Israel. Russell Luke was supported in part by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation)-Project-ID 432680300-SFB 1456 and Project-ID LU 1702/1-1.

Conflict of interest

The authors declare no conflicts of interest.

ORCID iDs

Nguyen Hieu Thao https://orcid.org/0000-0002-1455-9169
Oleg Soloviev https://orcid.org/0000-0003-3761-9192
Russell Luke https://orcid.org/0000-0002-4508-7360
Michel Verhaegen https://orcid.org/0000-0002-7967-6884

References

[1] Antonello J and Verhaegen M 2015 Modal-based phase retrieval for adaptive optics *J. Opt. Soc. Am.* A **32** 1160–70
[2] Arridge S R 1999 Optical tomography in medical imaging *Inverse Problems* **15** R41–93
[3] Bauschke H H and Borwein J M 1996 On projection algorithms for solving convex feasibility problems *SIAM Rev.* **38** 367–426
[4] Bauschke H H and Combettes P L 2017 *Convex Analysis and Monotone Operator Theory in Hilbert Spaces* (*CMS Books in Mathematics/Ouvrages de Mathématiques de la SMC*) 2nd edn (Berlin: Springer)
[5] Bauschke H H, Combettes P L and Luke D R 2002 Phase retrieval, error reduction algorithm, and Fienup variants: a view from convex optimization *J. Opt. Soc. Am.* A **19** 1334–45
[6] Bauschke H H, Combettes P L and Luke D R 2003 Hybrid projection-reflection method for phase retrieval *J. Opt. Soc. Am.* A **20** 1025–34
[7] Bauschke H H, Noll D and Phan H M 2015 Linear and strong convergence of algorithms involving averaged nonexpansive operators *J. Math. Anal. Appl.* **421** 1–20
[8] Borwein J M and Tam M K 2014 A cyclic Douglas–Rachford iteration scheme *J. Optim. Theory Appl.* **160** 1–29
[9] Braat J J M, Dirksen P, Janssen A J E M, van Haver S and van de Nes A S 2005 Extended Nijboer–Zernike approach to aberration and birefringence retrieval in a high-numerical-aperture optical system *J. Opt. Soc. Am.* A **22** 2635–50
[10] Candès E J, Eldar Y C, Strohmer T and Voroninski V 2013 Phase retrieval via matrix completion *SIAM J. Imag. Sci.* **6** 199–225
[11] Cegielski A 2012 *Iterative Methods for Fixed Point Problems in Hilbert Spaces* (*Lecture Notes in Mathematics*) vol 2057 (Heidelberg: Springer)
[12] Dainty J C and Fienup J R 1987 Phase retrieval and image reconstruction for astronomy *Image Recovery: Theory Appl.* **13** 231–75
[13] de Visser C C, Brunner E and Verhaegen M 2016 On distributed wavefront reconstruction for large-scale adaptive optics systems *J. Opt. Soc. Am.* A **33** 817–31
[14] de Visser C C and Verhaegen M 2013 Wavefront reconstruction in adaptive optics systems using nonlinear multivariate splines *J. Opt. Soc. Am.* A **30** 82–95
[15] Doelman R, Hieu Thao N and Verhaegen M 2018 Solving large-scale general phase retrieval problems via a sequence of convex relaxations *J. Opt. Soc. Am.* A **35** 1410–9
[16] Dontchev A L and Rockafellar R T 2014 *Implicit Functions and Solution Mappings* 2nd edn (New York: Srpinger)
[17] Elser V, Lan T-Y and Bendory T 2018 Benchmark problems for phase retrieval *SIAM J. Imag. Sci.* **11** 2429–55
[18] Fienup J R 1982 Phase retrieval algorithms: a comparison *Appl. Opt.* **21** 2758–69
[19] Fienup J R 2013 Phase retrieval algorithms: a personal tour *Appl. Opt.* **52** 45–56
[20] Gerchberg R W and Saxton W O 1972 A practical algorithm for the determination of phase from image and diffraction plane pictures *Optik* **35** 237–46
[21] Gonsalves R A 1982 Phase retrieval and diversity in adaptive optics *Opt. Eng.* **21** 829–32

