Coded Diffraction Pattern Design Algorithm for Phase Retrieval in Optical Imaging

Samuel Eduardo Pinilla Sánchez

Trabajo de Grado para optar al título de Doctor en Ingeniería

Director

Ph.D Henry Arguello Fuentes

Doctorado en Ingeniería

Universidad Industrial de Santander

Facultad de Ingenierías Fisicomecánicas

Escuela de Ingenierías Eléctrica, Electrónica y de Telecomunicaciones

Bucaramanga

2020

Dedication

This work is dedicated to all those people who supported the development and execution of this degree project.

Acknowledgments

I thank my family for the financial and moral support they had for me during the development of my career. I also thank my friends and colleagues for their experiences in university years.

Table of Content

List of Figures

List of Tables

List of Appendices

pág.

Glossary

ω Gaussian noise with distribution $\mathcal{N}(0,\sigma^2)$.

σ Noise level.

$(\cdot)^T$ matrix transpose operation.

$\overline{(\cdot)}$ Conjugate operation.

$(\cdot)^H$ Conjugate transpose operation.

$|\cdot|$ Smaller integer of the given number.

$card(\mathscr{R})$ Cardinality of the set.

$\|\mathbf{W}\|_p$ Cardinality of the set.

$card(\mathscr{R})$ p-norm of a matrix $\left(\|\mathbf{W}\|_p = \left[\sum_i \sigma_i^p(\mathbf{W})\right]^{1/p}\right)$

$\sigma_i(\mathbf{W})$ ith singular value.

$\|\mathbf{w}\|_p$ usual ℓ_p norm.

Resumen

Título: Coded Diffraction Pattern Design Algorithm for Phase Retrieval in Optical Imaging [*]

Autor: Samuel Eduardo Pinilla Sánchez [**]

Palabras Clave: Recuperación de fase, zona de difracción, matriz de detección, apertura codificada.

Descripción: La recuperación de fase es un problema inverso que consiste en estimar una escena a partir de intensidades de difracción. Este problema aparece en la formación de imágenes ópticas, que tiene tres zonas principales de difracción donde se pueden adquirir medidas, cerca, media y lejos. Trabajos recientes han empleado algoritmos de descenso de gradiente para resolver el problema de recuperación de fase relacionado con la zona lejana, creando redundancia en el proceso de medición al incluir una apertura codificada, que permite modular la escena y adquirir patrones de difracción codificados (CDP). Sin embargo, este problema no se ha estudiado teóricamente para CDP en las zonas cercana y media. Además, la estructura de la apertura codificada se selecciona al azar, lo que conduce a estimaciones subóptimas. Esta tesis proporciona garantías teóricas para la recuperación de una escena adquirida en las tres zonas de difracción utilizando modulaciones admisibles. Con base en los resultados teóricos, se demostrará que la calidad de reconstrucción de la imagen depende directamente de la estructura de apertura codificada; por lo tanto, el diseño de la matriz de detección es fundamental para obtener una alta calidad de reconstrucción. Específicamente, las aperturas codificadas se pueden diseñar para mejorar la calidad de la señal reconstruida. Además, cuando la escena se puede representar escasamente de alguna manera, su soporte se puede estimar mejor para una elección cuidadosa de los elementos de codificación. Los resultados numéricos muestran que la escena se recupera con éxito mediante el uso de aperturas codificadas diseñadas con hasta 40 % menos de medidas en comparación con conjuntos no diseñados.

---

[*] Trabajo de grado

[**] Facultad de Ingenierías Físico-Mecánicas. Escuela de Ingenierías Eléctrica, Electrónica y telecomunicaciones. Director: Ph.D Henry Arguello Fuentes, Doctorado en Ingeniería.

Abstract

Title: Coded Diffraction Pattern Design Algorithm for Phase Retrieval in Optical Imaging [*]

Author: Samuel Eduardo Pinilla Sánchez [**]

Keywords: Phase retrieval, diffraction zone, sensing matrix, coded aperture.

Description: Phase retrieval is an inverse problem that consists in estimating a scene from diffraction intensities. This problem appears in optical imaging, which has three main diffraction zones where the measurements can be acquired, i.e., near, middle and far. Recent works have employed gradient descent algorithms to solve the phase retrieval problem related to the far zone, creating redundancy in the measurement process by including a coded aperture, which allows to modulate the scene and acquire coded diffraction patterns (CDP). However, in the state-of-the-art, the PR problem has not been theoretically studied for CDP at the near and middle zones. Moreover, the structure of the coded aperture is selected at random, leading to suboptimal estimations. This thesis provides theoretical guarantees for the recovery of a scene from CDP acquired at the three diffraction zones using admissible modulations. Based on the theoretical results, it will be shown that the image reconstruction quality directly depends on the coded aperture structure; therefore, designing the sensing matrix is critical to obtain high reconstruction quality. Specifically, the coded apertures can be designed in order to boost the quality of the reconstructed signal. Moreover, when the scene can be sparsely represented in some basis, its support can be better estimated for a carefully choice of the coding elements in the modulation process. Numerical results show that the scene is successfully recovered by using designed coded apertures with up to 40 % less measurements compared to non-designed ensembles.

---

[*] Ph.D Thesis

[**] Facultad de Ingenierías Físico-Mecánicas. Escuela de Ingenierías Eléctrica, Electrónica y telecomunicaciones. Director: Ph.D Henry Arguello Fuentes, Doctorado en Ingeniería.

Introduction

Phase retrieval (PR) is an inverse problem of considerable importance in several areas of science, where measuring the phase information is hard or even infeasible. In particular, this problem appears in applications such as X-ray crystallography Millane (1990), astronomy Fienup and Dainty (1987), and diffractive optical imaging (DOI) Shechtman et al. (2015), with the latter being the object of study of this work. Exploring phase retrieval in optical settings, specifically, when the light originates from a laser, is natural since optical detection devices (*e.g.*, charge-coupled device (CCD) cameras, photosensitive films, and the human eye) cannot measure the phase of a light wave Shechtman et al. (2015). This is because, generally, optical measurement devices that rely on converting photons to electrons do not capture the phase directly since the electromagnetic field oscillates at rates of $\sim 10^{15}$ Hz, and no electronic measurement device can work that fast.

Mainly, DOI has three diffraction zones where the data can be acquired depending on the propagation distance, known as the *near*, *middle* and *far* zones Poon and Liu (2014). Important imaging applications have been developed by taking advantage of the particular properties of each diffraction zone. For instance, the near diffraction zone is considered in applications such as scanning near-field optical microscopy Dürig et al. (1986), near-field Raman Imaging Jahncke et al. (1995), and near-field spectroscopy Hess et al. (1994), since the spatial resolution for a nanostructure can be overcome if the sample is scanned at the near zone. This is possible because the optical resolution of the transmitted light depends on the diameter of the sample instead of the wavelength Pohl and Courjon (2012). On the other hand, applications such as Fresnel holography Poon and Liu

(2014) and lens-less imaging Shimano et al. (2018) take advantage of the middle diffraction zone, also known as Fresnel diffraction, to develop new acquisition imaging devices Shimano et al. (2018), and optical elements such as Fresnel lenses Sao et al. (2018). Finally, the far zone, also known as Fraunhofer diffraction, is the most popular diffraction phenomenon in optics since it allowed the development of applications such as crystallography, astronomical imaging, microscopy, among others Goodman (2005). In summary, all the aforementioned applications highlight the importance of analytically studying all the diffraction zones.

Mathematically, PR for the k-th diffraction zone consists in solving quadratic equations of the form $y_{i,k} = |\langle \mathbf{a}_{i,k}, \mathbf{x} \rangle|^2, \;\; i = 1, \cdots, m$, where $\mathbf{a}_{i,k} \in \mathbb{C}^n$ are the known sampling vectors, $\mathbf{x} \in \mathbb{C}^n$ is the unknown scene of interest, $y_{i,k}$ are the acquired diffraction patterns, and $k = 1, 2, 3$ indexes the near, middle and far zones respectively. Recent works have theoretically solved this inverse problem in the far zone ($k = 3$) Candes et al. (2015b); Gross et al. (2017); Candes et al. (2015c), creating redundancy in the measurement process by including a coded aperture, which allows to modulate the scene and acquire intensity measures known as coded diffraction patterns (CDP), as illustrated in Fig. 1. Analytically, the effect of the coded aperture in the modulation process is included in the sampling vectors, and has allowed to provide uniqueness guarantees (up to a unimodular constant) for a particular class of coded apertures Candes et al. (2015b). These theoretical results were not possible a decade ago. More details about the history of this problem can be found in Shechtman et al. (2015).

Several algorithms have been proposed to retrieve the phase by applying non-convex formulations. To name a few, the wirtinger flow (WF) Candes et al. (2015c), truncated wirtinger

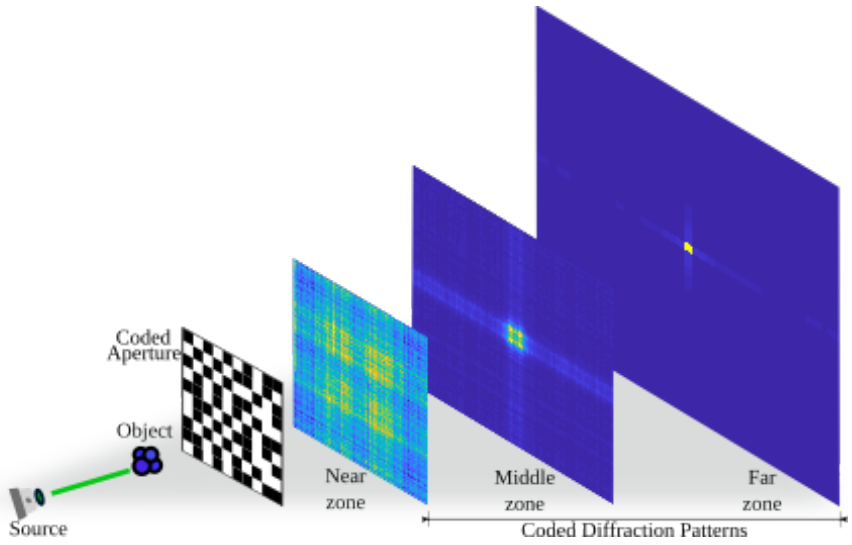

Figura 1. Illustration of a coded optical imaging system. A coded aperture is introduced to modulate the scene in order to acquire coded diffraction patterns.

flow (TWF) Chen and Candes (2015), truncated amplitude flow (TAF) Wang et al. (2018a), total-variation-based methods Chang et al. (2018), PR via smoothing function (PRSF) Pinilla et al. (2018a), and reweighted amplitude flow (RAF) Wang et al. (2018b). Additionally, in various applications the scene $\mathbf{x}$ is naturally sparse or admits a sparse representation on some basis Jaganathan et al. (2015). Indeed, recent algorithms have been developed to solve the PR problem under sparsity assumptions. Some examples of these methods are, sparse wirtinger flow (SWF) Yuan et al. (2017), sparse PR via truncated amplitude flow (SPARTA) Wang et al. (2016b) and sparse PR algorithm via smoothing function (SPRSF) Pinilla et al. (2018b). It is worth to mention that the sparse PR problem has been studied when the sampling vectors $\mathbf{a}_i$ follow a Gaussian distribution, implying that they do not model a realistic acquisition setup. Moreover, an important characteristic of all the aforementioned recovery methods is that they require a properly designed initialization strategy to guarantee convergence. In fact, different initialization methodologies have been propo-

sed to obtain an initial guess from measurements at the far zone, such as, spectral Candes et al. (2015c), orthogonality-promoting (OP) Wang et al. (2018a) and weighted maximal correlation (WMC) Wang et al. (2018b) initializations.

To date, theoretical recovery guarantees from CDP acquired at the near and middle diffraction zones have not been established. For instance, in Shevkunov et al. (2018) CDP are captured in the near zone with a lensless system, whereas, in Horisaki et al. (2017) CDP are captured in the middle zone with a single pixel system using structured light. Nevertheless, the theoretical guarantees in these cases are based on CDP acquired in the far zone Candes et al. (2015b), since it is the most popular scenario for phase retrieval.

One of the drawbacks in CDP is that state-of-the-art coded apertures allow coding elements with absolute value greater than 1 Candes et al. (2015b); Gross et al. (2017); Candes et al. (2015c), which is physically unfeasible because it increases the energy of the scene in the modulation process. Moreover, their spatial structure is selected at random, which limits the reconstruction quality, and also increases the required number of measurements to retrieve the phase. More precisely, the literature in areas such as compressive spectral imaging and computer tomography has shown that designing the coded apertures yields to better reconstructions Correa et al. (2016); Mojica et al. (2017). As a result, several coded aperture design strategies have been developed in the state-of-the-art, to name a few, based on gradient descend method Mojica et al. (2017), and greedy methodologies such as direct binary search (DBS) introduced in Chandu et al. (2013) and blue noise patterns in Correa et al. (2016).

Scope of the Thesis

Despite the satisfactory performance of reconstruction algorithms to solve the phase retrieval problem, a proper design of the sensing matrix is critical to obtain high image reconstruction quality Arguello and Arce (2014); Pinilla et al. (2018d). However, previous works have not focused on coding pattern designs, nor the developing of reconstruction algorithms that take into account the structure of the coded measurements. Additionally, in the state-of-the-art, the PR problem has not been theoretically studied for CDP at the near and middle zones.

Therefore, this thesis first proves that the phase of a scene can be recovered, with high probability, from CDP acquired at the three diffraction zones using feasible modulations. The recovery conditions provided in this work establish that image reconstruction quality directly depends on the coded aperture design. This theoretical analysis shows the crucial role of the coded aperture in reconstructing an image from coded diffraction patterns. Therefore, a greedy design strategy based on the theoretical result is also developed. This strategy consists in optimizing the concentration of measure of the sensing matrix. The resultant structures allow uniform sensing across the spatial dimensions of the scene and are physically implementable, improving image reconstruction quality compared with non-designed ensembles. In fact, in the case when the scene is sparsely represented in some basis, the support of the scene is better estimated for a suitable choice of the coding elements in the modulation process. Further, given the importance of a proper initialization to solve the PR problem, an extension of the Orthogonally-promoting initialization introduced in Wang et al. (2018a) for CDP is developed. Numerical results, based on admissible modulations,

show that using designed coded apertures the scene is successfully recovered using up to 40% less measurements compared with non-designed ensembles. Further, the relative error in the initialization stage using designed coded apertures decreases up to 50% compared with non-designed structures.

In summary, the contribution of this research includes the design, modeling, and testing, of the optimal sensing matrix and the reconstruction method for phase retrieval from coded measurements at the three diffraction zones.

Publications and author's contribution

Most of the material presented in this thesis appears in the following publications by the author:

- Samuel Pinilla, Jorge Bacca, and Henry Arguello. Phase retrieval algorithm via nonconvex minimization using a smoothing function. IEEE Transactions on Signal Processing, 66(17):4574-4584, 2018.
- Samuel Pinilla, Juan Poveda, and Henry Arguello. Coded diffraction system in x-ray crystallography using a boolean phase coded aperture approximation. Optics Communications, 410:707-716, 2018.
- Samuel Pinilla, Hans García, Luis Díaz, Juan Poveda, and Henry Arguello. Coded aperture design for solving the phase retrieval problem in x-ray crystallography. Journal of Computational and Applied Mathematics, 338:111-128, 2018.
- Bacca, Jorge, Samuel Pinilla, and Henry Arguello. Super-Resolution Phase Retrieval from Designed Coded Diffraction Patterns. IEEE Transactions on Image Processing (2019).

- Pinilla, Samuel, Tamir Bendory, Yonina C. Eldar, and Henry Arguello Frequency-Resolved Optical Gating Recovery via Smoothing Gradient. IEEE Transactions on Signal Processing 67.23 (2019): 6121-6132.
- Andrés Guerrero, Samuel Pinilla, and Henry Arguello. Phase recovery guarantees from designed coded diffraction patterns in optical imaging. IEEE Trans. on Image Proc, 29:5687-5697, 2020.
- Jerez, Andres, Samuel Pinilla, and Henry Arguello. Fast Target Detection via Template Matching in Compressive Phase Retrieval. IEEE Transactions on Computational Imaging (2020).

The whole thesis and the publications cited above represent original work, of which the author has been the main contributor or coauthor. However, this work would not have been possible without the support, criticism, and help of expert co-authors as well colleagues, among whom I wish to mention Yonina C. Eldar, Tamir Bendory, Jorge Bacca, and Andrés Jerez. The numerous suggestions which came from the anonymous reviewers of the above publications.

Structure of the Thesis

The document is organized as follows: Chapter 2 presents the fundamental theoretical background including details of coded aperture designs. Chapter 3 develops theoretical recovery guarantees from CDP acquired at the three diffraction zones. Chapter 4 contains the proposed phase recovery procedure from CDP. Chapter 9 includes numerical results along with the analysis of attained theoretical results. Finally, Chapter 10 contains analytical proofs of the results.

1. Objectives

General objective

To design the sensing matrix and a recovery algorithm to reduce the number of measurements to retrieve the phase from coded diffraction patterns in optical imaging .

Specific objectives

To establish a mathematical model of the acquisition process of coded diffraction patterns in optical imaging.

To develop a computational algorithm to simulate the modeled coded measurements in optical imaging.

To derive analytical conditions to optimally design the sensing matrix to improve the phase reconstruction quality in optical imaging.

To design a reconstruction algorithm that adjusts the acquisition process and the designed optimal coding patterns to retrieve the phase from coded measurements.

To evaluate the performance retrieving the phase of the designed reconstruction algorithm and the sensing matrix against non-designed of the state-of-the-art.

2. Theoretical Background

In this chapter the phase retrieval problem for each diffraction zone is exposed. Additionally, some details on optical setups to acquire coded diffraction patterns are provided.

2.1. Phase Retrieval from Coded diffraction patterns

In optical imaging, a coherent beam strikes the object and the phaseless measurements can be acquired, at a specific propagation distance z, at three diffraction zones known as near, middle and far zones Goodman (2005), as illustrated in Fig. 2. Specifically, Fig. 2(a) illustrates a lens-less diffractive imaging system, which traditionally has three diffraction zones, that are determined by

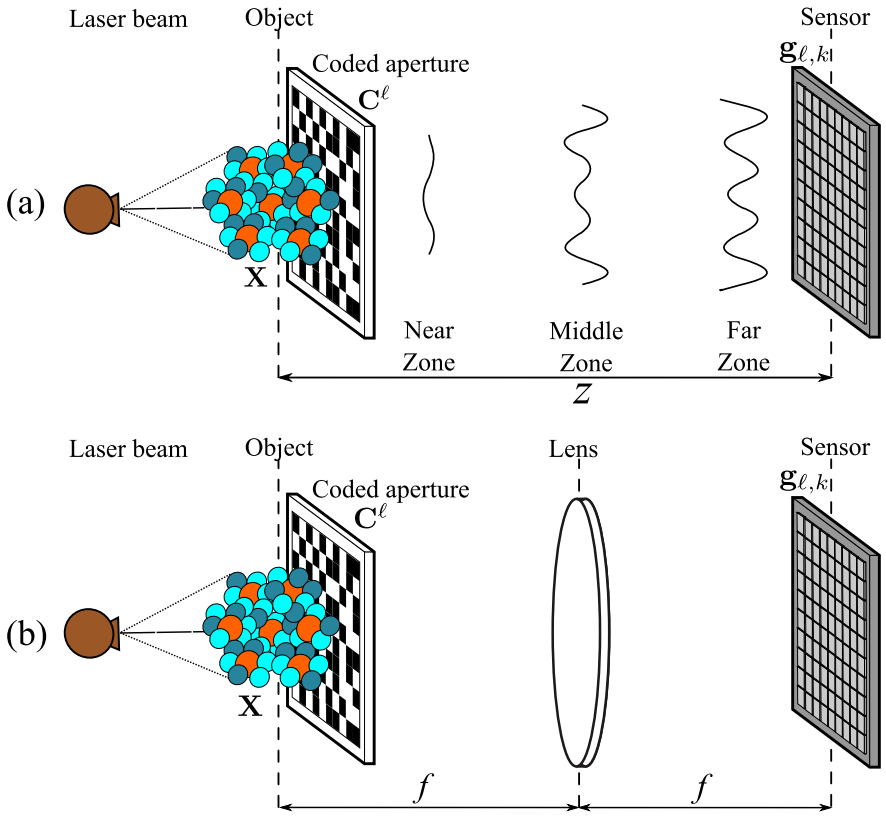

Figura 2. Optical setups to obtained coded diffraction patterns. (a) Lens-less imaging and (b) $2f$-optical systems.

the distance between the object and the sensor Goodman (2005). On the other hand, Fig. 2(b) represents a $2f$-optical setup, in which, the lens generates the diffraction patterns that are then recorded by the sensor. This system is equivalent to the first one when the detection distance corresponds to the far field.

Observe that a coded aperture is introduced in the object plane to modulate the scene $\mathbf{X} \in \mathbb{C}^{N\times N}$ and the CDP are acquired by the sensor. In fact, changing the spatial configuration of the coded aperture allows the system to acquire multiple projections of the scene. Mathematically, $\mathbf{C}^{\ell} \in \mathbb{C}^{N\times N}$, in Fig. 2, models the coded aperture at the ℓ-th projection where $\ell = 1, \cdots, L$ with L the total number of projections. There are several ways of achieving modulations of this type: using a phase mask, or using an optical grating to modulate the illumination beam as mentioned in Loewen and Popov (2018), or even by techniques from ptychography which scan an illumination patch on an extended specimen Rodenburg (2008); Thibault et al. (2009).

For ease of exposition, some definitions are considered through this chapter to facilitate the mathematical derivations. Let, $(\mathbf{C}^{\ell})_{s,u}$, and $(\mathbf{X})_{s,u}$ be the (s,u)-th spatial index of the coded aperture used for the ℓ-th projection and the scene, respectively. Further, the vector representation of the scene can be defined as $(\mathbf{x})_q = (\mathbf{X})_{q-vN,v+1}$, where $\mathbf{x} \in \mathbb{C}^n$, and one can take $\mathbf{D}_{\ell} \in \mathbb{C}^{n\times n}$ as the diagonal matrix whose entries are the elements of $\mathbf{C}^{\ell}$ given by $(\mathbf{D}_{\ell})_{q,q} = (\mathbf{C}^{\ell})_{q-vN,v+1}$, for $v = \lfloor \frac{q-1}{N} \rfloor$, $q = 1, \cdots, n$, and $n = N^2$. It is worth to point out that to mathematically model the CDP acquired at the three diffraction zones, and to establish the recovery guarantees the matrix $\mathbf{D}_{\ell}$ will be used, and the vector $\mathbf{x}$. In addition, to easily describe the design strategy of the coded aperture in Chapter 3 $\mathbf{C}^{\ell}$ will be used .

In order to state the mathematical model of the CDP acquired at the three diffraction zones, define $\mathbf{F} = [\mathbf{f}_1, \cdots, \mathbf{f}_n]^H \in \mathbb{C}^{n\times n}$ as the discrete Fourier transform matrix, where

$$\mathbf{f}_p^H = \frac{1}{\sqrt{n}}[\omega^{-0(p-1)}, \omega^{-1(p-1)}, \cdots, \omega^{-(n-1)(p-1)}], \tag{1}$$

with $p = 1, \cdots, n$, and $\omega = e^{\frac{2\pi j}{n}}$ is the n-th root of unity. Further, to the near and middle zones one have to consider two auxiliary orthogonal diagonal matrices $\mathbf{T} \in \mathbb{C}^{n\times n}$, and $\mathbf{Q} \in \mathbb{C}^{n\times n}$ that depend on the propagation distance z and the wavelength of the coherent beam, as shown in Fig. 1. The discrete version of $\mathbf{T}$ and $\mathbf{Q}$ are modeled in Shechtman et al. (2015)(Poon and Liu, 2014, Chapter 4). Then, the acquired CDP at the three diffraction zones for the ℓ-th projection are given by Poon and Liu (2014)

$$\begin{aligned}
\mathbf{g}_{\ell,1} &= |\mathbf{FTF}^H\mathbf{D}_\ell\mathbf{x}|^2 + \omega_{\ell,1} && (\textit{Near zone}) ,\\
\mathbf{g}_{\ell,2} &= |\mathbf{F}^H\mathbf{QD}_\ell\mathbf{x}|^2 + \omega_{\ell,2} && (\textit{Middle zone}) ,\\
\mathbf{g}_{\ell,3} &= |\mathbf{FD}_\ell\mathbf{x}|^2 + \omega_{\ell,3} && (\textit{Far zone}) ,
\end{aligned} \tag{2}$$

for $\ell = 1, \cdots, L$, with $\mathbf{x} \in \mathbb{C}^n$ the desired unknown scene, and $\omega_{\ell,k}$ is the observed additive noise. Thus, from (2) the PR problem for the k-th diffraction zone consists in estimating $\mathbf{x}$ from the set of phaseless measurements $\{\mathbf{g}_{\ell,k}\}_{\ell=1}^{L}$.

It is worth noticing that the propagation matrices for the near and middle diffraction zones significantly differ from the far zone which is the most studied scenario in the literature, therefore,

the theoretical results from Candes et al. (2015b); Gross et al. (2017) cannot be directly applied. In fact, the PR problem associated with the far zone ($k = 3$) as in (2) has been extensively analyzed Wang et al. (2018b); Jaganathan et al. (2015); Candes et al. (2015a); Bandeira et al. (2014); Kolte and Özgür (2016) assuming that the entries of $\mathbf{C}^{\ell}$ are *i.i.d* copies of a random variable $d \in \mathbb{C}$ satisfying the following condition

$$|d| \leq M, \;\; \mathbb{E}[d] = 0, \;\; \mathbb{E}[d^2] = 0, \;\; \mathbb{E}[|d|^4] = 2\mathbb{E}[|d|^2]^2, \tag{3}$$

where $\mathbb{E}[\cdot]$ represents the expected value. Some examples of coding elements that have been used in the-state-of-the-art based on (3) are listed in Table 1.

Table 1
State-of-the-art coding elements

Random Variable	**Coding Probability**	**Ref.**
$d = \{1, \sqrt{6}\}$	$\{\frac{4}{5}, \frac{1}{5}\}$	Candes et al. (2015b)
$d = \{\sqrt{2}, 0, -\sqrt{2}\}$	$\{\frac{1}{4}, \frac{1}{2}, \frac{1}{4}\}$	Gross et al. (2017)
$d = \{\frac{\sqrt{2}}{2}, \sqrt{3}\}$	$\{\frac{4}{5}, \frac{1}{5}\}$	Candes et al. (2015c)

In Table 1, only real numbers are included, however, in general the random variable d can take complex values. Note that the coding elements $\sqrt{2}, \sqrt{6}$, and $\sqrt{3}$ in Table 1 are unfeasible because their absolute value is greater than 1, which increases the power of the scene in the modulation process. Furthermore, even when $M = 1$ in (3), the three remaining conditions limit the set of possible coding elements because these only allow random variables that have an expected value equal to zero. Specifically, observe that $d = \{0, 1\}$, which models blocking and unblocking elements respectively, is the most natural coding variable for imaging applications, Correa et al.

(2016); Mojica et al. (2017); Arguello and Arce (2014), however it is not considered in (3), since its expected value is greater than zero. In addition to the limitations of the state-of-the-art theory of the PR problem from CDP at the far zone, it is worth to highlight that at this point the PR problems from CDP associated to the near and middle zones have not been theoretically established.

2.2. Motivation

State-of-the-art has successfully proved that the phase can be recovered from coded diffraction measurements. Also, different algorithms have been developed to recover the phase from generalized random quadratic projections, that is, assuming that the sampling vectors are normally distributed. However, it was previously mentioned both the sensing and the reconstruction process have been separately explored to retrieve the phase from coded measurements. Therefore, this dissertation proposal aims to theoretically study how much the quality of the recovered signal improves when the sensing matrix and the reconstruction strategy are jointly designed to recover the phase in an optical imaging system. Specifically, this thesis proves that the reconstruction quality directly depends on the coded aperture design. This analysis shows the crucial role of the coded aperture in reconstructing an image from coded diffraction patterns.

In addition, to avoid the previously mentioned drawbacks of the state-of-the-art, this work considers that a random variable d is admissible if the following definition is satisfied.

Definition 2.2.1. (*Admissible Random Variable*). A discrete random variable obeying $|d| \leq 1$, is said to be *admissible*.

It is important to remark that the inequality in Definition 2.2.1 imposes that d cannot in-

crease the power of the scene. Also, notice that Definition 2.2.1 considers a larger set of coding elements than (3). In addition, Chapter 3, in contrast to the state-of-the-art, theoretically establishes that the PR problem can be solved from CDP acquired at the three diffraction zones considering d as in Definition 2.2.1.

3. Theoretical Recovery Guarantees

This chapter presents the theoretical conditions under the phase retrieval problem from coded diffraction patterns at the different zones can be solved following Definition 2.2.1. As a result, a greedy design strategy of the coded aperture is developed that pursuits to fulfill these theoretical conditions.

3.1. Uniqueness Conditions

In order to provide theoretical guarantees for a scene $\mathbf{x} \in \mathbb{C}^n$ to be recovered from CDP at the three diffraction zones, we have to rewrite (2). Define the global noiseless measurement vectors $\mathbf{y}_k = [\mathbf{g}_{1,k}^T, \cdots, \mathbf{g}_{L,k}^T]^T \in \mathbb{R}^{m=nL}$ for $k = 1, 2, 3$, and consider the matrices $\mathbf{A}_k$ given by

$$\begin{aligned}
\mathbf{A}_1 &= \left[\overline{\mathbf{D}}_1 \mathbf{F}\overline{\mathbf{T}}\mathbf{F}^H, \cdots, \overline{\mathbf{D}}_L \mathbf{F}\overline{\mathbf{T}}\mathbf{F}^H\right]^H && (\textit{Near zone}) \,, \\
\mathbf{A}_2 &= \left[\overline{\mathbf{D}}_1 \overline{\mathbf{Q}}\mathbf{F}, \cdots, \overline{\mathbf{D}}_L \overline{\mathbf{Q}}\mathbf{F}\right]^H && (\textit{Middle zone}) \,, \\
\mathbf{A}_3 &= \left[\overline{\mathbf{D}}_1 \mathbf{F}^H, \cdots, \overline{\mathbf{D}}_L \mathbf{F}^H\right]^H && (\textit{Far zone}) \,.
\end{aligned} \tag{4}$$

Observe that from (4), one has that (2) can be succinctly expressed as

$$\mathbf{y}_k = |\mathbf{A}_k \mathbf{x}|^2, \tag{5}$$

for $k = 1, 2, 3$. Now, putting (1) and (5) together it can be concluded that the i-th entry of the vector $\mathbf{y}_k$, i.e. $y_{i,k}$, is given by

$$y_{i,k} = |\mathbf{a}_{i,k}^H \mathbf{x}|^2 = \mathbf{a}_{i,k}^H \mathbf{x}\mathbf{x}^H \mathbf{a}_{i,k}, \tag{6}$$

where the vector $\mathbf{a}_{i,k}$ is the i-th column of matrix $\mathbf{A}_k$ defined as

$$\begin{aligned} \mathbf{a}_{i,1} &= \overline{\mathbf{D}}_{r_i} \mathbf{F} \overline{\mathbf{T}} \mathbf{f}_{u_i}, && (\textit{Near zone}) \,, \\ \mathbf{a}_{i,2} &= \overline{\mathbf{D}}_{r_i} \overline{\mathbf{Q}\mathbf{f}}_{u_i}, && (\textit{Middle zone}) \,, \\ \mathbf{a}_{i,3} &= \overline{\mathbf{D}}_{r_i} \mathbf{f}_{u_i}, && (\textit{Far zone}) \,, \end{aligned} \tag{7}$$

while $r_i = \lfloor (i-1)/n \rfloor + 1$, $u_i = ((i-1) \mod n) + 1$, for $i = 1, \cdots, m$ with $m = nL$. Let $\mathscr{A}_k : \mathscr{S}^{n \times n} \to \mathbb{R}^{m=nL}$, where $\mathscr{S}^{n \times n}$ is the space of self-adjoint matrices, be the linear mapping yielding the linear equalities

$$\mathscr{A}_k(\mathbf{W}) = \left[\mathbf{a}_{1,k}^H \mathbf{W} \mathbf{a}_{1,k}, \cdots, \mathbf{a}_{m,k}^H \mathbf{W} \mathbf{a}_{m,k} \right]^T. \tag{8}$$

Note that combining (5), (6), and (8), one can conclude that the phase retrieval problem from CDP acquired in the different zones can be modeled as

$$\mathbf{y}_k = \mathscr{A}_k(\mathbf{x}\mathbf{x}^H), \tag{9}$$

for $k = 1,2,3$.

Taking the linear operators $\mathscr{A}_k$ in (8) into account, we have that the scene $\mathbf{x}$ can be recovered from CDP if all $\mathscr{A}_k$ are injective Candes et al. (2015b). More precisely, this work follows the strategy in Gross et al. (2017) that considers

$$\mathscr{T}_{\mathbf{x}} = \left\{ \mathbf{x}\mathbf{w}^H + \mathbf{w}\mathbf{x}^H | \mathbf{w} \in \mathbb{C}^n \right\}, \tag{10}$$

which is the tangent space of the manifold of all rank-1 Hermitian matrices at the point $\mathbf{x}\mathbf{x}^H$. Thus, if the operators $\mathscr{A}_k$ for $k = 1,2,3$ satisfy the Condition 10.0.5, which is proved in Theorem 3.1.1, one can guarantee recovery Candes et al. (2015b). It is worth mentioning that a proof for Theorem 3.1.1 is needed for two reasons. First, the coding random variable d assumed in this thesis follows Definition 2.2.1 that differs from (3), which are the conditions imposed by previous theoretical works such as Candes et al. (2015b); Gross et al. (2017). Second, remark that at this point in the literature the PR problems from CDP associated to the near and middle zones have not been theoretically studied.

Assumption 1. *For any $\delta \in (0,1)$, and some constant $\beta > 0$ the linear operator $\mathscr{A}_k$ satisfies*

$$(1-\delta)\|\mathbf{W}\|_1 \leq \frac{1}{\beta}\|\mathscr{A}_k(\mathbf{W})\|_1 \leq (1+\delta)\|\mathbf{W}\|_1, \tag{11}$$

for all matrices $\mathbf{W} \in \mathscr{T}_{\mathbf{x}}$ and for all $k = 1,2,3$.

Theorem 3.1.1. Fix any $\delta \in (0,1)$ and the set of coded apertures $\{\mathbf{D}_\ell | \ell = 1,...,L\}$ with *i.i.d* copies

of an admissible random variable d according to Definition 2.2.1. If for some constant $0 < r \leq L$,

$$\mathbb{E}\left[\sum_{\ell=1}^{L} \overline{\mathbf{D}}_\ell \mathbf{D}_\ell\right] = r\mathbf{I}, \tag{12}$$

where $L \geq c_0 n$ for some sufficiently large constant $c_0 > 0$, with $\mathbf{I}$ as the identity matrix, then

$$\mathscr{P}\left(\frac{1}{rnL}\|\mathbf{A}_k\|_\infty^2 \leq 1+\delta\right) \leq 1 - 2e^{-cnL\varepsilon^2}, \tag{13}$$

for all $k = 1,2,3$ and some constant $c > 0$. Also, Condition 10.0.5 is satisfied with the same probability taking $\beta = rnL$. Each matrix $\mathbf{A}_k$ is given as in (4).

Demostración. See Appendix A (see Chapter 10). □

Theorem 3.1.1 proves that a scene is uniquely determined from CDP acquired at the three diffraction zones for all admissible variables if the set of coded apertures satisfies (12). We point out that the importance of (12) is twofold. First, from a theoretical point of view, (12) does not limit the admissible variable as previous works in the area Wang et al. (2018b); Jaganathan et al. (2015); Candes et al. (2015a); Bandeira et al. (2014); Kolte and Özgür (2016) following (3), instead, (12) establishes a condition for the coded apertures to be satisfied in order to uniquely determine (up to a global unimodular constant) the image of interest. Second, since condition (12) is independent of the diffraction zones, Theorem 3.1.1 also guarantees that satisfying (12) is enough to uniquely determine an image from CDP regardless the diffraction zone. In fact, the ability of coded apertures satisfying (12) to better estimate the image regardless the diffraction zone is numerically validated

in Chapter 9.

Finally, since in general it is not expected that the coded apertures naturally satisfy (12), this fact implies the need to design them. Furthermore, from the previous analysis of Theorem 3.1.1 it is clear that a set of coded apertures that closely satisfies (12) is able to outperforms non-designed ensembles because the ability of a recovery procedure to uniquely determine the image is directly affected by the particular structure of the coded apertures. Thus, in Section 3.2 also introduces a strategy to design the set of coded apertures in order to satisfy (12).

3.2. Coded Aperture Design

This section describes the design principles of coded apertures to better satisfy (12) for uniform admissible random variable d, i.e. all the coding elements have the same probability. Further, even considering the fact that this is a specific case, it is worth to remark that the proposed strategy considers more possible coding variables than previous approaches based on (3), as it will be discussed in Chapter 9. For ease of exposition, the strategy is presented considering an admissible random variable $d = \{e_1, e_2, e_3, e_4\}$ with probability $\{\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}\}$, respectively, assuming that $L = C_d b$ for some integer $b > 0$, where $C_d = 4$ is the number of coding elements. This work follows a greedy methodology to design the set of coded apertures where the following criteria are taken into account:

(a) Temporal correlation: Notice that, the theoretical condition in (12) can be easily satisfied if the set of coded apertures is constrained to be complementary Pinilla et al. (2018c). In practical terms, this is equivalent to having all the coding elements of d along the L-projections at each particular spatial position of the ensemble. This guarantees that each pixel of the image is mo-

dulated by all the coding elements of d; mathematically it means that the set $\{(\mathbf{C}^{\ell})_{s,u}|\ell = 1,\cdots,L\}$, contains b times each possible value of d for any $s,u \in \{1,\cdots,N\}$.

(b) Spatial separation: Remark that from (13), we can conclude that the set of coded apertures defines the concentration of measure of the largest eigenvalue of the sensing matrices $\mathbf{A}_k$ in (4) for $k = 1,2,3$. Recently, it has been shown that building a set of coded apertures with an equi-spaced distribution of the coding elements optimizes the concentration of measure of $\mathbf{A}_k$ and increases the image reconstruction quality Correa et al. (2016); Mejia and Arguello (2018). In mathematical terms Mejia and Arguello (2018) minimizes the upper bounds of the Gershgorin theorem of a given matrix, which in this case are $\mathbf{A}_k$ for $k = 1,2,3$. This minimization process leads to a better condition number of $\mathbf{A}_k$ in order to satisfy (13).

Considering these design criteria, a strategy to build a complementary set of coded apertures where all the coding elements of d are equi-spaced when d is any admissible random variable, is described next. More precisely, it is worth to mention that Pinilla et al. (2018c); Correa et al. (2016); Mejia and Arguello (2018) have only studied the case when d has just two coding elements, *e.g.* $d = \{0,1\}$. Then, this thesis extends these works allowing to build a set of coded apertures where all the coding elements of d are equi-spaced when d has more than two possible values. The design strategy consists of five steps as shown in Fig. 3.

Step 1. The coded aperture $\mathbf{C}^{\ell}$ is divided in cells of size $\gamma \times \gamma$ (green squares highlighted) where γ is chosen as $\gamma^2 \geq C_d$. For the particular example in Fig. 3, $C_d = 4$ and $\gamma = 2$ because the chosen admissible random variable d has four coding elements. Then, the positions of all e_i of d in the first cell of $\mathbf{C}^{\ell}$ are chosen uniformly at random as $(\mathbf{C}^{\ell})_{s,u} \sim \mathscr{U}[\{e_1,\cdots,e_4\} -$

$\{(\mathbf{C}^1)_{s,u},\cdots,(\mathbf{C}^{\ell-1})_{s,u}\}]$ for $s,u\in\{1,\cdots,\gamma\}$. Observe that the set subtraction

$$\{e_1,\cdots,e_4\}-\{(\mathbf{C}^1)_{s,u},\cdots,(\mathbf{C}^{\ell-1})_{s,u}\}$$

guarantees that one chooses a different coding element of d with respect to the first $\ell-1$ coded apertures.

Step 2. Move to the cell on the right to determine at random a proper position for each coding element. These positions are determined by maximizing the distance between pixels given

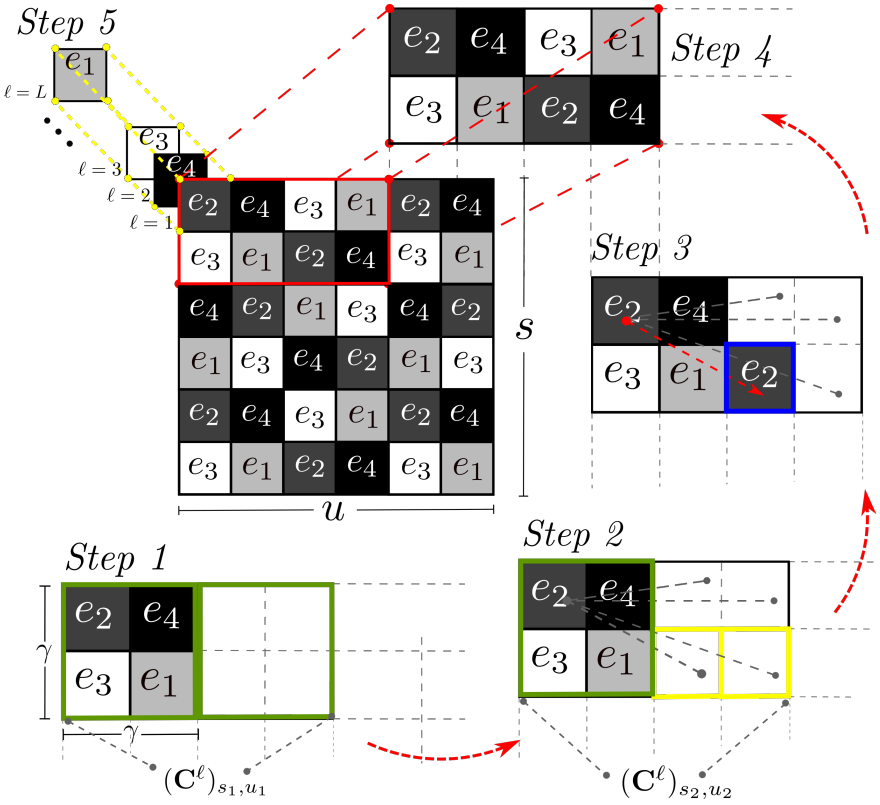

Figura 3. Coded aperture design strategy using the admissible random variables $d=\{e_1,e_2,e_3,e_4\}$ with probability $\{\frac{1}{4},\frac{1}{4},\frac{1}{4},\frac{1}{4}\}$, respectively and $C_d=4$.

by

$$P_{dist}((\mathbf{C}^{\ell})_{s_1,u_1},(\mathbf{C}^{\ell})_{s_2,u_2}) = min\{|s_1-s_2|,|u_1-u_2|\}, \tag{14}$$

where $(\mathbf{C}^{\ell})_{s_1,u_1}$ and $(\mathbf{C}^{\ell})_{s_2,u_2}$ are two pixels of the coded aperture at positions (s_1,u_1) in the first cell and (s_2,u_2) in the cell on the right, respectively. For instance, the positions maximizing this distance for the e_2 element is one of the two highlighted yellow squares.

Step 3. Let $\Omega_{e_i} = \{P_{dist}((\mathbf{C}^{\ell})_{s_1,u_1},(\mathbf{C}^{\ell})_{s_2,u_2})|(\mathbf{C}^{\ell})_{s_1,u_1} = e_i\}$ be the set of distances between the pixel that contains the coding element e_i in the first cell and the positions (s_2,u_2) of the next right cell. Then, the positions $\mathscr{R}_{e_i}$ that maximize (14) can be defined as $\mathscr{R}_{e_i} = \arg\max \Omega_{e_i}$. Define the set

$$\mathscr{B}_{e_i} = \{(s,u) \in \mathscr{R}_{e_i}|(\mathbf{C}^{l})_{s,u} = e_i, \text{ for } 1 \leq l \leq \ell-1\}.$$

From $\mathscr{F}_{e_i} = \mathscr{R}_{e_i} - \mathscr{B}_{e_i}$ it can be randomly determined the position of the coding element e_i in the next cell as $P_{e_i} \sim \mathscr{U}[\mathscr{F}_{e_i}]$. More precisely, the positions in $\mathscr{F}_{e_i}$ maximize (14), and guarantee choosing a different coding element than the first $\ell-1$ projections. For instance, for e_2 it would be the blue square highlighted in the next right cell.

Step 4. Steps 2 and 3 are repeated for all coding elements e_i of the admissible variable d. Thus, the spatial distribution of the ℓ-th coded aperture can be optimized.

Step 5. Note that the set of coded apertures is complementary until $\ell = C_d$. Then, if $\ell = C_d + 1$ a new coded aperture just optimizing the spatial distribution must be generated. Thus, from the $\ell = C_d + 2$ to $\ell = 2C_d$, the temporal and spatial correlation must be exploited considering the steps $1-4$, until the L-projections are completed.

In order to show an example of the outcome of the design procedure, Fig. 4 illustrates a designed coded aperture for $n = 16 \times 16$, $L = 4$ and $C_d = 4$ coding elements in a cell.

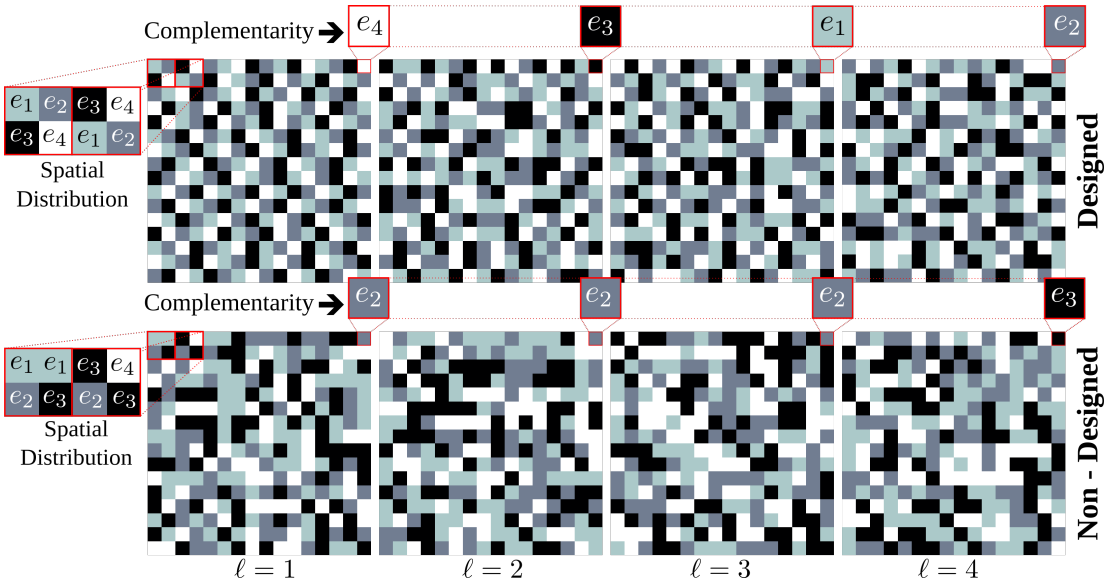

Figura 4. Comparison between a designed and non-designed coded apertures for $d = \{e_1, e_2, e_3, e_4\}$ with probability $\{\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}\}$, respectively.

Considering the above presented design strategy it is not intended to give the impression that the present thesis solves the problem of designing coded apertures for any admissible random variable. In fact, as mentioned before, this strategy is useful only for uniform admissible random variables. However, this particular scenario fulfills the theoretical purposes of this work to illustrate that instead of limiting the possible coding variables, as the current theory of the PR problem from CDP states, a design strategy for coded apertures is required. Specifically, Chapter 9 numerically validates that the design strategy enables phase retrieval using coding variables that cannot be considered by the current literature. Thus, the extension of the proposed design strategy for a

larger set of admissible random variables according to Definition 2.2.1 is relegated to future work.

4. Reconstruction Process

Considering that Chapter 3 established that a given scene $\mathbf{x}$ can be recovered from CDP acquired at the three diffraction zones with high probability, and taking the mathematical model in (9) into account, one can estimate the scene $\mathbf{x}$ by solving the following optimization problem

$$\underset{\mathbf{z}\in\mathbb{C}^n}{\text{minimize}} \quad f(\mathbf{z}) = \|\mathbf{y}_k - \mathscr{A}_k(\mathbf{z}\mathbf{z}^H)\|_2^2. \tag{15}$$

Observe that (15) is a non-convex problem. In fact, several algorithms have been developed to solve (15) such as WF Candès et al. (2015), TWF Chen and Candès (2015), TAF Wang et al. (2018a), PRSF Pinilla et al. (2018a), and RAF Wang et al. (2018b), among others. In fact, this thesis developed the PRSF method to solve the phase retrieval that will be detailed analyzed in Chapter 5. Further, these reconstruction algorithms require a proper initialization strategy to guarantee convergence. Thus, this work extends the orthogonality-promoting initialization introduced in Wang et al. (2018a), to CDP acquired at the three diffraction zones, as described in the following section.

4.1. Initialization procedure from CDP

According to Wang et al. (2018a), a motivating example that reveals the fundamental characteristics of high-dimensional random vectors for the three diffraction zones is presented. Fix any nonzero vector $\mathbf{x} \in \mathbb{C}^n$, and generate data as in (6) using the sampling vectors $\mathbf{a}_{i,k}$ as in (7). It is worth to remark that each vector $\mathbf{a}_{i,k}$ depends on the designed coded apertures. Thus, the following

squared normalized inner-product can be defined as

$$cos^2(\alpha_{i,k}) = \frac{|\langle \mathbf{a}_{i,k}, \mathbf{x} \rangle|^2}{\|\mathbf{a}_{i,k}\|^2 \|\mathbf{x}\|^2} \quad i = 1, \cdots, m, \tag{16}$$

where $\alpha_{i,k}$ is the angle between vectors $\mathbf{a}_{i,k}$ and $\mathbf{x}$. Consider ordering all $cos^2(\alpha_{i,k})$ in an ascending order, such that $cos^2(\alpha_{1,k}) \geq \cdots \geq cos^2(\alpha_{m,k})$.

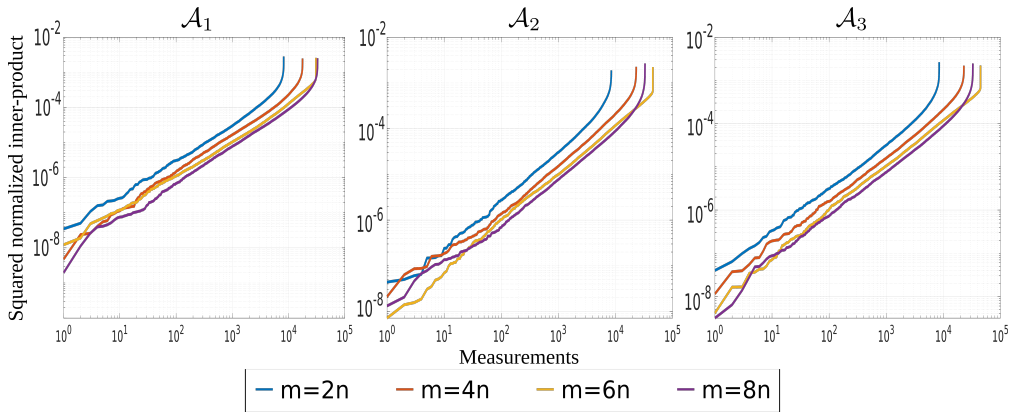

Figura 5. Ordered squared normalized inner-product for pairs $\mathbf{x}$ and $\mathbf{a}_i$, with m/n varying from 2 to 8, and $n = 64 \times 64$.

In Fig. 5 the squared normalized inner-products, varying m/n from 2 to 8, are illustrated for the three diffraction zones. Observe that all squared normalized inner-products are smaller than 10^{-2}, which implies that $\mathbf{x}$ is nearly orthogonal to a large number of $\mathbf{a}_{i,k}$ Wang et al. (2018a). Thus, in order to approximate $\mathbf{x}$ by a vector that is mostly orthogonal to a subset of vectors $\{\mathbf{a}_{i,k}\}$, an $\mathscr{I}_0$ index set with cardinality $card(\mathscr{I}_0) < m$, that includes indices of the smallest squared normalized inner-products $cos^2(\alpha_{i,k})$, where $i \in \mathscr{I}_0$, is introduced as follows. Define the set $\mathscr{I}_0 \subset \{1, \cdots, m\}$ as the collection of indices corresponding to the smallest values of $\{y_{i,k}/\|\mathbf{a}_{i,k}\|_2\}$. Thus, according

to Wang et al. (2018a) the OP initialization can be formulated as

$$\mathbf{z}_0 = \underset{\|\mathbf{w}\|_2=1}{\arg\min} \ \mathbf{w}^H \left(\frac{1}{card(\mathscr{I}_0)} \sum_{i \in \mathscr{I}_0} \frac{\mathbf{a}_{i,k}\mathbf{a}_{i,k}^H}{\|\mathbf{a}_{i,k}\|_2^2} \right) \mathbf{w}. \tag{17}$$

Notice that (17) implies finding the smallest eigenvalue, which calls for eigen-decomposition or matrix inversion, each typically requiring computational complexity $\mathscr{O}(n^3)$. However, we can avoid this step if we manipulate (17) as follows

$$\sum_{i \in \mathscr{I}_0} \frac{\mathbf{a}_{i,k}\mathbf{a}_{i,k}^H}{\|\mathbf{a}_{i,k}\|_2^2} = \sum_{i=1}^{nL} \frac{\mathbf{a}_{i,k}\mathbf{a}_{i,k}^H}{\|\mathbf{a}_{i,k}\|_2^2} - \sum_{i \in \mathscr{I}_0^c} \frac{\mathbf{a}_{i,k}\mathbf{a}_{i,k}^H}{\|\mathbf{a}_{i,k}\|_2^2}, \tag{18}$$

where $\mathscr{I}_0^c$ is the complement of $\mathscr{I}_0$. Further, in order to rewrite the term $\sum_{i=1}^{nL} \frac{\mathbf{a}_{i,k}\mathbf{a}_{i,k}^H}{\|\mathbf{a}_{i,k}\|_2^2}$ for each diffraction zone, we proceed by cases as follows.

- *Near zone:* First observe that $\|\mathbf{a}_{i,1}\|_2^2$ for this diffraction zone, according to (7), can be rewritten as

$$\begin{aligned} \|\mathbf{a}_{i,1}\|_2^2 &= \|\overline{\mathbf{D}}_{r_i}\mathbf{F}\overline{\mathbf{T}}\mathbf{f}_{u_i}\|_2^2 \\ &= \sum_{p=1}^{n} \left| (\overline{\mathbf{D}}_{r_i})_{p,p} \mathbf{f}_p^H \overline{\mathbf{T}}\mathbf{f}_{u_i} \right|^2 \\ &= \sum_{p=1}^{n} |(\mathbf{D}_{r_i})_{p,p}|^2 \left| \mathbf{f}_p^H \overline{\mathbf{T}}\mathbf{f}_{u_i} \right|^2 \approx \|\mathbf{D}_{r_i}\|_F^2, \end{aligned} \tag{19}$$

  where the third line comes from the fact that $\|\mathbf{f}_p\|_2 = 1$ for all $p \in \{1, \cdots, n\}$ (see (1)), and because $\mathbf{T}$ is an orthogonal diagonal matrix. Thus, considering (18) and (194) it can be

obtained that

$$\sum_{i=1}^{nL} \frac{\mathbf{a}_{i,1}\mathbf{a}_{i,1}^H}{\|\mathbf{a}_{i,1}\|_2^2} \approx \sum_{\ell=1}^{L}\sum_{p=1}^{n} \frac{\overline{\mathbf{D}}_\ell \mathbf{F}\overline{\mathbf{T}}\mathbf{f}_p\mathbf{f}_p^H\mathbf{T}\mathbf{F}^H\mathbf{D}_\ell}{\|\mathbf{D}_\ell\|_F^2}$$
$$\approx \sum_{\ell=1}^{L} \frac{\overline{\mathbf{D}}_\ell}{\|\mathbf{D}_\ell\|_F}\left(\sum_{p=1}^{n}\mathbf{F}\overline{\mathbf{T}}\mathbf{f}_p\mathbf{f}_p^H\mathbf{T}\mathbf{F}^H\right)\frac{\mathbf{D}_\ell}{\|\mathbf{D}_\ell\|_F} \approx \sum_{\ell=1}^{L}\frac{\overline{\mathbf{D}}_\ell\mathbf{D}_\ell}{\|\mathbf{D}_\ell\|_F^2}, \tag{20}$$

since $\mathbf{F}^H\mathbf{F} = \sum_{p=1}^{n}\mathbf{f}_p\mathbf{f}_p^H = \mathbf{I}$ and $\mathbf{T}$ is an orthogonal matrix.