[22] Goodman J W 2017 *Introduction to Fourier Optics* 4th edn (New York: W. H. Freeman and Company)
[23] Hanser B M, Gustafsson M G L, Agard D A and Sedat J W 2003 Phase retrieval for high-numerical-aperture optical systems *Opt. Lett.* **28** 801
[24] Hanser B M, Gustafsson M G L, Agard D A and Sedat J W 2004 Phase-retrieved pupil functions in wide-field fluorescence microscopy *J. Microsc.* **216** 32–48
[25] Hardy J W and Thompson L 2000 Adaptive optics for astronomical telescopes *Phys. Today* **53** 69
[26] Harrison R W 1993 Phase problem in crystallography *J. Opt. Soc. Am.* A **10** 1046–55
[27] Hauptman H 1986 The direct methods of x-ray crystallography *Science* **233** 178–83
[28] Hesse R and Luke D R 2013 Nonconvex notions of regularity and convergence of fundamental algorithms for feasibility problems *SIAM J. Optim.* **23** 2397–419
[29] Ioffe A D 2017 *Variational Analysis of Regular Mappings* (*Theory and Applications. Springer Monographs in Mathematics*) (Berlin: Springer)
[30] Kim T, Zhou R, Goddard L and Popescu G 2016 Solving inverse scattering problems in biological samples by quantitative phase imaging *Laser Photon. Rev* **10** 13–39
[31] Klatte D and Kummer B 2002 *Nonsmooth Equations in Optimization: Regularity, Calculus, Methods and Applications* (*Nonconvex Optimization and its Applications*) vol 60 (Dordrecht: Kluwer)
[32] Kruger A Y 2006 About regularity of collections of sets *Set-Valued Anal.* **14** 187–206
[33] Kruger A Y, Luke D R and Hieu Thao N 2018 Set regularities and feasibility problems *Math. Program.* B **168** 279–311
[34] Lin J, Rodríguez-Herrera O G, Kenny F, Lara D and Dainty J C 2012 Fast vectorial calculation of the volumetric focused field distribution by using a three-dimensional Fourier transform *Opt. Express* **20** 1060–9
[35] Luke D R 2005 Relaxed averaged alternating reflections for diffraction imaging *Inverse Problems* **21** 37–50
[36] Luke D R 2008 Finding best approximation pairs relative to a convex and prox-regular set in a Hilbert space *SIAM J. Optim.* **19** 714–39
[37] Luke D R, Burke J V and Lyon R G 2002 Optical wavefront reconstruction: theory and numerical methods *SIAM Rev.* **44** 169–224
[38] Luke D R and Martins A-L 2020 Convergence analysis of the relaxed Douglas–Rachford algorithm *SIAM J. Optim.* **30** 542–84
[39] Luke D R, Sabach S and Teboulle M 2019 Optimization on spheres: models and proximal algorithms with computational performance comparisons *SIAM J. Math. Data Sci.* **1** 408–45
[40] Luke D R, Teboulle M and Hieu Thao N 2020 Necessary conditions for linear convergence of iterated expansive, set-valued mappings *Math. Program.* A **180** 1–31
[41] Luke D R, Hieu Thao N and Tam M K 2018 Quantitative convergence analysis of iterated expansive, set-valued mappings *Math. Oper. Res.* **43** 1143–76
[42] Mansuripur M 2009 *Classical Optics and its Applications* (Cambridge: Cambridge University Press)
[43] McCutchen C W 2002 Generalized aperture and the three-dimensional diffraction image: erratum *J. Opt. Soc. Am.* A **19** 1721
[44] Millane R P 1990 Phase retrieval in crystallography and optics *J. Opt. Soc. Am.* A **7** 394–411
[45] Mordukhovich B S 2018 *Variational Analysis and Applications* (Berlin: Springer)
[46] Mugnier L M, Blanc A and Idier J 2006 Phase diversity: a technique for wave-front sensing and for diffraction-limited imaging *Adv. Imag. Electron Phys.* **141** 1–76
[47] Pierra G 1984 Decomposition through formalization in a product space *Math. Program.* **28** 96–115
[48] Poliquin R, Rockafellar R and Thibault L 2000 Local differentiability of distance functions *Trans. Am. Math. Soc.* **352** 5231–49
[49] Richards B and Wolf E 1959 Electromagnetic diffraction in optical systems: II. Structure of the image field in an aplanatic system *Proc. R. Soc. A Math. Phys. Eng. Sci.* **253** 358–79
[50] Rockafellar R T and Wets R J 1998 *Variational Analysis* (*Grundlehren Math. Wiss*) (Berlin: Springer)
[51] Shechtman Y, Eldar Y C, Cohen O, Chapman H N, Miao J and Segev M 2015 Phase retrieval with application to optical imaging: a contemporary overview *IEEE Signal Process. Mag.* **32** 87–109
[52] Soulez F, Thiébaut É, Schutz A, Ferrari A, Courbin F and Unser M 2016 Proximity operators for phase retrieval *Appl. Opt.* **55** 7412–21
[53] Southwell W H 1977 Wave-front analyzer using a maximum likelihood algorithm *J. Opt. Soc. Am.* **67** 396–9

[54] Hieu Thao N 2018 A convergent relaxation of the Douglas–Rachford algorithm *Comput. Optim. Appl.* **70** 841–63
[55] Hieu Thao N, Luke D R, Soloviev O and Verhaegen M 2020 Phase retrieval with sparse phase constraint *SIAM J. Math. Data Sci.* **2** 246–63
[56] Hieu Thao N, Soloviev O and Verhaegen M 2020 Phase retrieval based on the vectorial model of point spread function *J. Opt. Soc. Am.* A **37** 16–26
[57] Hieu Thao N, Soloviev O and Verhaegen M 2021 Convex combination of alternating projection and Douglas–Rachford operators for phase retrieval *Adv. Comput. Math.* **47** 33