- *Middle zone:* Notice that $\|\mathbf{a}_{i,2}\|_2^2$ for this diffraction zone, considering (7), can be rewritten as

$$\|\mathbf{a}_{i,2}\|_2^2 = \|\overline{\mathbf{D}}_{r_i}\overline{\mathbf{Q}\mathbf{f}}_{u_i}\|_2^2$$
$$= \sum_{p=1}^{n}\left|(\overline{\mathbf{D}}_{r_i})_{p,p}(\overline{\mathbf{Q}})_{p,p}(\bar{\mathbf{f}}_{u_i})_p\right|^2 = \frac{1}{n}\sum_{p=1}^{n}|(\mathbf{D}_{r_i})_{p,p}|^2 = \frac{1}{n}\|\mathbf{D}_{r_i}\|_F^2, \tag{21}$$

where the second equality comes from the fact that $\mathbf{Q}$ is diagonal orthogonal matrix, and that $|(\bar{\mathbf{f}}_{u_i})_p| = 1/\sqrt{n}$ according to (1). Thus, from (21) we have that

$$\sum_{i=1}^{nL}\frac{\mathbf{a}_{i,2}\mathbf{a}_{i,2}^H}{\|\mathbf{a}_{i,2}\|_2^2} = n\sum_{\ell=1}^{L}\sum_{p=1}^{n}\frac{\overline{\mathbf{D}}_\ell\overline{\mathbf{Q}\mathbf{f}}_p\mathbf{f}_p^T\mathbf{Q}\mathbf{D}_\ell}{\|\mathbf{D}_{r_i}\|_F^2}$$
$$= n\sum_{\ell=1}^{L}\frac{\overline{\mathbf{D}}_\ell}{\|\mathbf{D}_\ell\|_F}\left(\sum_{p=1}^{n}\overline{\mathbf{Q}\mathbf{f}}_p\mathbf{f}_p^H\mathbf{Q}\right)\frac{\mathbf{D}_\ell}{\|\mathbf{D}_\ell\|_F} = n\sum_{\ell=1}^{L}\frac{\overline{\mathbf{D}}_\ell\mathbf{D}_\ell}{\|\mathbf{D}_\ell\|_F^2}, \tag{22}$$

where the third equality comes from the fact that $\mathbf{Q}$ and $\mathbf{F}$ are orthogonal matrices.

- *Far zone:* Finally, observe that for the far zone it can be obtained

$$\|\mathbf{a}_{i,3}\|_2^2 = \|\overline{\mathbf{D}}_{r_i}\mathbf{f}_{u_i}\|_2^2 = \sum_{p=1}^{n} \left|(\overline{\mathbf{D}}_{r_i})_{p,p}(\mathbf{f}_{u_i})_p\right|^2 = \frac{1}{n}\|\mathbf{D}_{r_i}\|_F^2, \tag{23}$$

because $|(\mathbf{f}_{u_i})_p| = 1/\sqrt{n}$. Thus, from (23) we have that

$$\begin{aligned}\sum_{i=1}^{nL} \frac{\mathbf{a}_{i,3}\mathbf{a}_{i,3}^H}{\|\mathbf{a}_{i,3}\|_2^2} &= n\sum_{\ell=1}^{L}\sum_{p=1}^{n} \frac{\overline{\mathbf{D}}_\ell \mathbf{f}_p \mathbf{f}_p^H \mathbf{D}_\ell}{\|\mathbf{D}_\ell\|_F^2} \\ &= n\sum_{\ell=1}^{L} \frac{\overline{\mathbf{D}}_\ell}{\|\mathbf{D}_\ell\|_F}\left(\sum_{p=1}^{n}\mathbf{f}_p\mathbf{f}_p^H\right)\frac{\mathbf{D}_\ell}{\|\mathbf{D}_\ell\|_F} = n\sum_{\ell=1}^{L}\frac{\overline{\mathbf{D}}_\ell\mathbf{D}_\ell}{\|\mathbf{D}_\ell\|_F}.\end{aligned} \tag{24}$$

Now, if the set of coded apertures satisfies (12), then from (20), (22), and (24), it can be concluded that $\sum_{i=1}^{nL}\frac{\mathbf{a}_{i,k}\mathbf{a}_{i,k}^H}{\|\mathbf{a}_{i,k}\|_2^2} \approx c_k\mathbf{I}$ for some constants $c_k > 0$ with $k = 1,2,3$. Considering this observation, (17) can be approximated as

$$\mathbf{z}_0 = \underset{\|\mathbf{w}\|_2=1}{\arg\max}\ \mathbf{w}^H\left(\frac{1}{card(\mathscr{I}_0^c)}\sum_{i\in\mathscr{I}_0^c}\frac{\mathbf{a}_{i,k}\mathbf{a}_{i,k}^H}{\|\mathbf{a}_{i,k}\|_2^2}\right)\mathbf{w}, \tag{25}$$

which meets the numerical formulation in Wang et al. (2018a) for the OP initialization. Remark that the imposed condition over the set of coded apertures meets the theoretical recovery conditions of Theorem 3.1.1. This is an important result since for the first time the initialization procedure, required to solve (15), meets the recovery conditions of the PR problem. Therefore, Theorem 8.1.2 theoretically establishes that (116) can approximate the scene of interest with high probability.

Theorem 4.1.1. Consider (noise-free) measurements $\mathbf{y}_k$ as defined in (9). Then, with probability

of at least $1-2e^{-Cn}$ for some constant $C>0$, the vector $\mathbf{z}_0$ returned by (116) satisfies that

$$dist(\mathbf{z}_0,\mathbf{x}) \leq \rho\|\mathbf{x}\|_2, \tag{26}$$

for some constant $\rho \in (0,1)$, provided that L is sufficiently large.

Demostración. See Appendix B (see Chapter 10). □

To solve (15), based on extended OP initialization, Algorithm 1 is presented. Line 2 estimates the initial guess $\mathbf{z}_0$ in (116), which is the vector $\tilde{\mathbf{z}}_0$ scaled so that its norm matches approximately that of $\mathbf{x}$ based on the strong law of large numbers, where $\left(\frac{1}{r}\sum_{i=1}^{m} y_{i,k}\right)^{1/2} \approx \|\mathbf{x}\|_2$, assuming that the coded apertures are properly designed, based on (12). Further, the solution to (116) can be well approximated with a few power iterations at a much cheaper computational complexity $\mathscr{O}(n\ card(\mathscr{I}_0^c))$ than $\mathscr{O}(n^3)$ required for solving (17) Wang et al. (2018b), since it involves the estimation of the leading eigenvector of matrix $\mathbf{Y}_0$. Moreover, in Line 3 any reconstruction methods proposed in the literature to solve (15) can be used, which refines the initial guess solution from Line 2.

Algorithm 1

1: **Input:** Acquired data $\{(\mathbf{a}_{i,k}; y_{i,k})\}_{i=1}^{m}$ for $k = 1, 2, 3$.

2: Initial point $\mathbf{z}_0 = \left(\frac{1}{r}\sum_{i=1}^{m} y_{i,k}\right)^{1/2} \tilde{\mathbf{z}}_0$, where $\tilde{\mathbf{z}}_0$ is the leading eigenvector of

$$\mathbf{Y}_0 := \frac{1}{card(\mathscr{I}_0^c)} \sum_{i \in \mathscr{I}_0^c} \frac{\mathbf{a}_{i,k}\mathbf{a}_{i,k}^H}{\|\mathbf{a}_{i,k}\|_2^2}$$

and the set $\mathscr{I}_0^c$ includes the indices of the $\lfloor \frac{m}{2} \rfloor$ largest values of $y_{i,k}/\|\mathbf{a}_{i,k}\|_2$.

3: $\mathbf{z} \leftarrow$ **reconstruction-algorithm**$(\mathbf{z_0})$.

4: **Output:** $\mathbf{z}$

In summary, considering Theorem 8.1.2, it is important to remark that the extended initialization procedure is more general since, it does not limit the possible random coding variables; moreover, it is able to initialize the PR problem for CDP acquired at the three diffraction zones, and third, the numerical formulation in (116) links the theoretical assumption in (12) with the recovery guarantees from CDP.

4.2. Sparsity Assumptions

Until now it have been analytically studied the role of the coded aperture to recover an image $\mathbf{x}$ from CDP regardless the diffraction zone without any assumption over $\mathbf{x}$. This theoretical analysis has provided a coded aperture design strategy that pursuits to satisfy (12) to uniquely identify an image from CDP. However, there are optical applications such as astronomy Fienup and Dainty (1987) and microscopy Stephens and Allan (2003) where the scene $\mathbf{x}$ is naturally sparse in some

representation basis. Up to date, the PR problem from CDP assuming prior information on $\mathbf{x}$ has not been theoretically studied. Thus, this section provides some theoretical insights to successfully recover a scene $\mathbf{x}$ from CDP, which is assumed to be sparse on some basis.

Consider that $\mathbf{x} \in \mathbb{C}^n$ can be sparse in an orthogonal domain Ψ (*i.e* $\Psi^H\Psi = \mathbf{I}$) with a s-sparse representation $\theta \in \mathbb{C}^n$, where s is the sparsity level and $s \ll n$. In particular, in optical imaging, $\mathbf{x}$ can be sparsely represented in some domains such as Wavelet or Discrete Cosine Transform (DCT), in the sense that $\|\theta\|_0 \ll n$, where $\mathbf{x} = \Psi^H\theta$, and $\|\cdot\|_0$ is the ℓ_0 pseudo-norm. Thus, the acquisition process of CDP in (6) can be rewritten for each k-th diffraction zone under sparsity assumption as follows

$$y_{i,k} = |\mathbf{a}_{i,k}^H\mathbf{x}|^2 = |\mathbf{a}_{i,k}^H\Psi^H\theta|^2 = |\mathbf{b}_{i,k}^H\theta|^2, \tag{27}$$

where $\mathbf{b}_{i,k} = \Psi\mathbf{a}_{i,k}$ are the sparsity-based sampling vectors and the matrix which concatenates each sparsity-based sampling vector can be defined as $\mathbf{B}_k = \left[\mathbf{b}_{1,k}, \cdots, \mathbf{b}_{m,k}\right]^H$. Notice that the scene under sparsity assumptions meets required conditions of Theorem 3.1.1 from the measurements in (27), considering (12) and knowing that Ψ is orthogonal.

According to the measurements modeled in (27), the following optimization problem to recover θ can be formulated

$$\begin{aligned} \underset{\theta\in\mathbb{C}^n}{\text{minimize}} \quad & h(\theta) = \frac{1}{m}\sum_{i=1}^{m}\left(\sqrt{y_{i,k}} - |\mathbf{b}_{i,k}^H\theta|\right)^2, \\ \text{subject to} \quad & \|\theta\|_0 \leq s. \end{aligned} \tag{28}$$

Algorithm 2

1: **Input:** Data $\{(\mathbf{b}_{i,k}; y_{i,k})\}_{i=1}^{m}$, for $k = 1,2,3$, and the sparsity level s.
2: **Let** $\mathscr{S}_0$ be the set of s largest indices of $\{\frac{1}{m}\sum_{q=1}^{m} y_{q,k} b_{q,p}^{k}\}_{1\leq p\leq n}$.
3: **Set:** $\tilde{\mathbf{b}}_{i,k} = (\mathbf{b}_{i,k})_{\mathscr{S}_0}$
4: Initial point $\theta_0 = \left(\frac{1}{r}\sum_{i=1}^{m} y_{i,k}\right)^{1/2} \tilde{\theta}_0$, where $\tilde{\theta}_0$ is the leading eigenvector of

$$\mathbf{Y}_0 := \frac{1}{card(\mathscr{I}_0^c)} \sum_{i\in\mathscr{I}_0^c} \frac{\tilde{\mathbf{b}}_{i,k}\tilde{\mathbf{b}}_{i,k}^H}{\|\tilde{\mathbf{b}}_{i,k}\|_2^2}$$

and the set $\mathscr{I}_0^c$ includes the indices of the $\lfloor \frac{m}{2} \rfloor$ largest values of $y_{i,k}/\|\mathbf{a}_{i,k}\|_2$.

5: $\theta \leftarrow$ **reconstruction-algorithm**(θ_0).
6: $\mathbf{z} \leftarrow \Psi^H \theta$
7: **Output:** $\mathbf{z}$

The method to solve (28), based on sparsity assumption, is summarized in Algorithm 2. This procedure consists of three stages. First, in Lines 2-3, the non-zero coefficients are estimated, procedure that will be explained in detail in the following section; second, in Line 4 the initial guess of θ is estimated, which is then refined by some state-of-the-art sparse PR reconstruction methods.

4.3. Non-zero Coefficients Estimation

In order to obtain the estimation of the non-zero coefficients of θ the strategy developed in Yuan et al. (2017) is extended to CDP. Also, it is theoretically characterized the admissible random variables that attain the best performance estimating the support of θ.

Define $Z_{q,p}^{k} := y_{q,k}|b_{q,p}^{k}|^2$, $1 \leq q \leq m$, $1 \leq p \leq n$, where $b_{q,p}^{k}$ is the element at row q and column p of the matrix $\mathbf{B}_k$ for $k = 1,2,3$. Then, according to the random variable $Z_{p,q}^{k}$, it can be

obtained that

$$\mathbb{E}[Z_{q,p}^k] \geq c_1 \|\mathbf{x}\|_2^2 + c_2 |(\theta)_p|^2 + c_3, \tag{29}$$

where c_1, c_2 and c_3 are constants. Also, given the fact that θ is sparse, then $(\theta)_p \neq 0$ or $(\theta)_p = 0$. It is clear that as long as the constant c_2 is sufficiently large, the non-zero coefficients of θ can be recovered exactly in this way. Specifically, the following lemma theoretically proves (114) and establishes that an admissible random variable d satisfying $\mathbb{E}[d] \neq 0$ attains a better performance estimating the support of θ.

Lemma 4.3.1. An admissible random variable d satisfying $\mathbb{E}[d] \neq 0$ attains a better performance estimating the non-zero coefficients of θ.

Demostración. See Appendix C (see Chapter 10). □

It is important noticing that Lemma 8.1.1 provides a valuable theoretical observation related with the type of coding elements that improve the reconstruction performance when the image is sparse in some basis, showing the crucial role that plays the coded aperture to uniquely identify **x** from CDP.

5. Smoothing Gradient Algorithm

This thesis also develops the Phase Retrieval method via Smoothing function (PRSF) which uses an auxiliary differentiable function to retrieve the signal. PRSF is based on the smoothing projected gradient method which is useful for non-convex optimization problems. PRSF uses a nonlinear conjugate gradient of the smoothing function as the search direction to accelerate the convergence, as it will be explained in this chapter.

5.1. Smooth Optimization Problem

Consider the system of m quadratic equations of the form

$$y_k = |\langle \mathbf{a}_k, \mathbf{x} \rangle|^2, k = 1, \cdots, m, \tag{30}$$

where the data vector $\mathbf{y} := [y_1, \cdots, y_m]^T \in \mathbb{R}^m$ represents the measurements, $\mathbf{a}_k \in \mathbb{R}^n/\mathbb{C}^n$ are the known sampling vectors and $\mathbf{x} \in \mathbb{R}^n/\mathbb{C}^n$ is the desired unknown signal. For ease of exposition consider the complex-valued Gaussian design vectors as $\mathbf{a}_k \sim \mathscr{CN}(0, \mathbf{I}_n) = \mathscr{N}(0, \frac{1}{2}\mathbf{I}_n) + j\mathscr{N}(0, \frac{1}{2}\mathbf{I}_n)$, assumed to be independently and identically distributed (i.i.d.), where $j = \sqrt{-1}$. For the real Gaussian case the sampling vectors $\mathbf{a}_k$ are given by $\mathbf{a}_k \sim \mathscr{N}(0, \mathbf{I}_n)$, also assumed to be i.i.d.. Then, adopting the least-squares criterion, the task of recovering a solution from the phaseless measurements in (30) reduces to that of minimizing the amplitude-based loss function

$$\min_{\mathbf{x} \in \mathbb{C}^n} f(\mathbf{x}) = \frac{1}{m} \sum_{k=1}^{m} \left(f_k(\mathbf{x}) - q_k \right)^2, \tag{31}$$

where $f_k(\mathbf{x}) = |\langle \mathbf{a}_k, \mathbf{x} \rangle|$ and $q_k = \sqrt{y_k}$. Notice that the optimization problem in (31) is non-smooth and non-convex Candès and Li (2014). Then, this thesis proposes an algorithm based on an auxiliary smoothing function $g(\cdot)$ to approximate the original objective function, in order to solve the non-smooth and non-convex optimization problem. To do that, some conditions over the auxiliary function $g(\cdot)$ are required, that will be discussed in brief. Specifically, it is necessary to prove that the objective function $f(\mathbf{x})$ in (31) is locally Lipschitz continuous.

Definition 5.1.1. *Lipschitz continuous under dist*$(\cdot,\cdot)$: Let $f : (\mathbb{C}^n, dist(\cdot,\cdot)) \to \mathbb{R}$ be a function. The function f is called Lipschitz continuous if there exists a constant $L > 0$ such that, for all $\mathbf{w}_1, \mathbf{w}_2 \in \mathbb{C}^n$

$$|f(\mathbf{w}_1) - f(\mathbf{w}_2)| \leq L\, dist(\mathbf{w}_1, \mathbf{w}_2). \tag{32}$$

Definition 5.1.2. *Locally Lipschitz continuous under dist*$(\cdot,\cdot)$: Let $f : (\mathbb{C}^n, dist(\cdot,\cdot)) \to \mathbb{R}$ be a function. The function $f(\cdot)$ is called Locally Lipschitz continuous if for every $\mathbf{w} \in \mathbb{C}^n$ exists a neighborhood $\mathscr{U}$, such that, $f(\cdot)$ restricted to $\mathscr{U}$ is Lipschitz continuous.

The following lemma shows that $f(\mathbf{x})$ in (31) is locally Lipschitz according to Definition 5.1.2.

Lemma 5.1.1. The function $f(\mathbf{x})$ in (31) is locally Lipschitz continuous under the distance $dist(\cdot,\cdot)$.

Demostración. To prove the lemma, it is shown that for all $k \in \{1, \cdots, m\}$ the functions $f_k(\cdot)$ in (31) are Lipschitz continuous. Let $\mathbf{w}_1, \mathbf{w}_2 \in \mathbb{C}^n$ be two different vectors such that

$$|f_k(\mathbf{w}_1) - f_k(\mathbf{w}_2)| = ||\langle \mathbf{a}_k, \mathbf{w}_1\rangle| - |\langle \mathbf{a}_k, \mathbf{w}_2\rangle||. \tag{33}$$

By using the triangle inequality on the right hand side term of (33), one can write

$$||\langle \mathbf{a}_k, \mathbf{w}_1\rangle| - |\langle \mathbf{a}_k, \mathbf{w}_2\rangle|| \leq |\langle e^{-j\theta}\mathbf{w}_1, \mathbf{a}_k\rangle - \langle \mathbf{w}_2, \mathbf{a}_k\rangle|, \tag{34}$$

for any $\theta \in [0, 2\pi)$. Using the fact that $\langle \mathbf{w}, \mathbf{a}_k \rangle = \mathbf{a}_k^H \mathbf{w}$ and Eqs. (33), (34) yields

$$|f_k(\mathbf{w}_1) - f_k(\mathbf{w}_2)| \leq |e^{-j\theta}\left(\mathbf{a}_k^H \mathbf{w}_1\right) - \left(\mathbf{a}_k^H \mathbf{w}_2\right)| = |\mathbf{a}_k^H\left(e^{-j\theta}\mathbf{w}_1 - \mathbf{w}_2\right)|. \tag{35}$$

By definition $\mathbf{a}_k^H \mathbf{w} = \sum_{i=1}^n (\overline{\mathbf{a}}_k)_i (\mathbf{w})_i$, where $(\overline{\mathbf{a}}_k)_i$ is the i-th conjugate component of $\mathbf{a}_k$ and, $(\mathbf{w})_i$ is the i-th element of $\mathbf{w}$. Then, using the triangle inequality, (35) can be rewritten as

$$\begin{aligned} |f_k(\mathbf{w}_1) - f_k(\mathbf{w}_2)| &\leq |\textstyle\sum_{i=1}^n (\overline{\mathbf{a}}_k)_i \left(e^{-j\theta}\mathbf{w}_1 - \mathbf{w}_2\right)_i| \leq \textstyle\sum_{i=1}^n |(\mathbf{a}_k)_i||\left(e^{-j\theta}\mathbf{w}_1 - \mathbf{w}_2\right)_i| \\ &\leq a_{max}^k \textstyle\sum_{i=1}^n |\left(e^{-j\theta}\mathbf{w}_1 - \mathbf{w}_2\right)_i| \leq a_{max}^k \|e^{-j\theta}\mathbf{w}_1 - \mathbf{w}_2\|_1, \end{aligned} \tag{36}$$

where $a_{max}^k = \text{máx}\{|(\mathbf{a}_k)_i| : i = 1, \cdots, n\}$ and $\|\cdot\|_1$ is the ℓ_1 norm. Since ℓ_1 and ℓ_2 are equivalent norms, there exists a constant $\rho \in \mathbb{R}_{++}$ such that $\|\mathbf{w}\|_1 \leq \rho \|\mathbf{w}\|_2$ for all $\mathbf{w} \in \mathbb{R}^n / \mathbb{C}^n$ Candès and Wakin (2008). Thus, (36) becomes

$$|f_k(\mathbf{w}_1) - f_k(\mathbf{w}_2)| \leq a_{max}^k \|e^{-j\theta}\mathbf{w}_1 - \mathbf{w}_2\|_1 \leq \left(a_{max}^k \rho\right) \|e^{-j\theta}\mathbf{w}_1 - \mathbf{w}_2\|_2. \tag{37}$$

Notice that, for the $i.i.d.$ Gaussian vectors $\mathbf{a}_k$, $a_{max}^k = \|\mathbf{a}_k\|_\infty \leq \sqrt{2.3n}$, holds with probability at least $1 - me^{-n/2}$ Wang et al. (2016a). Taking the value of θ that minimizes the term $\|e^{-j\theta}\mathbf{w}_1 - \mathbf{w}_2\|_2$, (37) becomes

$$|f_k(\mathbf{w}_1) - f_k(\mathbf{w}_2)| \leq \left(\sqrt{2.3n}\rho\right) dist(\mathbf{w}_1, \mathbf{w}_2). \tag{38}$$

Therefore, it can be concluded that each $f_k(\cdot)$ is a Lipschitz continuous function with constant

$L_k = \sqrt{2.3n}\rho$, with probability at least $1 - me^{-n/2}$. Further, the function $\frac{1}{\sqrt{m}}(f_k(\mathbf{x}) - q_k)$ in (31) is also Lipschitz continuous with constant $\sqrt{\frac{2.3n}{m}}\rho$, with probability exceeding $1 - me^{-n/2}$, because the term q_k can be considered constant Eriksson et al. (2013).

On the other hand, take any $\mathbf{w} \in \mathbb{C}^n$ and define $\mathscr{U} = \{\mathbf{z} \in \mathbb{C}^n : dist(\mathbf{z}, \mathbf{w}) < \varepsilon\}$ for $\varepsilon > 0$. Note that $\mathscr{U}$ is the neighborhood of $\mathbf{w}$ and also $\mathscr{U}$ is a bounded set because $\|\mathbf{z}\|_2 \leq \|\mathbf{w}\|_2 + \varepsilon < \infty$, for all $\mathbf{z} \in \mathscr{U}$. Thus, given the fact that $\mathscr{U}$ is a bounded set and each function $\frac{1}{\sqrt{m}}(f_k(\mathbf{x}) - q_k)$ is Lipschitz continuous, then $\frac{1}{m}(f_k(\mathbf{x}) - q_k)^2$ restricted to the set $\mathscr{U}$ is a Lipschitz continuous function Eriksson et al. (2013) with probability at least $1 - me^{-n/2}$. Hence, since $f(\mathbf{x})$ defined in (31) is a sum of Lipschitz continuous functions in the set $\mathscr{U}$, then $f(\mathbf{x})$ is a Lipschitz continuous function in $\mathscr{U}$. Thus, it can be concluded that $f(\mathbf{x})$ is locally Lipschitz continuous according to Definition 5.1.2, with probability at least $1 - me^{-n/2}$. □

The concept of smoothing function was presented in Zhang and Chen (2009) as the following definition:

Definition 5.1.3. *Smoothing function:* Let $f : \mathbb{C}^n \to \mathbb{R}$ be a locally Lipschitz continuous function. Then $g : \mathbb{C}^n \times \mathbb{R}_+ \to \mathbb{R}$ is a smoothing function of $f(\cdot)$, if $g(\cdot, \mu)$ is smooth in $\mathbb{C}^n$ for any fixed $\mu \in \mathbb{R}_{++}$ and

$$\lim_{\mu \downarrow 0} g(\mathbf{w}, \mu) = f(\mathbf{w}), \tag{39}$$

for any fixed $\mathbf{w} \in \mathbb{C}^n$.

According to the above definition, consider the function $\varphi_\mu : \mathbb{R} \to \mathbb{R}_{++}$ defined as

$$\varphi_\mu(w) = \sqrt{w^2 + \mu^2}, \tag{40}$$

where $\mu \in \mathbb{R}_{++}$. The following lemma shows that $\varphi_\mu(\cdot)$ has important smooth properties to approximate the functions $f_k(\cdot)$, given that $\varphi_0(|\mathbf{a}_k^H \mathbf{x}|) = f_k(\mathbf{x})$.

Lemma 5.1.2. The function $\varphi_\mu(w)$, defined in (40), has the following properties.

1. $\varphi_\mu(w)$ is Lipschitz continuous function.

2. $\varphi_\mu(w)$ converges uniformly to $\varphi_0(w)$ on $\mathbb{R}$.

Demostración. 1. Since $\mu > 0$ then $\varphi_\mu(w)$ is smooth on $\mathbb{R}$, where $\varphi'_\mu(w)$ is given by

$$\varphi'_\mu(w) = \frac{w}{\sqrt{w^2 + \mu^2}}. \tag{41}$$

Notice that $\sqrt{w^2 + \mu^2} \geq w$ for all $w \in \mathbb{R}$, then $|\varphi'_\mu(w)| \leq 1$. Therefore, $\varphi_\mu(w)$ is a Lipschitz continuous function because its first derivative is bounded Eriksson et al. (2013). Further, the Lipschitz constant for the function $\varphi_\mu(\cdot)$ is $L_{\varphi_\mu} = 1$.

2. According to the definition of the function φ_μ in (40), it can be obtained that

$$|\varphi_\mu(w) - \varphi_0(w)| = |\sqrt{w^2 + \mu^2} - \sqrt{w^2}|. \tag{42}$$

Note that by the Minkowski inequality Kreyszig (1989), it can be concluded that $\sqrt{w^2+\mu^2} \leq \sqrt{w^2}+\mu$, therefore

$$|\varphi_\mu(w)-\varphi_0(w)| \leq |\sqrt{w^2}+\mu-\sqrt{w^2}| \leq \mu. \tag{43}$$

□

The first result in Lemma 5.1.2 is used to guarantee the convergence of the proposed algorithm in Subsection 5.2.1. Also, the second part of Lemma 5.1.2 establishes that the function $\varphi_\mu(|\mathbf{a}_k^H\mathbf{x}|)$ uniformly approximates $f_k(\mathbf{x})$, which is a desirable convergence, since it only depends on the value of μ. Therefore, a smooth optimization problem to recover the unknown desired signal $\mathbf{x} \in \mathbb{R}^n/\mathbb{C}^n$ from the measurements q_k in (31) can be formulated as

$$\min_{\mathbf{x}\in\mathbb{R}^n/\mathbb{C}^n} g(\mathbf{x},\mu) = \frac{1}{m}\sum_{k=1}^{m}\left(\varphi_\mu(|\mathbf{a}_k^H\mathbf{x}|)-q_k\right)^2, \tag{44}$$

where $g(\mathbf{x},\mu)$ is the smoothing function of $f(\mathbf{x})$. Theorem 5.1.3 shows that the function $g(\cdot)$ is a uniformly smooth approximation of the function $f(\cdot)$, which is a desired behavior in order to solve the optimization problem in (31).

Theorem 5.1.3. Let f and $g(\cdot,\mu)$ be defined as in (31) and (44), respectively. Then $g(\cdot,\mu)$ is smooth for any fixed $\mu > 0$, and there exists a constant $\kappa_1 > 0$ satisfying

$$|g(\mathbf{x},\mu)-f(\mathbf{x})| \leq \mu\kappa_1. \tag{45}$$

Demostración. See Appendix D (see Chapter 10). □

5.2. Gradient Update Step

This section introduces Algorithm 7 that summarizes the proposed method to recover the phase. This algorithm is a descent gradient method. Following the algorithm in each iteration, a backtracking line search strategy is used to choose a correct step size of the conjugate gradient update direction, which is calculated in Line 9. Further, the smoothing parameter μ is updated as in Zhang and Chen (2009), to obtain a new point. That is, if $\|\partial g(\mathbf{x}_{i+1}, \mu_i)\|_2 \geq \gamma\mu_i$ in Line 10 is not satisfied, then the smoothing parameter is updated using the new point in Line 13. Algorithm 7 calculates the conjugate direction in Line 18. Each vector $\tilde{\mathbf{g}}_i$ in Algorithm 7 is calculated using the Wirtinger derivative as was introduced in Hunger (2007). The following lemma introduces the Wirtinger derivative of the function $g(\mathbf{x}, \mu)$.

Lemma 5.2.1. The Wirtinger derivative of a real-valued function $h(\mathbf{z}) : \mathbb{C}^n \to \mathbb{R}$ with complex-valued argument $\mathbf{z} \in \mathbb{C}^n$ is obtained for

$$2\frac{\partial h(\mathbf{z})}{\partial \mathbf{z}^*} \triangleq 2\left[\frac{\partial h(\mathbf{z})}{\partial z_1^*}, \cdots, \frac{\partial h(\mathbf{z})}{\partial z_n^*}\right]^T, \tag{46}$$

where the variable z_i^* is the conjugate version of z_i. The proof of this lemma can be found in Corollary 5.0.1 in Hunger (2007). It is important to remark that this Wirtinger derivation has been recently used by the state-of-the-art methods to solve the phase retrieval problem Candès et al. (2015); Wang et al. (2016a); Chen and Candès (2015).

For simplicity, we denote the Wirtinger derivative of any function $h(\mathbf{z})$ as $\partial h(\mathbf{z})$, *i.e.* $\partial h(\mathbf{z}) =$

$2\frac{\partial h(\mathbf{z})}{\partial \mathbf{z}^*}$. Then, considering the result in Lemma 5.2.1, the Wirtinger derivative of $g(\mathbf{x},\mu)$ is given by

$$\partial g(\mathbf{x}_i,\mu_i) = \frac{2}{m}\sum_{k=1}^{m}(\varphi_{\mu_i}(|\mathbf{a}_k^H\mathbf{x}_i|) - q_k)\partial\varphi_{\mu_i}(|\mathbf{a}_k^H\mathbf{x}_i|), \tag{47}$$

where

$$\partial\varphi_{\mu_i}(|\mathbf{a}_k^H\mathbf{x}_i|) = \frac{\mathbf{a}_k^H\mathbf{x}_i}{\varphi_{\mu_i}(|\mathbf{a}_k^H\mathbf{x}_i|)}\mathbf{a}_k. \tag{48}$$

Notice that, in contrast to the gradient update steps for the TAF and TWF methods introduced in Wang et al. (2017a) and Chen and Candès (2015) respectively, $\partial g(\mathbf{x}_i,\mu_i)$ in (95) is always continuous because $\varphi_\mu(|\mathbf{a}_k^H\mathbf{w}|) \neq 0$ for any $\mathbf{w} \in \mathbb{C}^n$. Therefore, the proposed PRSF method does not require truncation parameters.

5.2.1. Gradient Consistency Property and Converge Conditions. This section presents the convergence of Algorithm 7, by proving that any stationary point $\mathbf{x}^*$ of the sequence $\{\mathbf{x}_i\}$ is a Clarke stationary point, that is, $\mathbf{0} \in \partial^c f(\mathbf{x}^*)$ Bagirov et al. (2014). To do that, we first introduce the Clarke subdifferential definition. Specifically, the Clarke subdifferential of a locally Lipschitz continuous function $h : \mathbb{C}^n \to \mathbb{R}$, at point $\mathbf{x}$, denoted as $\partial^c h(\mathbf{x})$, is defined as

$$\partial^c h(\mathbf{x}) = \text{conv}\left\{\lim_{k\to\infty}\nabla h(\mathbf{x}_k) : \mathbf{x}_k \to \mathbf{x}, \mathbf{x}_k \notin D_h\right\}, \tag{49}$$

where “conv"denotes the convex hull of a set and D_h is the set of points at which h fails to be smooth Clarke (1990). Considering, the Clarke subdifferential definition, the *Gradient consistency property* is introduced as follows, to prove the convergence of Algorithm 7.

Algorithm 3

Input: Data $\{(\mathbf{a}_k; q_k)\}_{k=1}^m$ and $\varepsilon_0 = 10^{-10}$. Choose constants $\delta_1 = 0.9, \delta_2 = 0.4, \gamma_1 = 0.5, \mu_0 = 5 \times 10^4/m, \gamma = 0.01$ and T maximum number of iterations.

Initial point $\mathbf{x}_0 = \sqrt{\frac{\sum_{k=1}^m q_k^2}{m}}\tilde{\mathbf{z}}_0$. where $\tilde{\mathbf{z}}_0$ is the leading eigenvector of $\mathbf{Y}_0 := \frac{1}{|I_0|}\sum_{k\in I_0}\frac{\mathbf{a}_k\mathbf{a}_k^H}{\|\mathbf{a}_k\|_2^2}$.

Set $\mathbf{d}_0 = -\tilde{\mathbf{g}_0} = -\partial g(\mathbf{x}_0, \mu_0)$.

for $i = 0 : T-1$ **do**

Compute the stepsize α_i by backtracking

Set $\rho = 1$.

while $g(\mathbf{x}_i + \rho\mathbf{d}_i, \mu_i) > g(\mathbf{x}_i, \mu_i) + \delta_1\rho\mathscr{R}\left(\tilde{\mathbf{g}}_i^H\mathbf{d}_i\right)$ **do**

$\rho = \delta_2\rho$

end while

$\alpha_i = \rho$ and $\mathbf{x}_{i+1} = \mathbf{x}_i + \alpha_i\mathbf{d}_i$

if $\|\partial g(\mathbf{x}_{i+1}, \mu_i)\|_2 \geq \gamma\mu_i$ **then**

$\mu_{i+1} = \mu_i$

else

$\mu_{i+1} = \gamma_1\mu_i$

end if

$\tilde{\mathbf{g}}_{i+1} = \partial g(\mathbf{x}_{i+1}, \mu_{i+1})$

$\tilde{\mathbf{p}}_i = \tilde{\mathbf{g}}_{i+1} - \tilde{\mathbf{g}}_i$ and $\mathbf{s}_i = \mathbf{x}_{i+1} - \mathbf{x}_i$.

$\tilde{\mathbf{z}}_i = \tilde{\mathbf{p}}_i + \left(\varepsilon_0\|\tilde{\mathbf{g}}_{i+1}\|_2^2 + \max\{0, -\frac{\mathscr{R}(\mathbf{s}_i^H\tilde{\mathbf{p}}_i)}{\|\mathbf{s}_i\|_2^2}\}\right)\mathbf{s}_i$.

$$\mathbf{d}_{i+1} = -\tilde{\mathbf{g}}_{i+1} + \mathscr{R}\left(\frac{\tilde{\mathbf{g}}_{i+1}^H\tilde{\mathbf{z}}_i}{\mathbf{d}_i^H\tilde{\mathbf{z}}_i} - \frac{2\|\tilde{\mathbf{z}}_i\|_2^2\tilde{\mathbf{g}}_{i+1}^H\mathbf{d}_i}{|\mathbf{d}_i^H\tilde{\mathbf{z}}_i|^2}\right)\mathbf{d}_i + \mathscr{R}\left(\frac{\tilde{\mathbf{g}}_{i+1}^H\mathbf{d}_i}{\mathbf{d}_i^H\tilde{\mathbf{z}}_i}\right)\tilde{\mathbf{z}}_i.$$

end for

return: $\mathbf{x}_T$

Notation: $\mathscr{R}(\cdot)$ represents the real part function.

Definition 5.2.1. *Gradient consistency property* Zhang and Chen (2009): The function $g(\cdot,\mu)$ satisfies the gradient consistency property if

$$\left\{ \lim_{\mu\downarrow 0,\mathbf{x}\to\mathbf{x}^*} \partial g(\mathbf{x},\mu) \right\} = \partial^c f(\mathbf{x}^*). \tag{50}$$

Given the fact that Algorithm 7 is a conjugate gradient method, two conditions are required to guarantee its convergence. First, the function $g(\mathbf{x},\mu)$ must satisfy Assumption 10.0.5, which will be introduced in shortly, and is used in the analysis of convergence for nonlinear conjugate gradient methods. Second, the function $g(\mathbf{x},\mu)$ needs to satisfy the gradient consistency property, which guarantees that any accumulation point of the sequence $\{\mathbf{x}_i\}$ generated by Algorithm 7 is a Clarke stationary point, as it is proven in Theorem 5.2.4

Assumption 2.

1. For any $(\mathbf{w},\mu)\in\mathbb{C}^n\times\mathbb{R}_{++}$, the level set

$$S_\mu(\mathbf{w}) = \{\mathbf{z}\in\mathbb{C}^n | g(\mathbf{z},\mu) \le g(\mathbf{w},\mu)\}, \tag{51}$$

   is bounded.

2. The Wirtinger derivative $\partial g(\mathbf{x},\mu)$ with respect to $\mathbf{x}$ is smooth and exists a constant $L_g > 0$,

such that, for any $\mathbf{w} \in \mathbb{C}^n$ and fixed $\mu \in \mathbb{R}_{++}$

$$dist(\partial g(\mathbf{z}_1, \mu), \partial g(\mathbf{z}_2, \mu)) \leq L_g dist(\mathbf{z}_1, \mathbf{z}_2), \tag{52}$$

holds for all $\mathbf{z}_1, \mathbf{z}_2 \in S_\mu(\mathbf{w})$ with probability at least $1 - me^{-n/2}$.

The following theorem, which uses the first result in Lemma 5.1.2, shows that the objective function g defined in (44) satisfies the Assumption 10.0.5.

Theorem 5.2.2. Assuming that $\text{span}(\mathbf{a}_1, \cdots, \mathbf{a}_m) = \{\sum_{k=1}^m \lambda_k \mathbf{a}_k : \lambda_k \in \mathbb{C}\} = \mathbb{C}^n$, then functions $\varphi'_\mu(x)$, and $g(\mathbf{x}, \mu)$ defined in (41) and (44) respectively, satisfy the following properties:

1. Assumption 1 is satisfied.

2. The function $\varphi'_\mu(|\mathbf{a}_k^H \mathbf{z}|)$ is Lipschitz continuous on any level set $S_\mu(\mathbf{w})$, for a fixed $\mu \in \mathbb{R}_{++}$ with probability at least $1 - me^{-n/2}$.

Demostración. Appendix E (See Chapter 10). □

The following theorem shows that the sequence $\{\mathbf{x}_i\}$ generated by Algorithm 7 converges to a stationary point.

Theorem 5.2.3. In the setup of Theorem 5.2.2 the sequences $\{\mu_i\}$ and $\{\mathbf{x}_i\}$ generated by Algorithm 7 satisfy

$$\lim_{i \to \infty} \mu_i = 0, \text{ and } \liminf_{i \to \infty} \|\partial g(\mathbf{x}_i, \mu_{i-1})\|_2 = 0. \tag{53}$$

Thus, there exists $\mathbf{x}^* \in \mathbb{C}^n$ such that $\lim_{i\to\infty} \mathbf{x}_i = \mathbf{x}^*$.

Demostración. Appendix F (See Chapter 10). □

Theorem 5.2.4, which uses the result in Theorem 5.2.3, establishes that the function $g(\mathbf{x}, \mu)$ satisfies the gradient consistency property defined in (50). Theorem 5.2.4 also shows that any stationary point of the sequence $\{\mathbf{x}_i\}$ in Algorithm 7 is a Clarke stationary point.

Theorem 5.2.4. In the setup of Theorem 5.2.2, the function $g(\cdot, \mu)$ defined in (44) satisfies the gradient consistency property as introduced in Definition 5.2.1. Also, any limit point $\mathbf{x}^*$ of the sequence $\{\mathbf{x}_i\}$ in Algorithm 7 is a Clarke stationary point, *i.e.* $\mathbf{0} \in \partial^c f(\mathbf{x}^*)$.

Demostración. Appendix G (See Chapter 10). □

Finally, it can be observed that Theorem 5.2.4 essentially proves that the Clarke derivative notion can be viewed as a limit case of the Wirtinger derivate $\partial g(\mathbf{z}, \mu)$. This theoretical result is very interesting from an optimization point of view, because two completely different derivative notions that have been used to solve the phase retrieval are compared, and it can be concluded that one is a limit case of the other.

5.3. Theoretical Advantages of the Proposed Smoothing Approach

This section analyzes why the smooth cost function in (44) performs better in comparison with its non-smooth counterparts TAF, STAF, TWF, RWF and RAF. To this end, the proposed descent direction given in (95) is analyzed. Notice that the Wirtinger derivative defined in (95) for each

iteration i in Algorithm 7 is given by

$$\partial g(\mathbf{x}_i, \mu_i) = \frac{2}{m} \sum_{k=1}^{m} \left(\mathbf{a}_k^H \mathbf{x}_i - q_k \frac{\mathbf{a}_k^H \mathbf{x}_i}{\sqrt{|\mathbf{a}_k^H \mathbf{x}_i|^2 + \mu_i^2}} \right) \mathbf{a}_k. \tag{54}$$

Then, observe that (54) can be rewritten as

$$\partial g(\mathbf{x}_i, \mu_i) = \frac{2}{m} \sum_{k=1}^{m} \left(1 - \frac{|\mathbf{a}_k^H \mathbf{x}|}{\sqrt{|\mathbf{a}_k^H \mathbf{x}_i|^2 + \mu_i^2}} \right) \mathbf{a}_k \mathbf{a}_k^H \mathbf{x}_i. \tag{55}$$

According to the update procedure of the variable μ in Algorithm 7 it can be observed

$$\|\partial g(\mathbf{x}_i, \mu_{i-1})\|_2 \leq \gamma \mu_i \leq \mu_i, \tag{56}$$

for some $\gamma \in (0,1)$. Further, in Theorem 5.2.3 establishes that, considering (56), the Wirtinger derivative in (55) tends to zero. Then, combining the result in Theorem 5.2.3 and inequality (56), it can be concluded that

$$\left| 1 - \frac{|\mathbf{a}_k^H \mathbf{x}|}{\sqrt{|\mathbf{a}_k^H \mathbf{x}_i|^2 + \mu_i^2}} \right| < 1, \tag{57}$$

for all $k \in \{1, \cdots, m\}$, otherwise inequality (56) does not hold (see Appendix C in the Supplementary material). Considering this fact, the Wirtinger gradient in (55), used by the proposed method, does not need truncation thresholds because $\frac{|\mathbf{a}_k^H \mathbf{x}|}{\sqrt{|\mathbf{a}_k^H \mathbf{x}_i|^2 + \mu_i^2}} < 2$ (it is bounded). Indeed, in the case of

STAF and TAF, the variable $\mu_i = 0$ for all $i > 0$, which implies that their gradients are given by

$$\partial g(\mathbf{x}_i, 0) = \frac{2}{m} \sum_{k=1}^{m} \left(1 - \frac{|\mathbf{a}_k^H \mathbf{x}|}{|\mathbf{a}_k^H \mathbf{x}_i|} \right) \mathbf{a}_k \mathbf{a}_k^H \mathbf{x}_i. \tag{58}$$

Notice that (58) could lead to excessively large size of the Wirtinger derivative because the term $\frac{|\mathbf{a}_k^H \mathbf{x}|}{|\mathbf{a}_k^H \mathbf{x}_i|}$, introduces bias in the update direction Wang et al. (2016a). This fact is the main reason why (58) (the Wirtinger gradient used in TAF and STAF) requires a truncation procedure in order to avoid a deviation in the update direction Wang et al. (2016a); Chen and Candès (2015).

On the other hand, since the proposed update direction in (95) does not need truncation thresholds, then the proposed cost function $g(\mathbf{z}, \mu)$ is locally smooth. In fact, the following Theorem 5.3.1 establishes that the whole Wirtinger derivative $\partial g(\mathbf{z}, \mu)$ in (95) does not vary too much around the curve of optimizers.

Theorem 5.3.1. *(Local smoothness property Candès et al. (2015))* The Wirtinger gradient defined in (95) satisfies the following property

$$\|\partial g(\mathbf{z}, \mu)\|_2 \leq \beta dist(\mathbf{z}, \mathbf{x}) + \frac{\rho}{m} \sum_{k=1}^{m} |\mathbf{a}_k^H \mathbf{h}|, \tag{59}$$

where $\rho, \beta \in \mathbb{R}_{++}$ with probability at least $1 - me^{-n/2}$ when $m \geq C(\varepsilon_0)n$ for some constant $C(\varepsilon_0)$ depending on $\varepsilon_0 > 0$ and $\mathbf{h} = \mathbf{x} - e^{-j\theta(z)}\mathbf{z}$ with $\theta(z) = \arg\min_{\theta \in [0, 2\pi)} \|\mathbf{x} - e^{-j\theta}\mathbf{z}\|_2$.

Demostración. Appendix H (See Chapter 10). □

Considering the result in Theorem 5.3.1 it can be obtained that the local smoothness pro-

perty is preserved for the whole Wirtinger derivative $\partial g(\mathbf{z}, \mu)$. In contrast, for methods such as TAF, STAF, RAF, TWF and RWF, that truncate or have a non-smooth update direction, the local smoothness property is preserved just for a piece of the update direction, thus introducing an important deviation of their search directions Wang et al. (2016a); Zhang and Liang (2016); Wang et al. (2017a), which leads to a reduced performance to solve the phase retrieval problem as it will be illustrated in Chapter 9.

6. Extension to Frequency-resolved optical gating

This chapter presents an extension of the theoretical results of this thesis to a real phase retrieval problem present in the Frequency-resolved optical gating (FROG) phenomenon. Specifically, FROG is a popular technique for complete characterization of ultrashort laser pulses. The acquired data in FROG, called FROG trace, is the Fourier magnitude of the product of the unknown pulse with a time-shifted version of itself, for several different shifts. Figure illustrates the acquisition system in FROG.

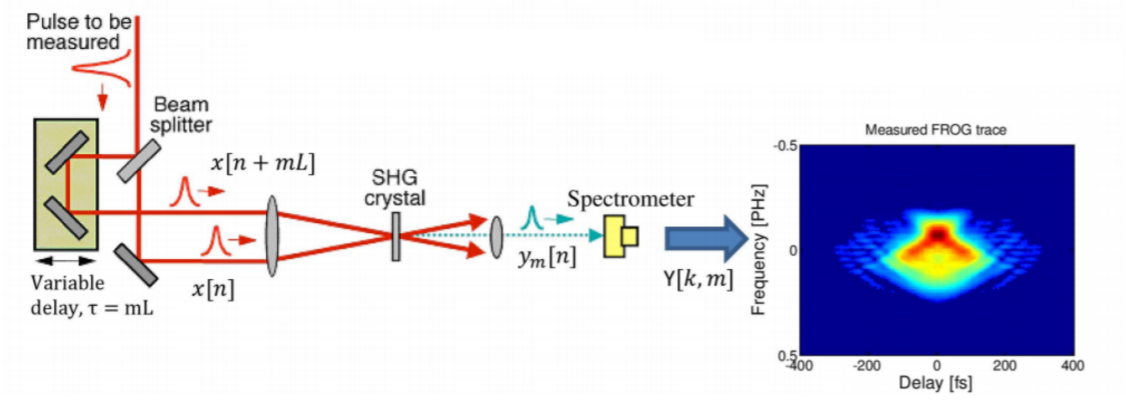

Figura 6. Illustration of the SHG FROG technique Bendory et al. (2017b).

To estimate the pulse from the FROG trace, this chapter introduces an algorithm that mini-

mizes a smoothed non-convex least-squares objective function. The method consists of two steps. First, the pulse is approximated by an iterative spectral algorithm. Then, the attained initialization is refined based upon a sequence of block stochastic gradient iterations. The algorithm is theoretically simple, numerically scalable, and easy-to-implement. Empirically, our approach outperforms the state-of-the-art when the FROG trace is incomplete, that is, when only few shifts are recorded. Simulations also suggest that the proposed algorithm exhibits similar computational cost compared to a state-of-the-art technique for both complete and incomplete data. In addition, we prove that in the vicinity of the true solution, the algorithm converges to a critical point. A Matlab implementation is publicly available at [1].

6.1. FROG Phase Retrieval Problem

Mathematically, the FROG trace of a signal $\mathbf{x} \in \mathbb{C}^N$ is defined as

$$\mathbf{Z}[p,k] := \left| \sum_{n=0}^{N-1} \mathbf{x}[n]\mathbf{x}[n+pL]e^{-2\pi ink/N} \right|^2,$$
$$k = 0, \cdots, N-1,\ p = 0, \cdots, R-1, \tag{60}$$

with $R = \lceil N/L \rceil$ where $L < N$ and $i := \sqrt{-1}$. This work assumes that the signal $\mathbf{x}$ is periodic, that is, $\mathbf{x}[n] = \mathbf{x}[n+lN]$ for any $l \in \mathbb{Z}$.

The FROG trace defined in (60) can be considered as a map $\mathbb{C}^N \to \mathbb{R}_+^{\lceil N/L \rceil}$ that has three types of symmetry, usually called *trivial ambiguities* in the PR literature. These ambiguities are

---

[1] `https://github.com/samuelpinilla/FROG`

summarized in Proposition 3, using the following definition of a bandlimited signal.

Definition 6.1.1. We say that $\mathbf{x} \in \mathbb{C}^N$ is a $B-bandlimited$ signal if its Fourier transform $\tilde{\mathbf{x}} \in \mathbb{C}^N$ contains $N-B$ consecutive zeros. That is, there exists k such that $\tilde{\mathbf{x}}[k] = \cdots = \tilde{\mathbf{x}}[N+k+B-1] = 0$.

Proposition 1. (Bendory et al. (2018a)) Let $\mathbf{x} \in \mathbb{C}^N$ be the underlying signal and let $\tilde{\mathbf{x}} \in \mathbb{C}^N$ be its Fourier transform. Let $\mathbf{Z}[p,k]$ be the FROG trace of $\mathbf{x}$ defined as in (60) for some fixed L. Then, the following signals have the same FROG trace as $\mathbf{x}$:

1. the rotated signal $\mathbf{x}e^{i\phi}$ for some $\phi \in \mathbb{R}$;
2. the translated signal $\mathbf{x}^\ell$ obeying $\mathbf{x}^\ell[n] = \mathbf{x}[n-\ell]$ for some $\ell \in \mathbb{Z}$ (equivalently, a signal with Fourier transform $\tilde{\mathbf{x}}^\ell$ obeying $\tilde{\mathbf{x}}^\ell[k] = \tilde{\mathbf{x}}[k]e^{-2\pi i\ell k/N}$ for some $\ell \in \mathbb{Z}$);
3. the reflected signal $\hat{\mathbf{x}}$ obeying $\hat{\mathbf{x}}[n] := \overline{\mathbf{x}}[-n]$.

If $\mathbf{x}$ is a B-bandlimited signal for some $B \leq N/2$, then the translation ambiguity is continuous. Namely, any signal with a Fourier transform such that $\tilde{\mathbf{x}}^\psi[k] := \tilde{\mathbf{x}}[k]e^{i\psi k}$ for some $\psi \in \mathbb{R}$, has the same FROG trace as $\mathbf{x}$.

Our goal is to estimate the signal $\mathbf{x}$, up to trivial ambiguities, from the FROG trace $\mathbf{Z}$. The work Bendory et al. (2018a) established that the pulse $\mathbf{x}$ can be uniquely identified (up to trivial ambiguities) from the FROG trace under rather mild conditions as summarized in the following proposition.

Proposition 2. (Bendory et al. (2018a)) Let $\mathbf{x} \in \mathbb{C}^N$ be a B-bandlimited signal as in Definition 7.1.1 for some $B \leq N/2$. If $N/L \geq 4$, then almost all signals are determined uniquely from their FROG trace $\mathbf{Z}[p,k]$, up to trivial ambiguities, from $m \geq 3B$ measurements. If in addition we have access to the signal's power spectrum and $N/L \geq 3$, then $m \geq 2B$ measurements suffice.

Proposition 4 has been recently extended to the case of blind ptychography, or blind FROG, in which the goal is to estimate two signals simultaneously Bendory et al. (2019a). Evidently, Proposition 4 allows choices of $L > 1$ meaning that not all the delay steps are needed to recover the pulse, and therefore a method that works in this regime as well is desired.

To take the ambiguities into account, we measure the relative error between the true signal $\mathbf{x}$ and any $\mathbf{w} \in \mathbb{C}^N$ as

$$\operatorname{dist}(\mathbf{x},\mathbf{w}) := \frac{\left\| \sqrt{\mathbf{Z}} - \sqrt{\mathbf{W}} \right\|_{\mathrm{F}}}{\left\| \sqrt{\mathbf{Z}} \right\|_{\mathrm{F}}}, \tag{61}$$

where $\mathbf{Z}$ is the FROG trace of $\mathbf{x}$ according to (60), $\sqrt{\cdot}$ is the point-wise square root, $\mathbf{W}$ is the FROG trace of $\mathbf{w}$, and $\|\cdot\|_{\mathrm{F}}$ denotes the Frobenius norm. Note that if $\operatorname{dist}(\mathbf{x},\mathbf{w}) = 0$, and the uniqueness conditions of Proposition 4 are met, then for almost all signals $\mathbf{w}$ is equal to $\mathbf{x}$ up to trivial ambiguities.

In recent years, many papers have examined the problem of recovering a signal from phaseless quadratic random measurements. A popular approach is to minimize the intensity least-squares objective; see for instance Candès et al. (2015). Recent works have shown that minimizing the amplitude least-squares objective leads to better reconstruction under noisy scenarios Pauwels et al. (2018); Wang et al. (2016a); Zhang and Liang (2016). However, the latter cost function is non-

smooth and thus may lead to a biased descent direction Pinilla et al. (2018a). To overcome the non-smoothness of the objective function, we follow the smoothing strategy proposed in Pinilla et al. (2018a).

The smooth objective to recover the underlying pulse considered in this work is

$$\min_{\mathbf{z}\in\mathbb{C}^n} h(\mathbf{z},\mu) = \min_{\mathbf{z}\in\mathbb{C}^n} \frac{1}{NR}\sum_{k=0}^{N-1}\sum_{p=0}^{R-1} \ell_{k,p}(\mathbf{z},\mu), \tag{62}$$

where

$$\ell_{k,p}(\mathbf{z},\mu) := \left[\varphi_\mu\left(\left|\sum_{n=0}^{N-1}\mathbf{z}[n]\mathbf{z}[n+pL]e^{-2\pi ink/N}\right|\right) - \sqrt{\mathbf{Z}[p,k]}\right]^2. \tag{63}$$

The function $\varphi_\mu : \mathbb{R} \to \mathbb{R}_{++}$ in (88) is defined as $\varphi_\mu(w) := \sqrt{w^2+\mu^2}$, with $\mu \in \mathbb{R}_{++}$ (a tunable parameter). Notice that if $\mu = 0$, then (88) reduces to the non-smooth formulation. In Wang et al. (2016a), the authors addressed the non-smoothness by introducing truncation parameters into the gradient step in order to eliminate the errors in the estimated descent direction. However, this procedure can modify the search direction and increase the sample complexity of the phase retrieval problem Pinilla et al. (2018a).

In this chapter a block stochastic gradient algorithm (BSGA) is presented to solve (87), that is initialized by a spectral procedure which requires only a few iterations.

6.2. Reconstruction Algorithm

In order to solve the optimization problem in (87), we develop a gradient-based algorithm, called BSGA. The algorithm is initialized by the outcome of a spectral method approximating the signal $\mathbf{x}$ which will be explained in Section 7.3.

To refine the initial estimate we use the Wirtinger derivatives as introduced in Hunger (2007). Let us define the vector $\mathbf{f}_k^H$ as

$$\mathbf{f}_k^H := \left[\omega^{-0(k-1)}, \omega^{-1(k-1)}, \cdots, \omega^{-(n-1)(k-1)}\right], \tag{64}$$

with $\omega = e^{\frac{2\pi i}{n}}$ the nth root of unity. Then, the Wirtinger derivative of $h(\mathbf{z}, \mu)$ in (87) with respect to $\overline{\mathbf{z}}[\ell]$ is given by

$$\frac{\partial h(\mathbf{z}, \mu)}{\partial \overline{\mathbf{z}}[\ell]} := \frac{1}{NR} \sum_{k=0}^{N-1} \sum_{p=1}^{R-1} \left(\mathbf{f}_k^H \mathbf{g}_p - \upsilon_{k,p}\right) \overline{q}_{\ell,p} e^{2\pi i \ell k/N}, \tag{65}$$

where $\upsilon_{k,p} := \sqrt{\mathbf{Z}[p,k]} \frac{\mathbf{f}_k^H \mathbf{g}_p}{\varphi_\mu\left(|\mathbf{f}_k^H \mathbf{g}_p|\right)}$, and

$$\begin{aligned} \overline{q}_{\ell,p} :=& \overline{\mathbf{z}}[\ell + p] + \overline{\mathbf{z}}[\ell - p] e^{-2\pi i k p/N}, \\ \mathbf{g}_p :=& \left[\mathbf{z}[0]\mathbf{z}[pL], \cdots, \mathbf{z}[N-1]\mathbf{z}[N-1+pL]\right]^T. \end{aligned} \tag{66}$$

The gradient of $h(\mathbf{z}, \mu)$ is then

$$\frac{\partial h(\mathbf{z}, \mu)}{\partial \overline{\mathbf{z}}} := \left[\frac{\partial h(\mathbf{z}, \mu)}{\partial \overline{\mathbf{z}}[0]}, \cdots, \frac{\partial h(\mathbf{z}, \mu)}{\partial \overline{\mathbf{z}}[N-1]}\right]^H. \tag{67}$$

Using (95), we define a standard gradient algorithm, taking the form of

$$\mathbf{x}^{(t+1)} := \mathbf{x}^{(t)} - \alpha \frac{\partial h(\mathbf{x}^{(t)}, \mu^{(t)})}{\partial \overline{\mathbf{z}}}, \tag{68}$$

where α is the step size.

To alleviate the memory requirements and computational complexity required for large N, we suggest a block stochastic gradient descent strategy. Instead of calculating (93), we choose only a random subset of the sum for each iteration t, that is,

$$\mathbf{d}_{\Gamma_{(t)}}[\ell] = \sum_{p,k\in\Gamma_{(t)}} \left(\mathbf{f}_k^H \mathbf{g}_p^{(t)} - \upsilon_{k,p,t}\right) \overline{q}_{\ell,p}^{(t)} e^{2\pi i \ell k/N}, \tag{69}$$

where the set $\Gamma_{(t)}$ is chosen uniformly and independently at random at each iteration t from subsets of $\{1,\cdots,N\}\times\{1\cdots,R\}$ with cardinality Q. Specifically, the gradient in (95) is uniformly sampled using a minibatch of data, in this case of size Q for each step update, such that in expectation is (93) (Spall, 2005, page 130).

As mentioned in Chapter 5, choosing $\mu > 0$ prevents bias in the update direction. Since the function h is smooth, we are able to construct a descent rule for μ (Line 13 of Algorithm 4) in order to guarantee convergence to a first-order optimal point, that is, a point with zero gradient, in the vicinity of the solution.

Algorithm 4

Input: Data $\{\mathbf{Z}[p,k] : k = 0,\cdots,N-1, p = 0,\cdots,R-1\}$. Choose constants $\gamma_1, \gamma, \alpha \in (0,1)$, $\mu^{(0)} \geq 0$, cardinality $Q \in \{1,\cdots,NR\}$, and tolerance $\varepsilon > 0$.

if $L = 1$ **then**

    Initial point $\mathbf{x}^{(0)} \leftarrow$ Algorithm 2$(\mathbf{Z}[p,k], T)$.

else

    Initial point $\mathbf{x}^{(0)} \leftarrow$ Algorithm 3$(\mathbf{Z}[p,k], T)$.

end if

while $\left\|\mathbf{d}_{\Gamma_{(t)}}\right\|_2 \geq \varepsilon$ **do**

Choose $\Gamma_{(t)}$ uniformly at random from the subsets of $\{1,\cdots,N\} \times \{1\cdots,R\}$ with cardinality Q per iteration $t \geq 0$.

    $\mathbf{x}^{(t+1)} = \mathbf{x}^{(t)} - \alpha \mathbf{d}_{\Gamma_{(t)}}$,

where

    $\mathbf{d}_{\Gamma_{(t)}}[\ell] = \sum_{p,k \in \Gamma_{(t)}} \left(\mathbf{f}_k^H \mathbf{g}_p^{(t)} - \upsilon_{k,p,t}\right) \overline{q}_{\ell,p}^{(t)} e^{2\pi i \ell k/N}$.

    $\upsilon_{k,p,t} = \sqrt{\mathbf{Z}[p,k]} \frac{\mathbf{f}_k^H \mathbf{g}_p^{(t)}}{\varphi_{\mu^{(t)}}\left(\left|\mathbf{f}_k^H \mathbf{g}_p^{(t)}\right|\right)}$.

    $\mathbf{g}_p^{(t)} = \left[\mathbf{x}^{(t)}[0]\mathbf{x}^{(t)}[pL],\cdots,\mathbf{x}^{(t)}[N-1]\mathbf{x}^{(t)}[N-1+pL]\right]^T$.

    $q_{\ell,p}^{(t)} = \mathbf{x}^{(t)}[\ell+p] + \mathbf{x}^{(t)}[\ell-p] e^{2\pi i k p/N}$.

    if $\left\|\mathbf{d}_{\Gamma_{(t)}}\right\|_2 \geq \gamma \mu^{(t)}$ **then**

        $\mu^{(t+1)} = \mu^{(t)}$.

    else

        $\mu^{(t+1)} = \gamma_1 \mu^{(t)}$.

    end if

end while

return: $\mathbf{x}^{(T)}$.

Theorem 6.2.1. Let $\mathbf{x}$ be B-bandlimited for some $B \leq N/2$, satisfying $\operatorname{dist}(\mathbf{x}, \mathbf{x}^{(t)}) \leq \rho$ for some sufficiently small constant $\rho > 0$. Suppose that $L = 1$ and $\Gamma_{(t)}$ is sampled uniformly at random from all subsets of $\{1, \cdots, N\} \times \{1 \cdots, R\}$ with cardinality Q, independently for each iteration. Then for almost all signals, Algorithm 4 with step size $\alpha \in (0, \frac{2}{U}]$ satisfies

$$\lim_{t\to\infty} \mu^{(t)} = 0, \text{ and } \lim_{t\to\infty} \left\| \frac{\partial h(\mathbf{x}^{(t)}, \mu^{(t)})}{\partial \overline{\mathbf{z}}} \right\|_2 = 0, \tag{70}$$

for some constant $U > 0$ depending on ρ.

Demostración. See Chapter 8 (Appendix H). □

6.3. Initialization Strategy

In this section we devise a method to initialize the gradient iterations. This strategy approximates the signal $\mathbf{x}$ from the FROG trace as the leading eigenvector of a carefully designed matrix. We divide the exposition of the initialization procedure into two cases, $L = 1$ and $L > 1$, explained in Sections 6.3.1 and 6.3.2, respectively.

6.3.1. Initialization for $L = 1$. Instead of directly dealing with the FROG trace in (60), we consider the acquired data in a transformed domain by taking its 1D DFT with respect to the frequency variable (normalized by $1/N$). Our measurement model is then

$$\begin{aligned} \mathbf{Y}[p, \ell] &= \frac{1}{N} \sum_{k=0}^{N-1} \mathbf{Z}[p, k] e^{-2\pi i k\ell/N} = \frac{1}{N} \sum_{k,n,m=0}^{N-1} \mathbf{x}[n]\overline{\mathbf{x}}[m]\mathbf{x}[n + pL]\overline{\mathbf{x}}[m + pL] e^{-2\pi i k \frac{(m-n-\ell)}{N}} \\ &= \sum_{n=0}^{N-1} \mathbf{x}[n]\overline{\mathbf{x}}[n + \ell]\mathbf{x}[n + pL]\overline{\mathbf{x}}[n + \ell + pL], \end{aligned} \tag{71}$$

where $p,\ell = 0,\cdots,N-1$. Observe that for fixed p, $\mathbf{Y}[p,\ell]$ is the autocorrelation of $\mathbf{x}\odot\mathbf{x}_{pL}$, where $\mathbf{x}_{pL}[n] = \mathbf{x}[n+pL]$.

Let $\mathbf{D}_{pL} \in \mathbb{C}^{N\times N}$ be a diagonal matrix composed of the entries of $\mathbf{x}_{pL}$, and let $\mathbf{C}_\ell$ be a circulant matrix that shifts the entries of a vector by ℓ locations, namely, $(\mathbf{C}_\ell\mathbf{x})[n] = \mathbf{x}[n+\ell]$. Then, the matrix $\mathbf{X} := \mathbf{x}\mathbf{x}^H$ is linearly mapped to $\mathbf{Y}[p,\ell]$ as follows:

$$\begin{aligned}\mathbf{Y}[p,\ell] &= \left(\mathbf{D}_{pL+\ell}\overline{\mathbf{D}}_{pL}\mathbf{C}_\ell\mathbf{x}\right)^H\mathbf{x} = \mathbf{x}^H\mathbf{A}_{p,\ell}\mathbf{x}\\ &= tr(\mathbf{X}\mathbf{A}_{p,\ell}),\end{aligned} \tag{72}$$

where $\mathbf{A}_{p,\ell} = \mathbf{C}_{-\ell}\mathbf{D}_{pL}\overline{\mathbf{D}}_{pL+\ell}$, and $tr(\cdot)$ denotes the trace function. Observe that $\mathbf{C}_\ell^T = \mathbf{C}_{-\ell}$. Thus, we have that

$$\mathbf{y}_\ell = \mathbf{G}_\ell\mathbf{x}_\ell, \tag{73}$$

for a fixed $\ell \in \{0,\cdots,N-1\}$, where $\mathbf{y}_\ell[n] = \mathbf{Y}[n,\ell]$ and $\mathbf{x}_\ell = \mathrm{diag}(\mathbf{X},\ell)$. The (p,n)th entry of the matrix $\mathbf{G}_\ell \in \mathbb{C}^{\lceil\frac{N}{L}\rceil\times N}$ is given by

$$\mathbf{G}_\ell[p,n] := \mathbf{x}_{pL}[n]\overline{\mathbf{x}}_{pL}[n+\ell]. \tag{74}$$

Since $L = 1$, it follows from (101) that $\mathbf{G}_\ell$ is a circulant matrix. Therefore, $\mathbf{G}_\ell$ is invertible if and only if the DFT of its first column, in this case $\mathbf{x}\odot(\mathbf{C}_\ell\overline{\mathbf{x}})$, is non-vanishing.

Using (100), we propose a method to estimate the signal $\mathbf{x}$ from measurements (60) using

an alternating scheme: fixing $\mathbf{G}_\ell$, solving for $\mathbf{x}_\ell$, updating $\mathbf{G}_\ell$ and so forth. The new methodology proposed in Bendory et al. (2018b) cannot be directly employed since here the matrices $\mathbf{G}_\ell$ are also unknown. Thus, our approach estimates the matrices $\mathbf{G}_\ell$ together with $\mathbf{x}_\ell$.

We start the alternating scheme by the initialization suggested in Sidorenko et al. (2016)

$$\mathbf{x}_{ini\_pty}[r] := \mathbf{v}[r]\exp(i\theta[r]), \tag{75}$$

where $\theta[r] \in [0, 2\pi)$ is chosen uniformly at random for all $r \in \{0, \cdots, N-1\}$. The rth entry of $\mathbf{v}$ corresponds to the summation of the measured FROG trace over the frequency axis:

$$\mathbf{v}[r] := \frac{1}{N}\sum_{k=0}^{N-1}\mathbf{Z}[r,k] = \sum_{k=0}^{N-1}\left|\sum_{n=0}^{N-1}\mathbf{x}[n]\mathbf{x}[n+rL]e^{-2\pi ink/N}\right|^2 := \sum_{n=0}^{N-1}|\mathbf{x}[n]|^2|\mathbf{x}[n+rL]|^2. \tag{76}$$

Once the vector $\mathbf{x}_{ini\_pty}$ is constructed, the vectors $\mathbf{x}_\ell^{(t)}$ at $t = 0$ can be built as

$$\mathbf{x}_\ell^{(0)} = \mathrm{diag}(\mathbf{X}_0^{(0)}, \ell), \tag{77}$$

where

$$\mathbf{X}_0^{(0)} = \mathbf{x}_{ini\_pty}\mathbf{x}_{ini\_pty}^H. \tag{78}$$

Then, from (104) we proceed with an alternating procedure between estimating the matrix $\mathbf{G}_\ell$, and updating the vector $\mathbf{x}_\ell$ as follows.

Update rule for $\mathbf{G}_\ell$*:* In order to update $\mathbf{G}_\ell$, we update the matrix $\mathbf{X}_0^{(t)}$ as

$$\operatorname{diag}(\mathbf{X}_0^{(t)}, \ell) = \mathbf{x}_\ell^{(t)}. \tag{79}$$

Observe that if $\mathbf{x}_\ell^{(t)}$ is close to $\mathbf{x}_\ell$ for all ℓ, then $\mathbf{X}_0^{(t)}$ is close to $\mathbf{x}\mathbf{x}^H$. Letting $\mathbf{w}^{(t)}$ be the leading (unit-norm) eigenvector of the matrix $\mathbf{X}_0^{(t)}$ constructed in (106), from (101) each matrix $\mathbf{G}_\ell^{(t)}$ at iteration t is given by

$$\mathbf{G}_\ell^{(t)}[p,n] = \mathbf{x}_{pL}^{(t)}[n]\overline{\mathbf{x}}_{pL}^{(t)}[n+\ell], \tag{80}$$

where $\mathbf{x}_{pL}^{(t)}[n] = \mathbf{w}^{(t)}[n+pL]$.

Optimization with respect to $\mathbf{x}_\ell$*:* Fixing $\mathbf{G}_\ell^{(t-1)}$, one can estimate $\mathbf{x}_\ell^{(t)}$ at iteration t by solving the linear least-squares (LS) problem

$$\min_{\mathbf{p}_\ell \in \mathbb{C}^N} \quad \|\mathbf{y}_\ell - \mathbf{G}_\ell^{(t-1)}\mathbf{p}_\ell\|_2^2. \tag{81}$$

The relationship between the vectors $\mathbf{x}_\ell^{(t)}$ is ignored at this stage. If $\mathbf{G}_\ell^{(t-1)}$ is invertible, then the solution to this problem is given by $(\mathbf{G}_\ell^{(t-1)})^{-1}\mathbf{y}_\ell$. Since $\mathbf{G}_\ell^{(t-1)}$ is a circulant matrix, it is invertible if and only if the DFT of $\mathbf{x}^{(t-1)} \odot (\mathbf{C}_\ell \overline{\mathbf{x}}^{(t-1)})$ is non-vanishing. This condition cannot be ensured in general. Thus, we propose a surrogate proximal optimization problem

to estimate $\mathbf{x}_\ell^{(t)}$ by

$$\min_{\mathbf{p}_\ell \in \mathbb{C}^N} \quad \|\mathbf{y}_\ell - \mathbf{G}_\ell^{(t-1)}\mathbf{p}_\ell\|_2^2 + \frac{1}{2\lambda}\|\mathbf{p}_\ell - \mathbf{x}_\ell^{(t-1)}\|_2^2, \tag{82}$$

where $\lambda > 0$ is a regularization parameter. In practice λ is a tunable parameter Parikh and Boyd (2014). In particular, for this work the value of λ was determined using a cross-validation strategy such that each simulation uses the value that results in the smallest relative error according to (86). The surrogate optimization problem in (109) is strongly convex Parikh and Boyd (2014), and admits the following closed form solution $\mathbf{x}_\ell^{(t)} = \mathbf{B}_{\ell,t}^{-1}\mathbf{e}_{\ell,t}$, where

$$\begin{aligned} \mathbf{B}_{\ell,t} &= \left(\mathbf{G}_\ell^{(t-1)}\right)^H \left(\mathbf{G}_\ell^{(t-1)}\right) + \frac{1}{2\lambda}\mathbf{I}, \\ \mathbf{e}_{\ell,t} &= \left(\mathbf{G}_\ell^{(t)}\right)^H \mathbf{y}_\ell + \frac{1}{2\lambda}\mathbf{x}_\ell^{(t-1)}, \end{aligned} \tag{83}$$

with $\mathbf{I} \in \mathbb{R}^{N\times N}$ the identity matrix. Clearly $\mathbf{B}_{\ell,t}$ in (111) is always invertible. The update step for each $\mathbf{x}_\ell^{(t)}$ is computed in Line 9 of Algorithm 9.

Finally, in order to estimate $\mathbf{x}$, the (unit-norm) principal eigenvector of $\mathbf{X}_0^{(T)}$ is normalized by

$$\beta = \sqrt{\sum_{n\in\mathscr{S}} \left(\mathbf{B}_{0,T}^{-1}\mathbf{e}_{0,T}\right)[n]}, \tag{84}$$

where $\mathscr{S} := \left\{n : \left(\mathbf{B}_{0,T}^{-1}\mathbf{e}_{0,T}\right)[n] > 0\right\}$. Observe that (112) results from the fact that $\sum_{n=0}^{N-1} \mathrm{diag}(\mathbf{X},0)[n] =$

Algorithm 5

Input: The measurements $\mathbf{Z}[p,k]$, T the number of iterations, and $\lambda > 0$.
Output: $\mathbf{x}^{(0)}$ (estimation of $\mathbf{x}$).
Initialize: $\mathbf{x}_{ini\_pty}[r] = \mathbf{v}[r]\exp(i\theta[r])$, and $\mathbf{v}[r] = \frac{1}{N}\sum_{k=0}^{N-1}\mathbf{Z}[r,k]$, $\theta[r] \in [0,2\pi)$ is chosen uniformly and independently at random.
Compute $\mathbf{Y}[p,\ell]$ the 1D inverse DFT with respect to k of $\mathbf{Z}[p,k]$.
for $t = 1$ to T **do**
    Construct $\mathbf{G}_\ell^{(t)}$ according to (107).
    Compute $\mathbf{B}_{\ell,t} = (\mathbf{G}_\ell^{(t)})^H(\mathbf{G}_\ell^{(t)}) + \frac{1}{2\lambda}\mathbf{I}$.
    Compute $\mathbf{e}_{\ell,t} = (\mathbf{G}_\ell^{(t)})^H\mathbf{y}_\ell + \frac{1}{2\lambda}\mathbf{x}_\ell^{(t-1)}$.
    Construct the matrix $\mathbf{X}_0^{(t)}$ such that

$$\operatorname{diag}(\mathbf{X}_0^{(t)},\ell) = \mathbf{B}_{\ell,t}^{-1}\mathbf{e}_{\ell,t}, \quad \ell = 0,\cdots,N-1.$$

    Let $\mathbf{w}^{(t)}$ be the leading (unit-norm) eigenvector of $\mathbf{X}_0^{(t)}$.
    Take $\mathbf{x}_{pL}^{(t)}[n] = \mathbf{w}^{(t)}[n+pL]$.
end for
Compute vector $\mathbf{x}^{(0)}$ as

$$\mathbf{x}^{(0)} := \sqrt{\sum_{n\in\mathscr{S}}\left(\mathbf{B}_{0,T}^{-1}\mathbf{e}_{0,T}\right)[n]}\,\mathbf{w}^{(T)},$$

where $\mathscr{S} := \left\{n : \left(\mathbf{B}_{0,T}^{-1}\mathbf{e}_{0,T}\right)[n] > 0\right\}$.
return: $\mathbf{x}^{(0)}$.

$\|\mathbf{x}\|_2^2$.

After a few iterations of this two-step procedure, the output is used to initialize the gradient algorithm described in Section 7.2. This alternating scheme is summarized in Algorithm 9.

6.3.2. FROG initialization step for $L > 1$. Until now we focused on the case $L = 1$. If $L > 1$, then the linear system in (100) is underdetermined and $\mathbf{y}_\ell$ can be viewed as a subsampled version of (98) by a factor L. Therefore, in order to increase the number of equations when $L > 1$, we up-sample $\mathbf{y}_\ell$ by a factor L. Specifically, we follow the proposed scheme in Bendory et al.

(2018b) that expands the measurement vector $\mathbf{y}_\ell$ by low-pass interpolation. Once the measurements are upsampled, we proceed as for $L = 1$. This initialization, for $L > 1$, is summarized in Algorithm 6. From Line 3 to Line 5 the low-pass interpolation by a factor L is computed, and then in Line 6, Algorithm 9 generates the initial estimation of the underlying signal.

Algorithm 6

1: **Input:** The measurements $\mathbf{Z}[p,k]$, T the number of iterations, and a smooth interpolation filter $\mathbf{s}_L$ that approximates a lowpass filter with bandwidth $\lceil N/L \rceil$.

2: **Output:** $\mathbf{x}^{(0)}$ (estimation of $\mathbf{x}$).

3: Compute $\mathbf{Y}[p,\ell]$ as the 1D DFT with respect to k of $\mathbf{Z}[p,k]$.

4:

- *Expansion:*

$$\grave{\mathbf{y}}_\ell[n] = \begin{cases} \mathbf{y}_\ell[p] & \text{if } n = pL \\ 0 & \text{otherwise.} \end{cases}$$

- *Interpolation:*

$$\mathbf{y}_\ell^{(I)} = \grave{\mathbf{y}}_\ell * \mathbf{s}_L.$$

5: Compute $\mathbf{Y}^{(I)}[p,\ell] = \mathbf{y}_\ell^{(I)}[p]$.

6: Compute $\mathbf{Z}^{(I)}[p,k] = |\tilde{\mathbf{Y}}^{(I)}[p,k]|^2$ where $\tilde{\mathbf{Y}}^{(I)}[p,k]$ is the 1D inverse DFT with respect to ℓ of $\mathbf{Y}^{(I)}[p,\ell]$.

7: Compute $\mathbf{x}^{(0)} \leftarrow$ Algorithm 2$(\mathbf{Z}^{(I)}, T)$

8: **return:** $\mathbf{x}^{(0)}$.

7. Extension to Radar Waveform Design

This chapter presents an extension of the theoretical results of this thesis to a real phase retrieval problem present in radar. Specifically, radar signals play a central role in applications such as wireless systems, surveillance and vehicle-to-vehicle communications. In these applications the correlation between the signal emitted by the radar transmitter and its echo is analyzed by building the *radar ambiguity function*. Mathematically, this function is the Fourier magnitude of the product of the unknown signal with a conjugate time-shifted version of itself, for several different shifts. To estimate a (time) band-limited signal from the radar ambiguity function, this chapter presents a uniqueness theoretical result which states that the underlying signal can be recovered from at least (3S) 3B measurements where (S) B is the (timewidth) bandwidth, respectively. Additionally, a trust region algorithm that minimizes a smoothed non-convex least-squares objective function is proposed to iteratively estimate the band-limited signal of interest. The method consists of two steps. First, we approximate the signal by an iterative spectral algorithm. Then, the attained initialization is refined based upon a sequence of gradient iterations. To the best of our knowledge this work is seminal in the sense of solving the radar phase retrieval problem for both time-limited and band-limited signals. Simulations results suggest that the proposed algorithm is able to estimate time-limited and band-limited signal from the radar ambiguity function for both complete and incomplete radar ambiguity function. The radar function is incomplete when only few shifts or Fourier frequencies are considered.

7.1. Radar Phase Retrieval Problem

Mathematically, the radar *ambiguity function* of a signal $\mathbf{x} \in \mathbb{C}^N$ is defined as

$$\mathbf{A}[p,k] := \left| \sum_{n=0}^{N-1} \mathbf{x}[n]\overline{\mathbf{x}[n-p]}e^{2i\pi nk/N} \right|^2, \tag{85}$$

where $\overline{\mathbf{x}}$ is the conjugate of $\mathbf{x}$, and $i = \sqrt{-1}$.

The ambiguity function defined in (85) can be considered as a map $\mathbb{C}^N \to \mathbb{R}_+^{N\times N}$ that has four types of symmetry, usually called *trivial ambiguities* in the radar PR literature. These ambiguities are summarized in Proposition 3.

Proposition 3. (Jaming (2010)) Let $\mathbf{x} \in \mathbb{C}^N$ be the underlying signal and let $\tilde{\mathbf{x}} \in \mathbb{C}^N$ be its Fourier transform. Let $\mathbf{A}[p,k]$ be the ambiguity function of $\mathbf{x}$ defined as in (85). Then, the following signals have the same ambiguity function as $\mathbf{x}$:

1. the rotated signal $\mathbf{x}e^{i\phi}$ for some $\phi \in \mathbb{R}$;
2. the translated signal $\mathbf{x}^\ell$ obeying $\mathbf{x}^\ell[n] = \mathbf{x}[n-\ell]$ for some $\ell \in \mathbb{Z}$ (equivalently, a signal with Fourier transform $\tilde{\mathbf{x}}^\ell$ obeying $\tilde{\mathbf{x}}^\psi[k] = \tilde{\mathbf{x}}[k]e^{i\psi k}$ for some $\psi \in \mathbb{R}$);
3. the reflected signal $\hat{\mathbf{x}}$ obeying $\hat{\mathbf{x}}[n] := \mathbf{x}[-n]$.
4. the scaled signal $\grave{\mathbf{x}}$ obeying $\grave{\mathbf{x}}[n] := e^{ibn}\mathbf{x}[n]$ for some $b \in \mathbb{R}$.

Our goal is to estimate the signal $\mathbf{x}$, up to trivial ambiguities, from the ambiguity function $\mathbf{A}$. In this work it is established that the signal $\mathbf{x}$ can be uniquely identified (up to trivial ambiguities)

from its ambiguity function under rather mild conditions as summarized in the Proposition 4 using the following definition of a band-limited signal.

Definition 7.1.1. We say that $\mathbf{x} \in \mathbb{C}^N$ is a $B-band-limited$ signal if its Fourier transform $\tilde{\mathbf{x}} \in \mathbb{C}^N$ contains $N-B$ consecutive zeros. That is, there exists k such that $\tilde{\mathbf{x}}[k] = \cdots = \tilde{\mathbf{x}}[N+k+B-1] = 0$.

Proposition 4. Let $\mathbf{x} \in \mathbb{C}^N$ be a B-band-limited signal as in Definition 7.1.1 for some $B \leq N/2$. Then almost all signals are uniquely determined from their ambiguity function $\mathbf{A}[p,k]$, up to trivial ambiguities, from $m \geq 3B$ measurements. If in addition we have access to the signal's power spectrum and $N \geq 3$, then $m \geq 2B$ measurements suffice.

Demostración. See Appendix 9. □

By almost all signals Theorem 4 means that the set of signals which cannot be uniquely determined, up to trivial ambiguities, is contained in the vanishing locus of a nonzero polynomial (see Appendix 9 for more details). Observe that evidently, Proposition 4 states that not all the delay steps are needed to recover the signal, and therefore a method that works in this regime as well is desired. Additionally, due the extension of the proof it is deferred to Appendix 9, however there are two aspects that it is important mentioning. First, the proof of Theorem 4 is a construction procedure that uses two classical results in phase retrieval, Corollary IV.3 in Bendory et al. (2019b), and Corollary 2 in Beinert and Plonka (2018). Second, the proof reveals that the first and the $(B-1)$-th rows of the ambiguity function in (85) must be perfectly preserved in order to ensure uniqueness (up to trivial ambiguities). Then, since the radar phase retrieval problem is a design

approach these two mentioned rows cannot be discarded or corrupted by any distortion noise in the design in order to guarantee uniqueness.

A direct consequence of Proposition 4 is the following corollary, under rather mild conditions, states that for almost all time-limited signals as in Definition 2 can be recovered.

Definition 7.1.2. We say that $\mathbf{x} \in \mathbb{C}^N$ is a $S-timelimited$ signal if $\mathbf{x} \in \mathbb{C}^N$ contains $N-S$ consecutive zeros. That is, there exists k such that $\mathbf{x}[k] = \cdots = \tilde{\mathbf{x}}[N+k+S-1] = 0$.

Corollary 1. Let $\mathbf{x} \in \mathbb{C}^N$ be a S-band-limited signal as in Definition 7.1.1 for some $S \leq N/2$. Then almost all signals are uniquely determined from their ambiguity function $\mathbf{A}[p,k]$, up to trivial ambiguities, from $m \geq 3S$ measurements. If in addition we have access to the signal's power spectrum and $N \geq 3$, then $m \geq 2S$ measurements suffice.

Demostración. See Appendix 10. □

We remark here that the notion almost all signals is the same as in Theorem 4. Additionally, the proof of Corollary 1 is also a construction procedure, and that the first and the $(B-1)$-th rows of the ambiguity function in (85) must be perfectly preserved in order to ensure uniqueness (up to trivial ambiguities).

To take the ambiguities into account, we measure the relative error between the true signal $\mathbf{x}$ and any $\mathbf{w} \in \mathbb{C}^N$ as

$$\operatorname{dist}(\mathbf{x},\mathbf{w}) := \frac{\left\| \sqrt{\mathbf{A}} - \sqrt{\mathbf{W}} \right\|_{\mathrm{F}}}{\left\| \sqrt{\mathbf{A}} \right\|_{\mathrm{F}}}, \tag{86}$$

where $\mathbf{A}$ is the ambiguity function of $\mathbf{x}$ according to (85), $\sqrt{\cdot}$ is the point-wise square root, $\mathbf{W}$ is

the ambiguity function of $\mathbf{w}$, and $\|\cdot\|_{\mathrm{F}}$ denotes the Frobenius norm. Note that if $\mathrm{dist}(\mathbf{x},\mathbf{w}) = 0$, and the uniqueness conditions of Proposition 4 are met, then for almost all signals $\mathbf{w}$ is equal to $\mathbf{x}$ up to trivial ambiguities.

In recent years, many papers have examined the problem of recovering a signal from phaseless quadratic random measurements. A popular approach is to minimize the intensity least-squares objective; see for instance Candès et al. (2015). Recent works have shown that minimizing the amplitude least-squares objective leads to better reconstruction under noisy scenarios Pauwels et al. (2018); Wang et al. (2016a); Zhang and Liang (2016). However, the latter cost function is non-smooth and thus may lead to a biased descent direction Pinilla et al. (2018a). To overcome the non-smoothness of the objective function, we follow the smoothing strategy proposed in Pinilla et al. (2018a).

The smooth objective to recover the underlying signal considered in this work is

$$\min_{\mathbf{z}\in\mathbb{C}^N} h(\mathbf{z},\mu) = \min_{\mathbf{z}\in\mathbb{C}^N} \frac{1}{N^2} \sum_{k,p=0}^{N-1} \ell_{k,p}(\mathbf{z},\mu), \tag{87}$$

where

$$\ell_{k,p}(\mathbf{z},\mu) := \left[\varphi_\mu \left(\left| \sum_{n=0}^{N-1} \mathbf{z}[n]\overline{\mathbf{z}[n-p]} e^{-2\pi i n k/N} \right| \right) - \sqrt{\mathbf{A}[p,k]} \right]^2. \tag{88}$$

The function $\varphi_\mu : \mathbb{R} \to \mathbb{R}_{++}$ in (88) is defined as $\varphi_\mu(w) := \sqrt{w^2+\mu^2}$, with $\mu \in \mathbb{R}_{++}$ (a tunable parameter). Notice that if $\mu = 0$, then (88) reduces to the non-smooth formulation. In Wang et al. (2016a), the authors addressed the non-smoothness by introducing truncation parameters into the gradient step in order to eliminate the errors in the estimated descent direction. However, this

procedure can modify the search direction and increase the sample complexity of the phase retrieval problem Pinilla et al. (2018a).

In this work we propose a trust region algorithm based on the Cauchy point to solve (87), that is initialized by a spectral procedure which requires only a few iterations. Section 7.2 explains in detail the proposed algorithm.

7.2. Reconstruction Algorithm

In order to solve the optimization problem in (87), we develop a trust region algorithm, based on the Cauchy point, that is initialized by the outcome of a spectral method approximating the signal $\mathbf{x}$ which will be explained in Section 7.3.

The standard update rule in this kind of methods takes the form of

$$\mathbf{x}^{(t+1)} := \mathbf{x}^{(t)} + \alpha^{(t)}\mathbf{b}^{(t)}, \tag{89}$$

where $\alpha^{(t)}$ is the step size at iteration t and the vector $\mathbf{b}^{(t)}$ is chosen in this work as

$$\begin{aligned} \mathbf{b}^{(t)} := \underset{\mathbf{b}\in\mathbb{C}^n}{\arg\min} \quad & h(\mathbf{x}^{(t)}, \mu^{(t)}) + 2\mathscr{R}\left(\mathbf{b}^H\mathbf{d}^{(t)}\right), \\ s.t \quad & \|\mathbf{b}\|_2 \leq \mu^{(t)} \end{aligned} \tag{90}$$

with $\mathscr{R}(\cdot)$ as the real part function, and $\mathbf{d}^{(t)}$ as the gradient of $h(\mathbf{z}, \alpha)$ with respect to $\overline{\mathbf{z}}$ at iteration

t. The solution to (90) is given by (Nocedal and Wright, 2006, Chapter 4)

$$\mathbf{b}^{(t)} = -\frac{\mu^{(t)}}{\|\mathbf{d}^{(t)}\|_2}\mathbf{d}^{(t)}. \tag{91}$$

To mathematically compute $\mathbf{d}^{(t)}$, the Wirtinger derivatives as introduced in Hunger (2007) are employed. Let us define the vector $\mathbf{f}_k^H$ as

$$\mathbf{f}_k^H := \left[\omega^{-0(k-1)}, \omega^{-1(k-1)}, \cdots, \omega^{-(N-1)(k-1)}\right], \tag{92}$$

with $\omega = e^{\frac{2\pi i}{n}}$ the nth root of unity. Then, the Wirtinger derivative of $h(\mathbf{z},\mu)$ in (87) with respect to $\overline{\mathbf{z}[\ell]}$ is given by

$$\frac{\partial h(\mathbf{z},\mu)}{\partial \overline{\mathbf{z}[\ell]}} := \frac{1}{N^2}\sum_{k,p=0}^{N-1}\left(\mathbf{f}_k^H\mathbf{g}_p - \upsilon_{k,p}\right)\mathbf{z}[\ell-p]e^{2\pi i\ell k/N} + \frac{1}{N^2}\sum_{k,p=0}^{N-1}\left(\mathbf{f}_k^T\overline{\mathbf{g}_p} - \upsilon_{k,p}\right)\mathbf{z}[\ell+p]e^{-2\pi i(\ell+p)k/N}, \tag{93}$$

where $\upsilon_{k,p} := \sqrt{\mathbf{A}[p,k]}\frac{\mathbf{f}_k^H\mathbf{g}_p}{\varphi_\mu\left(|\mathbf{f}_k^H\mathbf{g}_p|\right)}$, and

$$\mathbf{g}_p := \left[\mathbf{z}[0]\overline{\mathbf{z}[p]}, \cdots, \mathbf{z}[N-1]\overline{\mathbf{z}[N-1+p]}\right]^T. \tag{94}$$

The gradient $\mathbf{d}^{(t)}$ is then given by

$$\mathbf{d}^{(t)} := \left[\frac{\partial h(\mathbf{x}^{(t)},\mu)}{\partial\overline{\mathbf{x}^{(t)}[0]}}, \cdots, \frac{\partial h(\mathbf{x}^{(t)},\mu)}{\partial\overline{\mathbf{x}^{(t)}[N-1]}}\right]^T. \tag{95}$$

To alleviate the memory requirements and computational complexity required for large N, we suggest a block stochastic gradient descent strategy. Instead of calculating (93), we choose only a random subset of the sum for each iteration t, that is,

$$\begin{aligned}\mathbf{d}_{\Gamma_{(t)}}[\ell] = &\sum_{p,k\in\Gamma_{(t)}} \left(\mathbf{f}_k^H \mathbf{g}_p^{(t)} - \upsilon_{k,p,t}\right) \mathbf{z}^{(t)}[\ell - p] e^{2\pi i \ell k/N} \\ &+ \sum_{p,k\in\Gamma_{(t)}} \left(\mathbf{f}_k^T \overline{\mathbf{g}_p}^{(t)} - \upsilon_{k,p}\right) \mathbf{z}^{(t)}[\ell + p] e^{-2\pi i (\ell+p) k/N}, \end{aligned} \tag{96}$$

where the set $\Gamma_{(t)}$ is chosen uniformly and independently at random at each iteration t from subsets of $\{1,\cdots,N\}^2$ with cardinality Q. Specifically, the gradient in (95) is uniformly sampled using a minibatch of data, in this case of size Q for each step update, such that in expectation is (93) (Spall, 2005, page 130).

As mentioned in Section 7.2, choosing $\mu > 0$ prevents bias in the update direction. Since the function h is smooth, we are able to construct a descent rule for μ (Line 13 of Algorithm 7) in order to guarantee convergence to a first-order optimal point, that is, a point with zero gradient, in the vicinity of the solution.

Theorem 7.2.1. Let $\mathbf{x}$ be S-time-limited or B-band-limited for some $S \leq N/2$ or $B \leq N/2$, respectively, satisfying $\text{dist}(\mathbf{x}, \mathbf{x}^{(t)}) \leq \rho$ for some sufficiently small constant $\rho > 0$. Suppose that $\Gamma_{(t)}$ is sampled uniformly at random from all subsets of $\{1,\cdots,N\}^2$ with cardinality Q, independently

Algorithm 7

Input: Data $\{\mathbf{A}[p,k] : k,p = 0,\cdots,N-1\}$. Choose constants $\gamma_1,\gamma,\alpha \in (0,1)$, $\mu^{(0)} \geq 0$, cardinality $Q \in \{1,\cdots,N^2\}$, and tolerance $\varepsilon > 0$.

Initial point $\mathbf{x}^{(0)} \leftarrow$ Algorithm 2$(\mathbf{A}[p,k],T)$.

while $\left\|\mathbf{b}_{\Gamma_{(t)}}\right\|_2 \geq \varepsilon$ **do**

Choose $\Gamma_{(t)}$ uniformly at random from the subsets of $\{1,\cdots,N\}^2$ with cardinality Q per iteration $t \geq 0$.

$$\mathbf{x}^{(t+1)} = \mathbf{x}^{(t)} + \alpha^{(t)}\mathbf{b}_{\Gamma_{(t)}} = \mathbf{x}^{(t)} - \alpha^{(t)}\frac{\mu^{(t)}}{\|\mathbf{d}_{\Gamma_{(t)}}\|_2}\mathbf{d}_{\Gamma_{(t)}},$$

where

$$\begin{aligned}\mathbf{d}_{\Gamma_{(t)}}[\ell] = &\sum_{p,k\in\Gamma_{(t)}}\left(\mathbf{f}_k^H\mathbf{g}_p^{(t)} - \upsilon_{k,p,t}\right)\mathbf{x}^{(t)}[\ell-p]e^{2\pi i\ell k/N}\\ &+\sum_{p,k\in\Gamma_{(t)}}\left(\mathbf{f}_k^T\overline{\mathbf{g}_p}^{(t)} - \upsilon_{k,p}\right)\mathbf{x}^{(t)}[\ell+p]e^{-2\pi i(\ell+p)k/N}\end{aligned}$$

$$\upsilon_{k,p,t} = \sqrt{\mathbf{A}[p,k]}\frac{\mathbf{f}_k^H\mathbf{g}_p^{(t)}}{\varphi_{\mu^{(t)}}\left(\left|\mathbf{f}_k^H\mathbf{g}_p^{(t)}\right|\right)}.$$

$$\mathbf{g}_p^{(t)} = \left[\mathbf{x}^{(t)}[0]\overline{\mathbf{x}[p]}^{(t)},\cdots,\mathbf{x}^{(t)}[N-1]\overline{\mathbf{x}[N-1+p]}^{(t)}\right]^T.$$

if $\left\|\mathbf{d}_{\Gamma_{(t)}}\right\|_2 \geq \gamma\mu^{(t)}$ **then**

$\mu^{(t+1)} = \mu^{(t)}$.

else

$\mu^{(t+1)} = \gamma_1\mu^{(t)}$.

end if

end while

return: $\mathbf{x}^{(T)}$.

for each iteration. Then for almost all signals, Algorithm 7 with step size $\alpha \in (0, \frac{2}{U}]$ satisfies

$$\lim_{t\to\infty} \mu^{(t)} = 0, \text{ and } \lim_{t\to\infty} \left\| \mathbf{d}^{(t)} \right\|_2 = 0, \tag{97}$$

for some constant $U > 0$ depending on ρ.

Demostración. See Appendix 8. □

7.3. Initialization Algorithm

In this section we devise a method to initialize the gradient iterations. This strategy approximates the signal $\mathbf{x}$ from the ambiguity function as the leading eigenvector of a carefully designed matrix.

Instead of directly dealing with the ambiguity function in (85), we consider the acquired data in a transformed domain by taking its 1D DFT with respect to the frequency variable (normalized by $1/N$). Our measurement model is then

$$\begin{aligned}\mathbf{Y}[p,\ell] &= \frac{1}{N}\sum_{k=0}^{N-1} \mathbf{A}[p,k]e^{-2\pi ik\ell/N} = \frac{1}{N}\sum_{k,n,m=0}^{N-1} \mathbf{x}[n]\overline{\mathbf{x}[n-p]}\mathbf{x}[m-p]\overline{\mathbf{x}[m]}e^{-2\pi ik\frac{(m-n-\ell)}{N}} \\ &= \sum_{n=0}^{N-1} \mathbf{x}[n]\overline{\mathbf{x}[n-p]}\mathbf{x}[n+\ell-p]\overline{\mathbf{x}[n+\ell]},\end{aligned} \tag{98}$$

where $p, \ell = 0, \cdots, N-1$. Observe that for fixed p, $\mathbf{Y}[p,\ell]$ is the autocorrelation of $\mathbf{x} \odot \overline{\mathbf{x}_p}$, where $\mathbf{x}_p[n] = \mathbf{x}[n-p]$.

Let $\mathbf{D}_p \in \mathbb{C}^{N\times N}$ be a diagonal matrix composed of the entries of $\mathbf{x}_p$, and let $\mathbf{C}_\ell$ be a circulant matrix that shifts the entries of a vector by ℓ locations, namely, $(\mathbf{C}_\ell \mathbf{x})[n] = \mathbf{x}[n+\ell]$. Then, the matrix

$\mathbf{X} := \mathbf{x}\mathbf{x}^H$ is linearly mapped to $\mathbf{Y}[p,\ell]$ as follows:

$$\begin{aligned}\mathbf{Y}[p,\ell] &= \left(\overline{\mathbf{D}}_{p+\ell}\mathbf{D}_p\mathbf{C}_\ell\mathbf{x}\right)^H\mathbf{x} = \mathbf{x}^H\mathbf{A}_{p,\ell}\mathbf{x} \\ &= tr(\mathbf{X}\mathbf{A}_{p,\ell}),\end{aligned} \tag{99}$$

where $\mathbf{A}_{p,\ell} = \mathbf{C}_{-\ell}\overline{\mathbf{D}}_p\mathbf{D}_{p+\ell}$, and $tr(\cdot)$ denotes the trace function. Observe that $\mathbf{C}_\ell^T = \mathbf{C}_{-\ell}$. Thus, we have that

$$\mathbf{y}_\ell = \mathbf{G}_\ell\mathbf{x}_\ell, \tag{100}$$

for a fixed $\ell \in \{0,\cdots,N-1\}$, where $\mathbf{y}_\ell[n] = \mathbf{Y}[n,\ell]$ and $\mathbf{x}_\ell = \text{diag}(\mathbf{X},\ell)$. The (p,n)th entry of the matrix $\mathbf{G}_\ell \in \mathbb{C}^{\lceil\frac{N}{L}\rceil\times N}$ is given by

$$\mathbf{G}_\ell[p,n] := \overline{\mathbf{x}_p[n]}\mathbf{x}_p[n+\ell]. \tag{101}$$

From (101) it follows that $\mathbf{G}_\ell$ is a circulant matrix. Therefore, $\mathbf{G}_\ell$ is invertible if and only if the DFT of its first column, in this case $\overline{\mathbf{x}}\odot(\mathbf{C}_\ell\mathbf{x})$, is non-vanishing.

Using (100), we propose a method to estimate the signal $\mathbf{x}$ from measurements (85) using an alternating scheme: fixing $\mathbf{G}_\ell$, solving for $\mathbf{x}_\ell$, updating $\mathbf{G}_\ell$ and so forth.

We start the alternating scheme with the initial point

$$\mathbf{x}_{init}[p] := \mathbf{v}[p]\exp(i\theta[p]), \tag{102}$$

where $\theta[r] \in [0, 2\pi)$ is chosen uniformly at random for all $r \in \{0, \cdots, N-1\}$. The rth entry of $\mathbf{v}$ corresponds to the summation of the measured ambiguity function over the frequency axis:

$$\mathbf{v}[p] := \frac{1}{N}\sum_{k=0}^{N-1}\mathbf{A}[p,k] = \sum_{k=0}^{N-1}\left|\sum_{n=0}^{N-1}\mathbf{x}[n]\overline{\mathbf{x}[n-p]}e^{-2\pi ink/N}\right|^2 := \sum_{n=0}^{N-1}|\mathbf{x}[n]|^2|\mathbf{x}[n-p]|^2. \tag{103}$$

Once the vector $\mathbf{x}_{init}$ is constructed, the vectors $\mathbf{x}_\ell^{(t)}$ at $t = 0$ can be built as

$$\mathbf{x}_\ell^{(0)} = \mathrm{diag}(\mathbf{X}_0^{(0)}, \ell), \tag{104}$$

where

$$\mathbf{X}_0^{(0)} = \mathbf{x}_{init}\mathbf{x}_{init}^H. \tag{105}$$

Then, from (104) we proceed with an alternating procedure between estimating the matrix $\mathbf{G}_\ell$, and updating the vector $\mathbf{x}_\ell$ as follows.

- *Update rule for* $\mathbf{G}_\ell$*:* In order to update $\mathbf{G}_\ell$, we update the matrix $\mathbf{X}_0^{(t)}$ as

$$\mathrm{diag}(\mathbf{X}_0^{(t)}, \ell) = \mathbf{x}_\ell^{(t)}. \tag{106}$$

Observe that if $\mathbf{x}_\ell^{(t)}$ is close to $\mathbf{x}_\ell$ for all ℓ, then $\mathbf{X}_0^{(t)}$ is close to $\mathbf{x}\mathbf{x}^H$. Letting $\mathbf{w}^{(t)}$ be the leading (unit-norm) eigenvector of the matrix $\mathbf{X}_0^{(t)}$ constructed in (106), from (101) each matrix $\mathbf{G}_\ell^{(t)}$ at iteration t is given by

$$\mathbf{G}_\ell^{(t)}[p,n] = \overline{\mathbf{x}_p}^{(t)}[n]\mathbf{x}_p^{(t)}[n+\ell], \tag{107}$$

where $\mathbf{x}_p^{(t)}[n] = \mathbf{w}^{(t)}[n-p]$.

- *Optimization with respect to* $\mathbf{x}_\ell$*:* Fixing $\mathbf{G}_\ell^{(t-1)}$, one can estimate $\mathbf{x}_\ell^{(t)}$ at iteration t by solving the linear least-squares (LS) problem

$$\min_{\mathbf{p}_\ell \in \mathbb{C}^N} \quad \|\mathbf{y}_\ell - \mathbf{G}_\ell^{(t-1)}\mathbf{p}_\ell\|_2^2. \tag{108}$$

The relationship between the vectors $\mathbf{x}_\ell^{(t)}$ is ignored at this stage. If $\mathbf{G}_\ell^{(t-1)}$ is invertible, then the solution to this problem is given by $(\mathbf{G}_\ell^{(t-1)})^{-1}\mathbf{y}_\ell$. Since $\mathbf{G}_\ell^{(t-1)}$ is a circulant matrix, it is invertible if and only if the DFT of $\overline{\mathbf{x}}^{(t-1)} \odot (\mathbf{C}_\ell \mathbf{x}^{(t-1)})$ is non-vanishing. This condition cannot be ensured in general. Thus, we propose a surrogate proximal optimization problem to estimate $\mathbf{x}_\ell^{(t)}$ by

$$\min_{\mathbf{p}_\ell \in \mathbb{C}^N} \quad \|\mathbf{y}_\ell - \mathbf{G}_\ell^{(t-1)}\mathbf{p}_\ell\|_2^2 + \frac{1}{2\lambda_{(t)}}\|\mathbf{p}_\ell - \mathbf{x}_\ell^{(t-1)}\|_2^2, \tag{109}$$

where $\lambda_{(t)} > 0$ is a regularization parameter. In practice $\lambda_{(t)}$ is a tunable parameter Parikh

and Boyd (2014). In particular, for this work the value of $\lambda_{(t)}$ was determined using a cross-validation strategy such that each simulation uses the value that results in the smallest relative error according to (86). The surrogate optimization problem in (109) is strongly convex Parikh and Boyd (2014), and admits the following closed form solution

$$\mathbf{x}_\ell^{(t)} = \mathbf{B}_{\ell,t}^{-1}\mathbf{e}_{\ell,t}, \tag{110}$$

where

$$\begin{aligned}\mathbf{B}_{\ell,t} &= \left(\mathbf{G}_\ell^{(t-1)}\right)^H \left(\mathbf{G}_\ell^{(t-1)}\right) + \frac{1}{2\lambda}\mathbf{I},\\ \mathbf{e}_{\ell,t} &= \left(\mathbf{G}_\ell^{(t)}\right)^H \mathbf{y}_\ell + \frac{1}{2\lambda_{(t)}}\mathbf{x}_\ell^{(t-1)},\end{aligned} \tag{111}$$

with $\mathbf{I} \in \mathbb{R}^{N\times N}$ the identity matrix. Clearly $\mathbf{B}_{\ell,t}$ in (111) is always invertible. The update step for each $\mathbf{x}_\ell^{(t)}$ is computed in Line 9 of Algorithm 9.

Finally, in order to estimate $\mathbf{x}$, the (unit-norm) principal eigenvector of $\mathbf{X}_0^{(T)}$ is normalized by

$$\beta = \sqrt[4]{\sum_{n\in\mathscr{S}} \left(\mathbf{B}_{0,T}^{-1}\mathbf{e}_{0,T}\right)[n]}, \tag{112}$$

where $\mathscr{S} := \left\{n : \left(\mathbf{B}_{0,T}^{-1}\mathbf{e}_{0,T}\right)[n] > 0\right\}$. Observe that (112) results from the fact that $\sum_{n=0}^{N-1} \operatorname{diag}(\mathbf{X},0)[n] = \|\mathbf{x}\|_2^4$.

Algorithm 8

Input: The measurements $\mathbf{A}[p,k]$, T the number of iterations, and $\lambda > 0$.
Output: $\mathbf{x}^{(0)}$ (estimation of $\mathbf{x}$).
Initialize: $\mathbf{x}_{init}[p] = \mathbf{v}[p]\exp(i\theta[p])$, and $\mathbf{v}[p] = \dfrac{1}{N}\displaystyle\sum_{k=0}^{N-1}\mathbf{A}[p,k]$, $\theta[p] \in [0, 2\pi)$ is chosen uniformly and independently at random.
Compute $\mathbf{Y}[p,\ell]$ the 1D inverse DFT with respect to k of $\mathbf{A}[p,k]$.
for $t = 1$ to T **do**
    Construct $\mathbf{G}_\ell^{(t)}$ according to (107).
    Compute $\mathbf{B}_{\ell,t} = (\mathbf{G}_\ell^{(t)})^H(\mathbf{G}_\ell^{(t)}) + \frac{1}{2\lambda}\mathbf{I}$.
    Compute $\mathbf{e}_{\ell,t} = (\mathbf{G}_\ell^{(t)})^H\mathbf{y}_\ell + \frac{1}{2\lambda}\mathbf{x}_\ell^{(t-1)}$.
    Construct the matrix $\mathbf{X}_0^{(t)}$ such that

$$\operatorname{diag}(\mathbf{X}_0^{(t)}, \ell) = \mathbf{B}_{\ell,t}^{-1}\mathbf{e}_{\ell,t}, \quad \ell = 0, \cdots, N-1.$$

    Let $\mathbf{w}^{(t)}$ be the leading (unit-norm) eigenvector of $\mathbf{X}_0^{(t)}$.
    Take $\mathbf{x}_p^{(t)}[n] = \mathbf{w}^{(t)}[n-p]$.
end for
Compute vector $\mathbf{x}^{(0)}$ as

$$\mathbf{x}^{(0)} := \sqrt[4]{\sum_{n\in\mathscr{S}}\left(\mathbf{B}_{0,T}^{-1}\mathbf{e}_{0,T}\right)[n]}\,\mathbf{w}^{(T)},$$

where $\mathscr{S} := \left\{n : \left(\mathbf{B}_{0,T}^{-1}\mathbf{e}_{0,T}\right)[n] > 0\right\}$.
return: $\mathbf{x}^{(0)}$.

After a few iterations of this two-step procedure, the output is used to initialize the gradient algorithm described in Section 7.2. This alternating scheme is summarized in Algorithm 9.

8. Extension to Detection Tasks

Template matching (TM) is a common methodology for target detection (TD) which allows detecting a target based on cross-correlation analysis between a reference pattern and the scene. State-of-the-art TD approaches do not consider the optical phase of the target as a discriminant in the detection process, because to recover the phase involves solving a computationally deman-

ding inverse problem known as phase retrieval (PR). However, in applications such as microscopy and optical imaging, the optical phase contains valuable information that describes the shape and depth of the object. This work proposes a method for fast TD via TM, which considers the optical phase of the object in the reference pattern as a discriminant in a setup that records coded diffraction patterns (CDP). Specifically, the proposed TD methodology is established for far-field imaging. This approach consists of two steps: (i) fast approximation of the optical field from CDP based on compressive PR, including its optical phase information; (ii) cross-correlation analysis to detect the target using its optical phase. The approximation of the optical field considering its phase is performed by low-pass-filtering the leading eigenvector of a designed matrix, overcoming traditional approaches in terms of relative error. Since no explicit TD methodology that includes the optical phase as a discriminant exists in the literature, the proposed approach is compared to a method that reconstructs the optical field and then performs the detection step. Numerical results suggest that the proposed methodology detects a target under noisy scenarios using up to 75 % fewer measurements in the tested datasets. Also, the proposed TD using the filtered spectral method reduces the detection time in up to 79 % in the tested datasets, compared to a methodology that requires the reconstruction of the phase.

8.1. Target Detection Methodology from CDP

This section describes the proposed TD methodology composed by two stages: (i) a fast optical field approximation strategy of the phase from CDP based on compressive PR literature introduced in Section 8.1.1, and (ii) a TD procedure via TM that includes the optical phase information, as explained in Section 8.1.2. Fig. 7 summarizes the proposed TD approach.

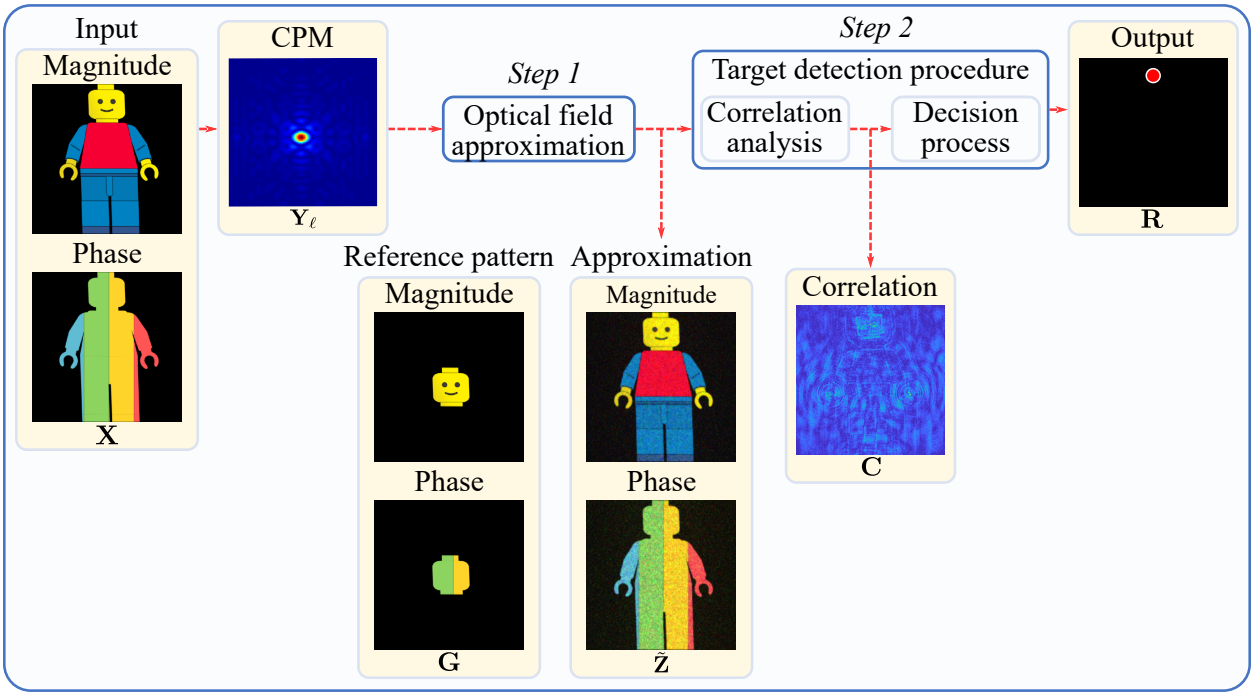

Figura 7. Flowchart of the proposed TD methodology using CDP. For the first step, the approximation is performed by low-pass-filtering the leading eigenvector of a designed matrix. For the second step, cross-correlation analysis is used to detect the target from its optical phase.

8.1.1. Step1: Fast Optical Field Approximation. In this section, the proposed phase approximation strategy is presented. It exploits the mathematical model of CDP in (2) and the sparsity property of natural scenes in the Fourier domain. Specifically, from the compressive PR imaging literature, it is known that an image can be accurately represented using a few coefficients in the Fourier domain Jensen and Lulla (1987). This implies that $\|\mathbf{F}\mathbf{x} - \theta\|_2 < \varepsilon$ for some small constant $\varepsilon > 0$ where $\|\theta\|_0 = s \ll n$ and $\|\cdot\|_0$ represents the ℓ_0 pseudo-norm that returns the number of non-zero elements (support) of a given vector. Considering this sparsity prior over $\mathbf{x}$, and the model in (2), we know that from the compressive sensing theory, the support of θ can be estimated from the

CDP Wang et al. (2017b). Specifically, define

$$\hat{\upsilon}_p = \frac{1}{m}\sum_{i=1}^{m}(\mathbf{y})_i|(\mathbf{B})_{i,p}|^2,\ 1 \leq p \leq n, \tag{113}$$

where $\mathbf{B} = \mathbf{AF}$, and the expected value of the random variable $\hat{\upsilon}_p$ is given by

$$\mathbb{E}[\hat{\upsilon}_p] \geq c_1\|\mathbf{x}\|_2^2 + c_2|(\theta)_p|^2 + c_3, \tag{114}$$

where c_1, c_2 and c_3 are constants. Then, given the fact that θ is sparse, it is clear that as long as the constant c_2 is sufficiently large, the non-zero coefficients of θ can be exactly recovered. In fact, appealing to the strong law of large numbers, the sample average, namely $\hat{\upsilon}_p \to \mathbb{E}[\hat{\upsilon}_p]$ as m increases, approaches the support of θ and is estimated as

$$\mathscr{S} := \left\{1 \leq p \leq n | \text{indices of top-s intances in } \{\hat{\upsilon}_p\}_{p=1}^{n}\right\}. \tag{115}$$

In summary, Lemma 8.1.1 in Wang et al. (2017b), theoretically states that (115) is able to recover the support of θ with high probability.

Lemma 8.1.1. ((Wang et al., 2017b, Lemma 1)) Consider any signal $\mathbf{x} \in \mathbb{C}^n$ with a s-sparse representation $\theta \in \mathbb{C}^n$ in the Fourier domain. Then, (115) recovers the support of θ with probability at least $1 - 6/m$ provided that $m \geq \kappa s^2 \log(mn)$ for some constant $\kappa > 0$.

Once the set $\mathscr{S}$ is estimated following (115), the non-zero entries of θ are approximated

solving the following optimization problem Wang et al. (2017b)

$$\hat{\theta}_{\mathscr{S}} = \underset{\|\theta_{\mathscr{S}}\|_2=1}{\arg\max} \ \theta_{\mathscr{S}}^H \left(\frac{1}{|\mathscr{I}_0|} \sum_{i \in \mathscr{I}_0} \frac{\mathbf{b}_{i,\mathscr{S}} \mathbf{b}_{i,\mathscr{S}}^H}{\|\mathbf{b}_{i,\mathscr{S}}\|_2^2} \right) \theta_{\mathscr{S}}, \tag{116}$$

where $\mathbf{b}_{i,\mathscr{S}}$ is the i-th row of matrix $\mathbf{B}$ which includes the p-th entry $(\mathbf{b}_i)_p$ of $\mathbf{b}_i$ if and only if $p \in \mathscr{S}$. Likewise, for $\theta_{\mathscr{S}}, \hat{\theta}_{\mathscr{S}} \in \mathbb{C}^s$, and $\mathscr{I}_0 \subset \{1, \ldots, nL\}$ is the collection of indices corresponding to the $\lfloor m/6 \rfloor$ largest values of $\{(\mathbf{y})_i / \|\mathbf{a}_i\|_2\}$ Guerrero et al. (2020). The optimization problem in (116) mathematically involves the computation of the leading eigenvector of the matrix $\mathbf{G}_0 := \frac{1}{|\mathscr{I}_0|} \sum_{i \in \mathscr{I}_0} \frac{\mathbf{b}_{i,\mathscr{S}} \mathbf{b}_{i,\mathscr{S}}^H}{\|\mathbf{b}_{i,\mathscr{S}}\|_2^2}$ Wang et al. (2017b). Usually, (116) is numerically solved via the power iteration method Saad (2003); Wang et al. (2018b,a). This method consists in recursively performing a matrix-vector multiplication between $\mathbf{G}_0$ and the iterative approximation of the optical field Wang et al. (2018b). Subsequently, a s-sparse n-dimensional approximation $\hat{\theta}$ is obtained by zero-padding $\hat{\theta}_{\mathscr{S}}$ at entries with indices not belonging to $\mathscr{S}$. Thus, since $\hat{\theta}$ is a sparse approximation of $\mathbf{x}$ in the Fourier domain, $\hat{\mathbf{z}} = \mathbf{F}^H \hat{\theta}$ approximates the optical field $\mathbf{x}$. It is worth mentioning that $\hat{\mathbf{z}}$ is a complex vector that approximates both the magnitude and phase of the optical field $\mathbf{x}$. Also, note that sparse PR requires at least $\mathscr{O}(s \log(n/s))$ measurements as in compressive sensing literature Wang et al. (2017b). In summary, Theorem 8.1.2 in Wang et al. (2017b), states that $\hat{\mathbf{z}}$ is a close approximation of $\mathbf{x}$ with high probability.

Theorem 8.1.2. ((Wang et al., 2017b, Theorem 1)) Consider noisy measurements $(\mathbf{y})_i = |\langle \mathbf{a}_i, \mathbf{x} \rangle|^2 + (\omega)_i$ such that $\|\omega\|_\infty \leq c\|\mathbf{x}\|_\infty$ for some $c > 0$. If the set of coded apertures satisfies (12), then with

high probability the vector $\hat{\mathbf{z}}$ as the solution of (116) satisfies

$$\text{dist}(\hat{\mathbf{z}}, \mathbf{x}) \leq \delta \|\mathbf{x}\|_2 + \mathscr{O}(\|\boldsymbol{\omega}\|_\infty)^1, \tag{117}$$

for some constant $\delta \in (0,1)$, provided that $m \geq \kappa s$, for $\kappa > 0$.

Considering that a fast approximation of the optical field from CDP including its phase information without full reconstruction time is desired, the main drawback of the above state-of-the-art strategy is the computational complexity to perform (115). To alleviate this limitation, we alternatively propose to solve the following optimization problem

$$\begin{aligned} \hat{\theta} = \quad & \underset{\|\theta\|_2=1}{\arg\max} \ \theta^H \left(\frac{1}{|\mathscr{I}_0|} \sum_{i \in \mathscr{I}_0} \frac{\mathbf{b}_i \mathbf{b}_i^H}{\|\mathbf{b}_i\|_2^2} \right) \theta \\ \text{s.t} \quad & \|\theta\|_1 \leq \tau, \end{aligned} \tag{118}$$

for some $\tau > 0$. Observe that (118) instead of hardly removing those zero-frequencies that can be identified using (115) as (116) does, (118) relaxes the sparsity assumption introducing the ℓ_1-norm. This alternative optimization problem is motivated by the fact that the complexity to estimate the non-zero frequencies in (115) is $\mathscr{O}(n^2)$, since $\hat{\upsilon}_p$ in (113) is obtained performing matrix-vector multiplications. Also, solving the ℓ_1 constraint in (118) is less computationally expensive than computing (115) as it will be discussed in brief. To numerically solve (118), the *spectral filtered*

---

[1] The notation $\varphi(w) = \mathscr{O}(g(w))$ means there exists a numerical constant $c > 0$ such that $\varphi(w) \leq c g(w)$.

Algorithm 9

Input: Acquired data $\{(\mathbf{a}_i;(\mathbf{y})_i)\}_{i=1}^{m}$, maximum number of iterations T, and low-pass filter $\mathscr{G}$.
$\tilde{\mathbf{z}}^{(0)} \leftarrow$ Chosen randomly.
Set $\mathscr{I}_0$ as the set of indices corresponding to the $\lfloor m/6 \rfloor$ largest values of $\{(\mathbf{y})_i/\|\mathbf{a}_i\|_2\}$.

$$\mathbf{Y}_0 := \frac{1}{|\mathscr{I}_0|}\sum_{i\in\mathscr{I}_0}\frac{\mathbf{a}_i\mathbf{a}_i^H}{\|\mathbf{a}_i\|_2^2}$$

for $t = 0 : T-1$ **do**
    $\grave{\mathbf{z}}^{(t+1)} \leftarrow \mathscr{G}\left(\mathbf{Y}_0\tilde{\mathbf{z}}^{(t)}\right)$
    $\tilde{\mathbf{z}}^{(t+1)} \leftarrow \frac{\grave{\mathbf{z}}^{(t+1)}}{\|\grave{\mathbf{z}}^{(t+1)}\|_2}$
end forend for
Compute $\hat{\mathbf{z}} = \sqrt{\frac{\sum_{i=1}^{m}(\mathbf{y})_i}{m}}\tilde{\mathbf{z}}^{(T)}$
Return: $\hat{\mathbf{z}}$

method is introduced as summarized in Algorithm 9. This algorithm follows a power iteration methodology and reduces the computational complexity of estimating (115), employing a low-pass filter which allows to solve the inequality constraint in (118). Mathematically, the effect of this filter is the attenuation of the high-frequencies of the optical field in the Fourier domain. Additionally, it is well-known that the filtering process can be rapidly performed through the fast Fourier transform

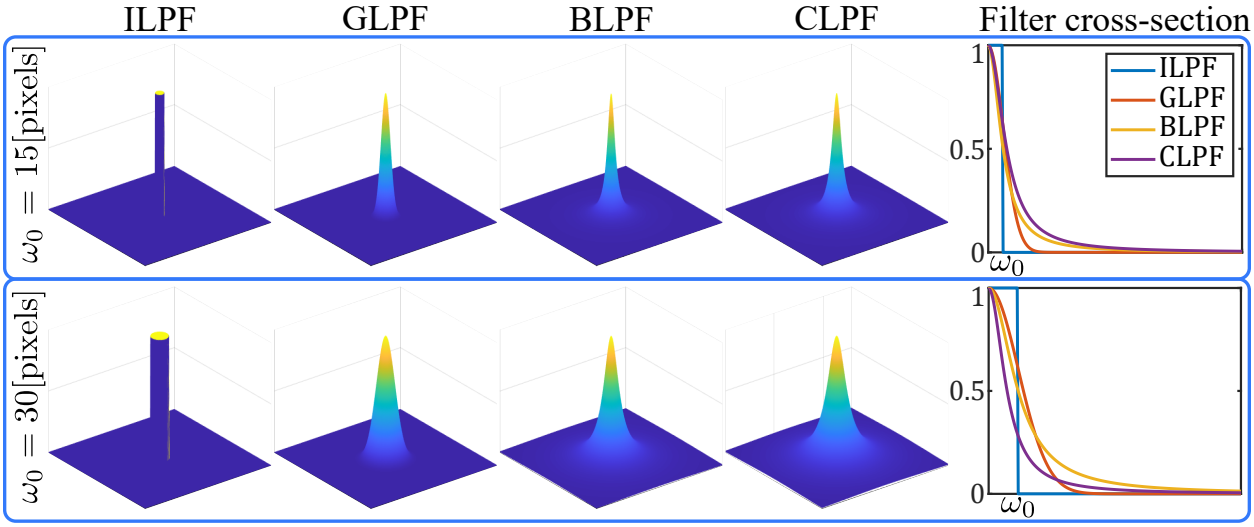

Figura 8. Sketch of different low-pass filters with cutoff frequency $\omega_0 \in \{15, 30\}$[pixels].

with a computational complexity $\mathscr{O}(n\log(n))$ Gonzalez and Wintz (1977), which is substantially lower than $\mathscr{O}(n^2)$. Notice that Algorithm 9 requires the sampling vectors, the acquired CDP and a low-pass filter $\mathscr{G}$. Among different filter types, e.g., ideal low-pass filter (ILPF), gaussian low-pass filter (GLPF), butterworth low-pass filter (BLPF), and chebyshev low-pass filter (CLPF) Gonzalez and Wintz (1977), this work employs a Gaussian filter with cutoff frequency $\omega_0 = 15$[pixels] to illustrate the effectiveness of Algorithm 9. Nevertheless, any other filter could be used. Fig. 8 illustrates the perspective plot of different low-pass filters using two different cutoff frequencies.

Following the iteration process, the characteristic matrix-vector multiplication of the power iteration method $\mathbf{Y}_0\tilde{\mathbf{z}}^{(t)}$ is performed in line 6. The result of this product is considered the current approximation of both the magnitude and phase of the optical field. Also, in line 6 a low-pass filtering process over $\mathbf{Y}_0\tilde{\mathbf{z}}^{(t)}$ is accomplished, where $\mathscr{G}$ represents the filter. The effect of iteratively applying $\mathscr{G}$ over the approximation of the image is the selection of those low-frequencies that sparsely represent the image in the Fourier domain. Observe that this selection is rapidly performed in comparison with (115) Gonzalez and Wintz (1977). Finally, Algorithm 9 returns the scaled complex vector $\hat{\mathbf{z}}$, which according to Theorem 8.1.2 is a close approximation of both the magnitude and phase of $\mathbf{x}$. The scaling factor $\sqrt{\frac{\sum_{i=1}^{m}(\mathbf{y})_i}{m}}$ in line 8 is a close approximation of $\|\mathbf{x}\|_2$ Wang et al. (2018b) and it has to be calculated because $\tilde{\mathbf{z}}$ is a unitary image.

To mathematically summarize the advantages of solving (118) compared to (116) to approximate the optical field $\mathbf{x}$, Theorem 8.1.3 is presented.

Theorem 8.1.3. Consider noisy measurements $(\mathbf{y})_i = |\langle \mathbf{a}_i, \mathbf{x}\rangle|^2 + (\boldsymbol{\omega})_i$, and a low-pass filter $\mathscr{G}$, such that $\|\omega\|_\infty \leq c\|\mathbf{x}\|_\infty$ for some $c > 0$. If the set of coded apertures satisfies (12), then with high

probability the vector $\hat{\mathbf{z}}$ returned by Algorithm 9 satisfies

$$\operatorname{dist}(\hat{\mathbf{z}}, \mathbf{x}) \leq \delta_{\mathscr{G}} \|\mathbf{x}\|_2 + \mathscr{O}(\|\boldsymbol{\omega}\|_\infty), \tag{119}$$

for some constant $\delta_{\mathscr{G}} \leq \delta < 1$ with δ as in (117), provided that $m \geq \kappa s$, for $\kappa > 0$.

Demostración. See Appendix A in the supplementary material. □

Notice that Theorem 8.1.3 guarantees that solving (118) via Algorithm 9 returns a more accurate approximation of $\mathbf{x}$ than that obtained by solving (116), since $\delta_{\mathscr{G}} \leq \delta$, for both noisy and noiseless scenarios. This advantage comes from the fact that the low-pass filter promotes a more accurate representation of $\mathbf{x}$ in the Fourier domain because it does not hardly remove any frequency. Additionally, since $\delta_{\mathscr{G}}$ depends on the chosen low-pass filter $\mathscr{G}$, it means that the accuracy of the approximation returned by Algorithm 9 is determined by $\mathscr{G}$. In fact, a filter $\mathscr{G}_1$ is able to better approximate the complex signal $\mathbf{x}$ compared to $\mathscr{G}_2$ if $\delta_{\mathscr{G}_1} < \delta_{\mathscr{G}_2}$. More details about how numerically compute $\delta_{\mathscr{G}}$ for a given filter $\mathscr{G}$ can be found in Appendix A in the supplementary material. Finally, note that (119) reveals that the amount of noise of the measurements $(\mathbf{y})_i$ is not affected by the low-pass filter. This is an expected result since $(\mathbf{y})_i$ does not intervene in the computation of the low-pass filtering step in Line 6 of Algorithm 9.

8.1.2. Step 2: Target Detection Procedure. This section describes a TD procedure following a template matching strategy that employs a circular harmonic filter (CHF) Prémont and Sheng (1993) to perform the detection. It is worth mentioning that the TM technique is invariant to rotations and scale-changes. In detail, the detection step is divided into two stages: (i) correlation

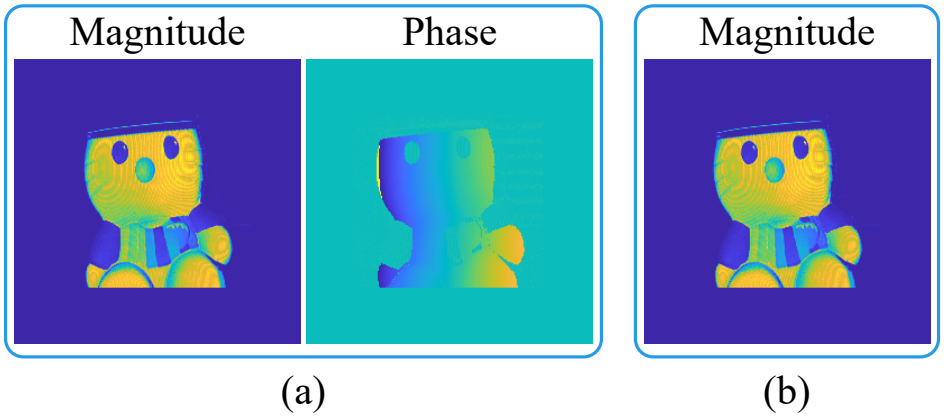

Figura 9. Example of two reference patterns. (a) Using both phase and magnitude information. (b) Using only magnitude information.

analysis based on CHF, and (ii) decision considering a thresholding procedure.

8.1.2.1. Cross-correlation Analysis. A circular filter using a reference pattern has to be designed to detect a target through cross-correlation. Cross-correlation is a metric commonly used in TM, which calculates the similarity between a CHF and a scene. Figure 9 illustrates two reference patterns, first using both phase and magnitude information, and second, magnitude-only information. It is worth mentioning that the magnitude-only reference pattern physically models a flat object, while a reference pattern as in Fig. 9(a) models a three-dimensional (3D) object. This implies that a TD methodology equipped with a complex reference pattern is able to differentiate between a flat and a 3D object.

In order to perform the detection, a CHF $\mathbf{H} \in \mathbb{C}^{n\times n}$, based on a reference pattern $\mathbf{G} \in \mathbb{C}^{n\times n}$ as in Fig. 9(a) is mathematically designed in polar coordinates on the Fourier domain Gualdron and Arsenault (1993). This system of coordinates is preferred in order to make the CHF invariant to rotations, such that, $\mathbf{H}$ is able to detect the object regardless any rotated version of it Prémont and Sheng (1993). Additionally, in order to build a filter $\mathbf{H}$ to be invariant to scale-changes sup-

pose that there are V different scale-changing patterns of one standard reference pattern Zi-Liang and Dalsgaard (1995), as in Fig. 9(a). Considering (ρ, ϕ) as the indexing variables of the polar coordinates, $\mathbf{H}$ is modeled as Prémont and Sheng (1993)

$$(\mathbf{H})_{\rho,\phi} = e^{2j\phi} \sum_{i=1}^{V} \sum_{l=1}^{n} \frac{(\mathbf{E}_i)_{\rho,(l-1)\Delta\phi}}{\left|(\mathbf{E}_i)_{\rho,(l-1)\Delta\phi}\right|} e^{-2j(l-1)\Delta\phi}, \tag{120}$$

where $\mathbf{E}_i \in \mathbb{C}^{n\times n}$ is given by

$$\mathbf{E}_i = \mathscr{F}(\mathbf{G}_i), \tag{121}$$

while $\mathbf{G}_i$ is a different scaled version of the standard reference pattern $\mathbf{G}$. We remark that (120) is a well-known model in the CHF literature where more details can be found in Zi-Liang and Dalsgaard (1995). In (120), the size constant $\Delta\phi$ allows to range the angular dimension of the polar coordinates Prémont and Sheng (1993) to make the CHF invariant to rotations. In practice, the value for $\Delta\phi$ is fixed as $\Delta\phi = \frac{\pi}{n}$. Thus, in order to accomplish the detection of the object, the correlation matrix $\mathbf{C} \in \mathbb{C}^{n\times n}$ is calculated between the CHF, $\mathbf{H}$, and the Fourier transform of the approximated optical field $\hat{\mathbf{Z}} \in \mathbb{C}^{n\times n}$, which is given by

$$\mathbf{C} = \mathscr{F}^{-1}\left(\overline{\mathbf{H}} \circ \hat{\mathbf{Z}}\right), \tag{122}$$

where $\hat{\mathbf{Z}} \in \mathbb{C}^{n\times n}$ is the Fourier transform of $\tilde{\mathbf{Z}}$, which is the matrix version of $\hat{\mathbf{z}}$ that represents the approximation of $\mathbf{x}$ obtained from Algorithm 9. The resultant matrix (122) is used in the following

section to determine the spatial location of the 3D object of interest.

8.1.2.2. Decision Process. Once the correlation matrix is calculated following (122), the target can be detected using a thresholding approach. Precisely, the threshold is defined as the maximum absolute value of the correlation matrix multiplied by a tolerance parameter $\varepsilon > 0$. Mathematically, the decision rule for a TD is given by

$$(\mathbf{R})_{u,v} = \begin{cases} 1, & \text{if } |(\mathbf{C})_{u,v}| \geq \varepsilon \cdot \max(\mathbf{C}) \\ 0, & \text{otherwise} \end{cases}, \tag{123}$$

where $(\mathbf{R})_{u,v} \in \{0,1\}$ represents the elements of the decision matrix, $\varepsilon \in (0,1]$ is a tolerance parameter and $\max(\cdot)$ is an operator that returns the element of a matrix with the largest magnitude value. In practice, ε is a tunable constant, which in this work is fixed as $\varepsilon = 0.9$. Thus, the object of interest is spatially located at an entry (u,v) if $(\mathbf{R})_{u,v} = 1$.

Algorithm 10

1: **Input:** data $\{(\mathbf{a}_i;(\mathbf{y})_i)\}_{i=1}^m$, the tolerance $\varepsilon > 0$, compute $\Delta\phi = \frac{\pi}{n}$, and the reference pattern $\mathbf{G}$.
2: $\hat{\mathbf{z}} \leftarrow$Algorithm 9$(\mathbf{a}_i;\mathbf{y})$.
3: $\mathbf{E} = \mathscr{F}(\mathbf{G})$
4: $(\mathbf{H})_{\rho,\phi} = e^{2j\phi}\sum_{l=1}^{n} \frac{(\mathbf{E})_{\rho,(l-1)\Delta\phi}}{|(\mathbf{E})_{\rho,(l-1)\Delta\phi}|} e^{-2j(l-1)\Delta\phi}$.
5: $\tilde{\mathbf{Z}} \leftarrow$ Matrix version of $\hat{\mathbf{z}}$.
6: Compute $\hat{\mathbf{Z}} = \mathscr{F}\{\tilde{\mathbf{Z}}\}$.
7: Compute $\mathbf{C} = \mathscr{F}^{-1}\{\overline{\mathbf{H}} \circ \hat{\mathbf{Z}}\}$.
8: Compute $(\mathbf{R})_{u,v} = \begin{cases} 1, & \text{if } |(\mathbf{C})_{u,v}| \geq \varepsilon \cdot \max(\mathbf{C}) \\ 0, & \text{otherwise} \end{cases}$
9: **Return: R**

To summarize the two-steps procedure, Algorithm 10 is introduced, which requires the ac-

quired CDP and the tolerance $\varepsilon > 0$. In line 2, the optical field is estimated from the phaseless measurements using Algorithm 9. Then, in line 3 the circular filter $\mathbf{H}$ is constructed from a reference pattern $\mathbf{G}$ following (120). The Fourier transform of the estimated optical field is calculated in line 4. In line 5, the correlation matrix is computed using (122). In line 6, a target is detected using the decision matrix described in Eq (123). Finally, the decision matrix is returned in line 7. The computational complexity of the detection procedure is $\mathscr{O}(n\log(n))$ according to the computed correlation in the Fourier domain.

It is worth highlighting that the above described detection algorithm is not able to detect a complex object (3D object) with its magnitude-only reference pattern. Mathematically, this issue is explained in the following. Suppose that in the pixels (u_1, v_1) and (u_2, v_2) of $\mathbf{X}$, a 3D object and its flat version are located, respectively. Define the correlation matrix $\mathbf{C}_a \in \mathbb{C}^{n\times n}$ as

$$\mathbf{C}_a = \mathscr{F}^{-1}\left(\overline{\mathbf{H}}_a \circ \hat{\mathbf{Z}}\right), \tag{124}$$

where the CHF, $\mathbf{H}_a$, of the magnitude-only reference pattern is constructed, according to (120), as

$$(\mathbf{H}_a)_{\rho,\phi} = e^{2j\phi}\sum_{i=1}^{V}\sum_{l=1}^{n}\frac{(\mathbf{E}_i^a)_{\rho,(l-1)\Delta\phi}}{\left|(\mathbf{E}_i^a)_{\rho,(l-1)\Delta\phi}\right|}e^{-2j(l-1)\Delta\phi}, \tag{125}$$

with $\mathbf{E}_i^a = \mathscr{F}(|\mathbf{G}_i|)$. Then, if the decision rule in (123) is applied over $\mathbf{C}_a$ the 3D object would not be detected, due to this object contains both phase and magnitude information. In mathematical

terms, the previous fact means that

$$|(\mathbf{C}_a)_{u_2,v_2}| > |(\mathbf{C}_a)_{u_1,v_1}|. \tag{126}$$

In fact, (126) is always valid since the cross-correlation between the CHF filter $\mathbf{H}_a$ and the $\hat{\mathbf{z}}$, will produce a higher magnitude at (u_2, v_2) Prémont and Sheng (1993), as it is theoretically stated in the following Lemma 8.1.4.

Lemma 8.1.4. Consider the CHF's $\mathbf{H}$ and $\mathbf{H}_a$ as modeled in (120), and (125), respectively. Suppose that a 3D object is located in the pixel (u_1, v_1) of the complex optical field. Then, $|(\mathbf{C})_{u_1,v_1}| > |(\mathbf{C}_a)_{u_1,v_1}|$ holds.

Demostración. See Appendix B in the supplementary material. □

Observe that Lemma 8.1.4 theoretically guarantees that the magnitude-only decision rule in (123) is enough to detect a 3D object. Finally, to complement the mathematical property in (126), and the result of Lemma 8.1.4, Section 9 numerically validates that Algorithm 10 effectively uses the optical phase as discriminant.

9. Numerical Results with Synthetic Data

In this chapter, the performance of the designed coded apertures for the three diffraction zones, and Algorithm 7 are evaluated. The used performance metric is

$$\text{relative error} := \frac{dist(\mathbf{z}, \mathbf{x})}{\|\mathbf{x}\|_2}$$

where $dist(\mathbf{z},\mathbf{x})$ is defined as

$$dist(\mathbf{z},\mathbf{x}) = \min_{\theta\in[0,2\pi)} \|\mathbf{x}e^{-j\theta} - \mathbf{z}\|_2, \quad j = \sqrt{-1}. \tag{127}$$

Five different tests are performed to analyze the effect of the coded apertures in the reconstruction quality. First, the initialization methodology is evaluated for designed and non-designed coded apertures. Second, some examples of reconstructed images using designed and non-designed coded apertures, based on admissible random variables, are shown. Third, the empirical success rate of some state-of-the-art reconstruction algorithms using designed and non-designed coded apertures is analyzed. The fourth experiment determines the robustness of designed coded apertures under noisy scenarios for different values of Signal-to-Noise-Ratio (SNR), defined as SNR$= 20\log_{10}(\|\mathbf{y}_k\|_2/\|\zeta\|_2)$, where ζ is the variance of the noise. Finally, under sparsity assumptions, the performance of different admissible random variables to estimate the non-zero coefficients of θ and to solve (28) is evaluated. Particularly, for this test, the average error over 100 tests was calculated. The admissible random variables tested are shown in Table 2.

Table 2
Admissible random variables used for simulations

Random Variable	**Coding Probability**	**Expected Value**	**Cardinality**
$d_1 = \{1,0\}$	$\{\frac{1}{2},\frac{1}{2}\}$	$\mathbb{E}[d_1] = \frac{1}{2}$	$\gamma^2 = 16$
$d_2 = \{1,0,j\}$	$\{\frac{1}{3},\frac{1}{3},\frac{1}{3}\}$	$\mathbb{E}[d_2] = \frac{1}{3} + \frac{1}{3}j$	$\gamma^2 = 9$
$d_3 = \{1,j,-j,-1\}$	$\{\frac{1}{4},\frac{1}{4},\frac{1}{4},\frac{1}{4}\}$	$\mathbb{E}[d_3] = 0$	$\gamma^2 = 16$

Note that the expected values of d_1, d_2 in Table 2 are non-zero. Specifically, these tested coding variables do not satisfy (3) and, in contrast to Table 1, these variables do not increase the

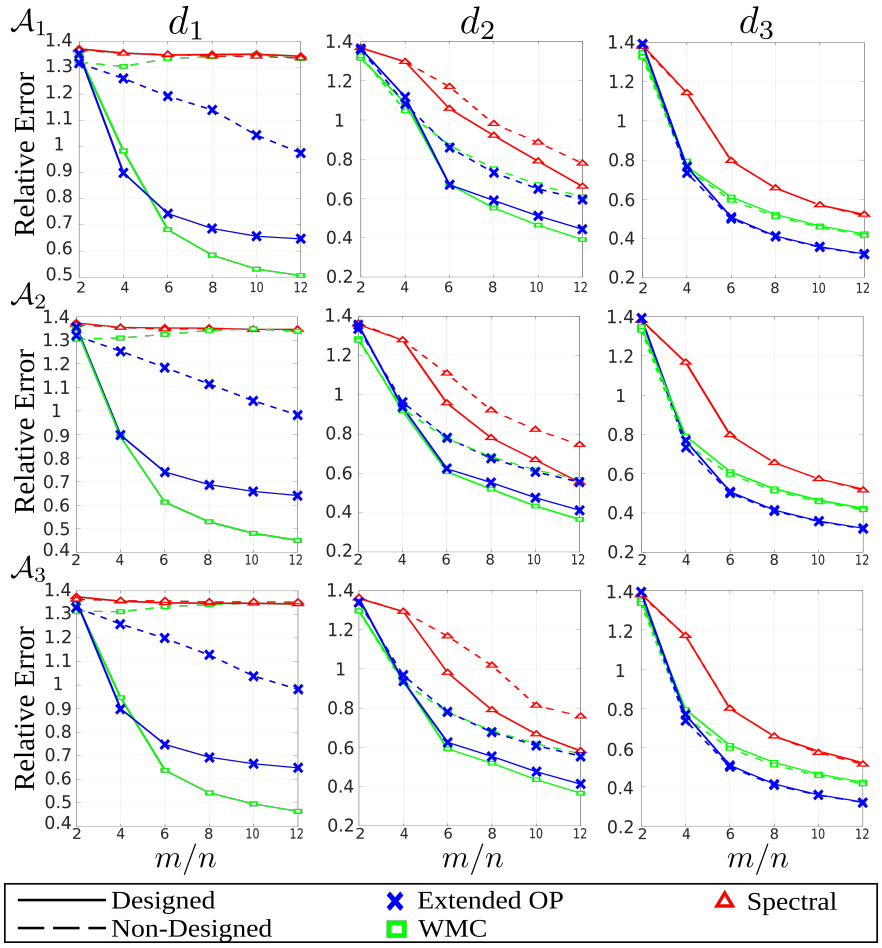

Figura 10. Performance of different initialization methods using designed coded apertures in terms of the relative error vs the number of projections. Rows: Diffraction zones, columns: admissible random variables.

power of the scene during the modulation process.

9.1. Designed Coded Aperture Analysis

9.1.1. Initialization Stage Performance. In order to evaluate the performance of designed coded apertures at the initialization stage, different state-of-the-art initializations, such as spectral, extended OPI and WMC are here employed the attained relative error is shown in Fig. 10. For this test $card(\mathscr{I}_0^c) = \lceil nL/2 \rceil$, while the number of projections is varied from $L = 2$ to $L = 12$. The scene is generated as a complex Gaussian random vector $\mathbf{x} \sim \mathscr{N}(0, \frac{1}{2}\mathbf{I}_n) + j\mathscr{N}(0, \frac{1}{2}\mathbf{I}_n)$. Figure 10

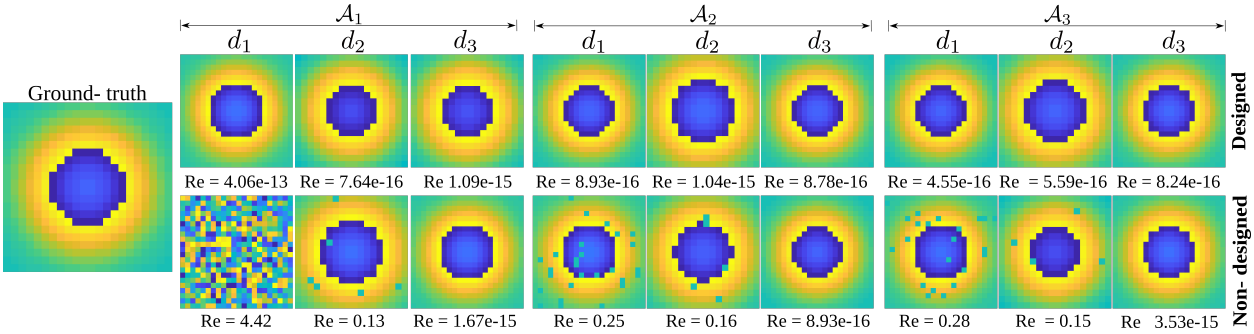

Figura 11. Reconstructed phase from CDP acquired at the different diffraction zones using the admissible random variables in Table 2 and $L = 4$ for designed and non-designed coded apertures.

shows that the designed coded apertures exhibit better performance for all the diffraction zones and all the initialization methodologies when compared to non-designed ensembles.

9.1.2. Reconstructions. To illustrate the attained reconstructions for each diffraction zone, using the admissible random variables listed in Table 2, the complex image of size $n = 100 \times 100$, illustrated in Fig. 11, was used as the ground truth. All the reconstructions are performed using the PRSF method in Pinilla et al. (2018a), with designed and non-designed coded apertures, fixing $L = 4$. Note that the designed coded apertures are able to better estimate the image for all diffraction zones than non-designed structures. This experiment corresponds to the reconstructed phase of the image for the three diffraction zones.

9.1.3. Sampling Complexity. Experiments are conducted to determine the empirical success rate of the TWF, PRSF and RAF reconstruction methods, using designed and non-designed coded apertures, as shown in Fig. 12. The scene is generated as a complex Gaussian random vector $\mathbf{x} \sim \mathcal{N}(0, \frac{1}{2}\mathbf{I}_n) + j\mathcal{N}(0, \frac{1}{2}\mathbf{I}_n)$ where $n = 100 \times 100$. In addition, it is established that a trial is successful when the returned relative error is less than 10^{-5}. The tested methods require up to 40 % less number of measurements to recover the image using designed coded apertures for the three

diffraction zones, attaining a success rate of 100% when $m/n = 4$ using the PRSF reconstruction algorithm. Also, combining Theorem 3.1.1 and the attained results in Fig. 12 the constant c_0 for these experiments is on the order of 10^{-3}, meaning that the number of projections L required to retrieve the phase is limited.

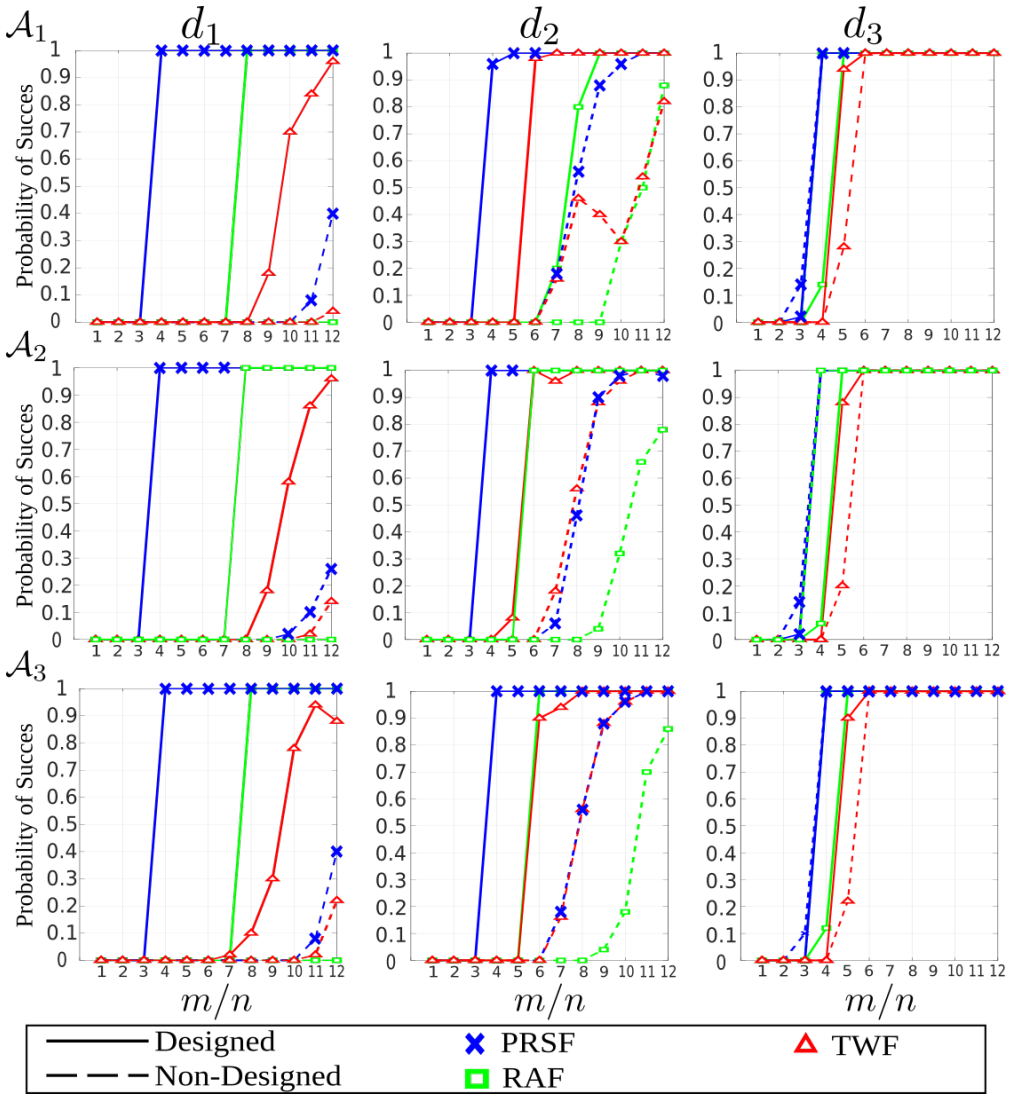

Figura 12. Empirical success rate of different reconstruction methods using designed coded apertures vs the number of projections. Rows: Diffraction zones, columns: admissible random variables.

9.1.4. Noise Robustness. This section characterizes the robustness of designed coded apertures to recover an image from CDP when the measurements are corrupted by additive Gaussian noise for different values of SNR and for the three diffraction zones. Figure 13 presents the attained relative error using the PRSF method in Pinilla et al. (2018a) for $L = 4$, when the SNR is varied from 5 to 50 dB. Figure 13 suggests the effectiveness of the designed coded apertures to better estimate the image from noisy CDP with a gain of up to 0.4 of relative error compared with non-designed ensembles.

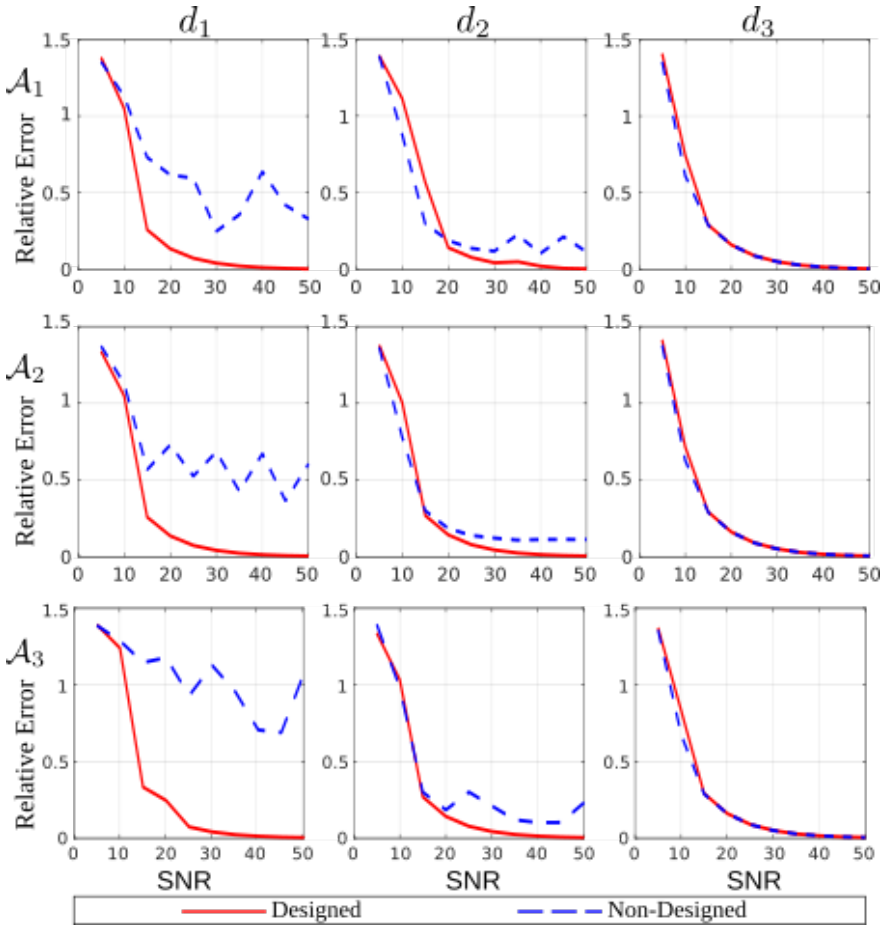

Figura 13. Recovery performance of designed coded apertures from noisy coded diffraction patterns when SNR is varied from 5 to 50dB. Rows: Diffraction zones. Columns: admissible random variables.

9.1.5. Support Estimation. This section presents the performance of the admissible random variables in Table 2 to estimate the non-zero coefficients of θ for each diffraction zone. The numerical results are summarized in Fig. 14. The color bar represents the support estimation percentage for the different admissible random variables over 100 trials, where 1 represents the best support that can be obtained. For each admissible random variable the sparsity is varied from $s = 0.1n$ to $s = 0.5n$ as shown in Fig. 14. Note that the admissible random variables d_1 and d_2 in Table 2 attain the highest performance compared with d_3. In fact, this observation validates the

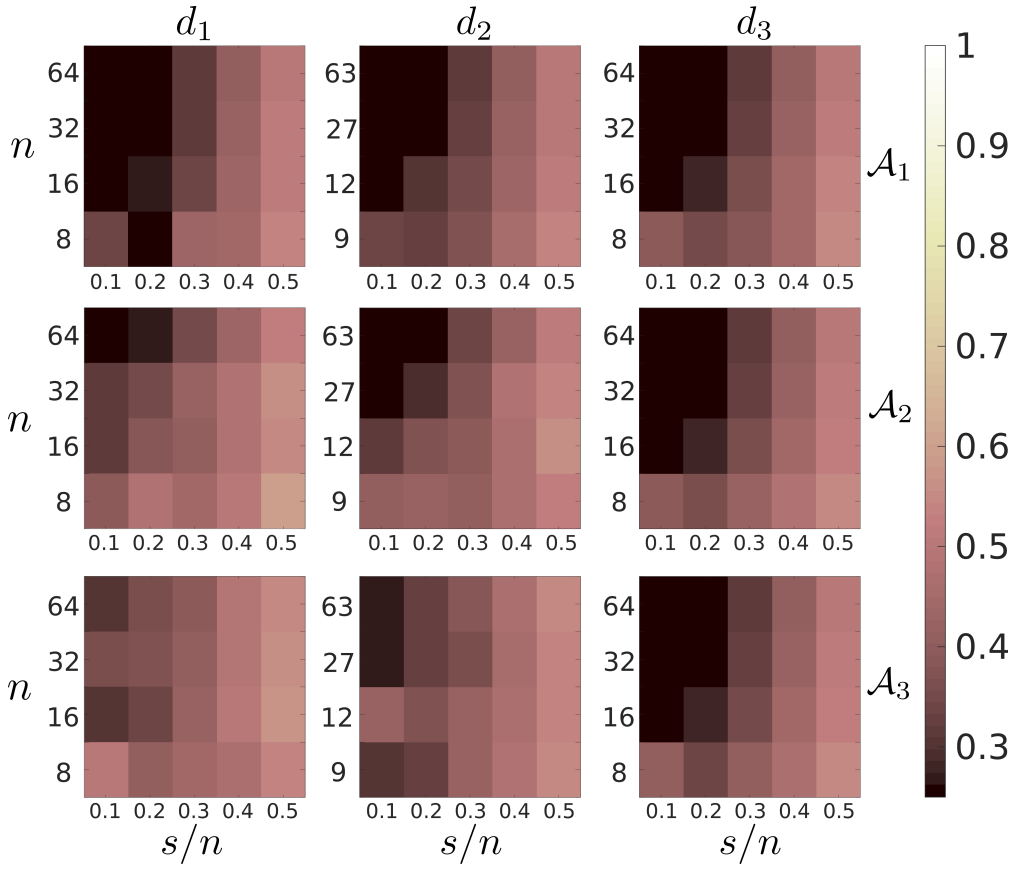

Figura 14. Empirical success rate estimating the non-zero coefficients of θ varying the image size from $n = 8 \times 8$ to $n = 64 \times 64$ and level sparsity from $s = 0.1n$ to $s = 0.5n$. Rows: Diffraction zones, columns: admissible random variables.

theoretical result established in Lemma 8.1.1, because $\mathbb{E}[d_1] \neq 0$ and $\mathbb{E}[d_2] \neq 0$ while $\mathbb{E}[d_3] = 0$.

To complement the results in Fig. 14, Table 3 reports the performance of the admissible random variables d_1 and d_3 to recover an image from CDP at the middle zone using Algorithm 2, assuming sparsity constrains for $L = 1$, $s = 0.1n$ and $n = 64 \times 64$. The average error over 100 tests was calculated. Note that Algorithm 2 requires fewer iterations to converge when a designed coded aperture is employed. In addition, it can be seen that d_1 $(\mathbb{E}[d_1] \neq 0)$ achieves better reconstruction performance than d_3 $(\mathbb{E}[d_3] = 0)$, fact that validates the theoretical result in Lemma 8.1.1.

Table 3
Recovery Performance of two Admissible Random Variables from Sparsity Constraints

	Relative Error		**# Iterations**	
	Designed	No-designed	Designed	No-designed
d_1	$2.33e^{-16}$	$2.26e^{-16}$	**100**	173
d_3	**0.53**	0.79	**14**	96

9.2. Analysis of the Proposed Phase Retrieval Algorithm

9.2.1. Sampling Complexity and Speed of Convergence. These experiments are performed for the noiseless real/complex Gaussian model as shown in Figs. 15 and 16, using the Truncated Spectral initialization proposed in Chen and Candes (2015) for all the algorithms under analysis, *i.e*. TAF, RWF, TWF, and PRSF in order to analyze the sampling complexity and the speed of convergence of those methods.

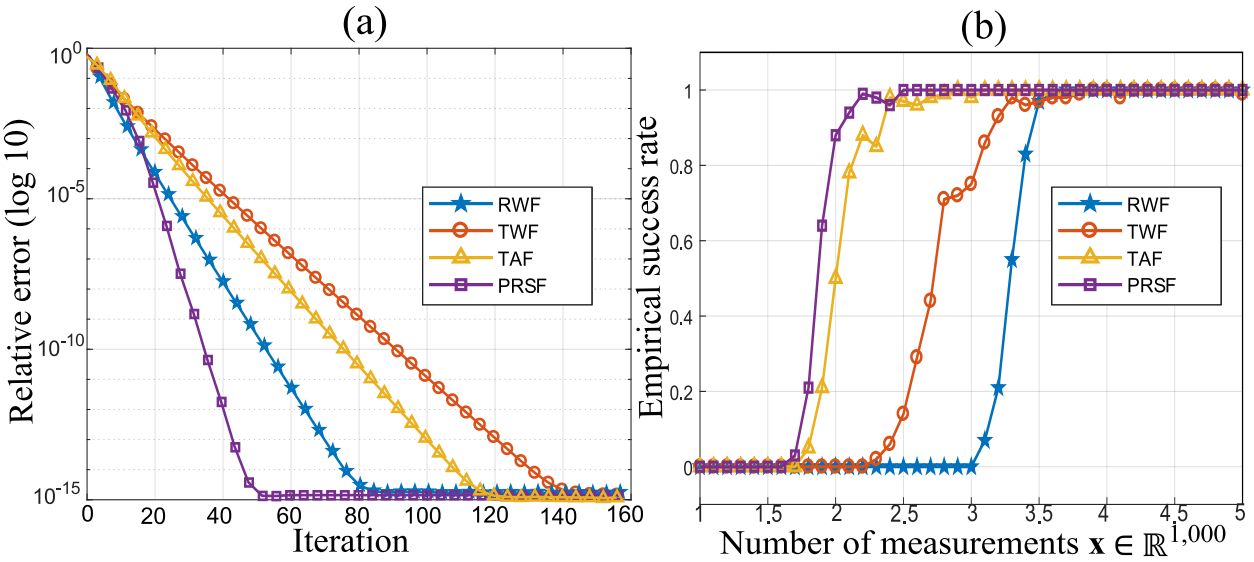

Figura 15. Relative error versus iteration for $m/n = 8$. (b) Empirical success rate versus number of measurements with m/n varying 0.1 from 0 to 5 under their own initialization.

These results suggest that PRSF exhibits a higher performance compared with TAF, TWF and RWF for both real and complex cases, in terms of sampling complexity and speed of convergence.

9.3. Numerical Results for FROG

This section evaluates the numerical performance of BSGA and compares the results with the stochastic gradient algorithm Ptych proposed in Sidorenko et al. (2016). We used the following parameters for Algorithm 4: $\gamma_1 = 0.1$, $\gamma = 0.1$, $\alpha = 0.6$, $\mu_0 = 65$, and $\varepsilon = 1 \times 10^{-10}$. The number of indices that are chosen uniformly at random is fixed as $Q = N$. A cubic interpolation was used in Algorithm 6 (see Line 4), and the regularization parameter was fixed to $\lambda = 0.5$.

Five tests were conducted to evaluate the performance of the proposed method under noisy

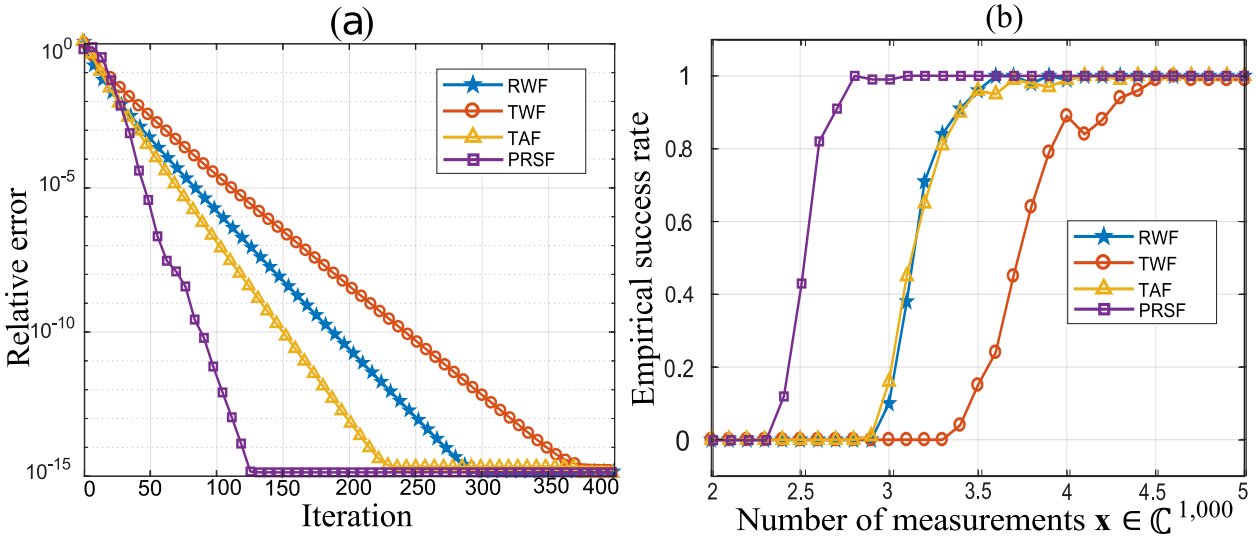

Figura 16. (a) Relative error versus iteration for $m/n = 8$. (b) Empirical success rate versus number of measurements with m/n varying 0.1 from 0 to 5 under their own initialization.

and noiseless scenarios at different values of signal-to-noise-ratio (SNR), defined as

$$SNR = 10\log_{10}(\|\mathbf{Z}\|_{\mathrm{F}}^2/\|\boldsymbol{\sigma}\|_2^2)$$

, where σ is the variance of the noise. First, we examine the empirical success rate of BSGA for different values of L. The second experiment assesses the performance of the initialization technique and its impact on the reconstruction quality. Third, we show several examples of reconstructed pulses attained with BSGA and Ptych under noisy and noiseless scenarios, when the complete FROG trace is used. The fourth experiment investigates the performance of the proposed method and Ptych in reconstructing the pulses when $L > 1$ and the FROG trace is corrupted by noise. The last test compares the computational complexity between the reconstruction methods in terms of their running time to reach a given relative error.

The signals used in the simulations were constructed as follows. For all tests, we built a set of $\left\lceil \frac{N-1}{2} \right\rceil$-bandlimited pulses that conform to a Gaussian power spectrum centered at 800 nm. Specifically, each pulse ($N = 128$ grid points) is produced via the Fourier transform of a complex vector with a Gaussian-shaped amplitude with a cutoff frequency of 150 femtoseconds^{-1} (fsec^{-1}). Next, we multiply the obtained power spectrum by a uniformly distributed random phase. In the experiments we used the inverse Fourier of this signal as the underlying pulse.

9.3.1. Empirical Probability of Success. This section numerically evaluates the success rate of BSGA. To this end, BSGA and Ptych are initialized at $\mathbf{x}^{(0)} = \mathbf{x} + \delta\zeta$, where δ is a fixed constant and ζ takes values on $\{-1, 1\}$ with equal probability, while L ranges from 1 to 6. A trial is declared successful when the returned estimate attains a relative error as in (86) that is smaller than 10^{-6}. We numerically determine the empirical success rate among 100 trials. Fig. 30 summarizes these results, and shows that BSGA performs better than Ptych, since it is able to retrieve the signal for larger values of L.

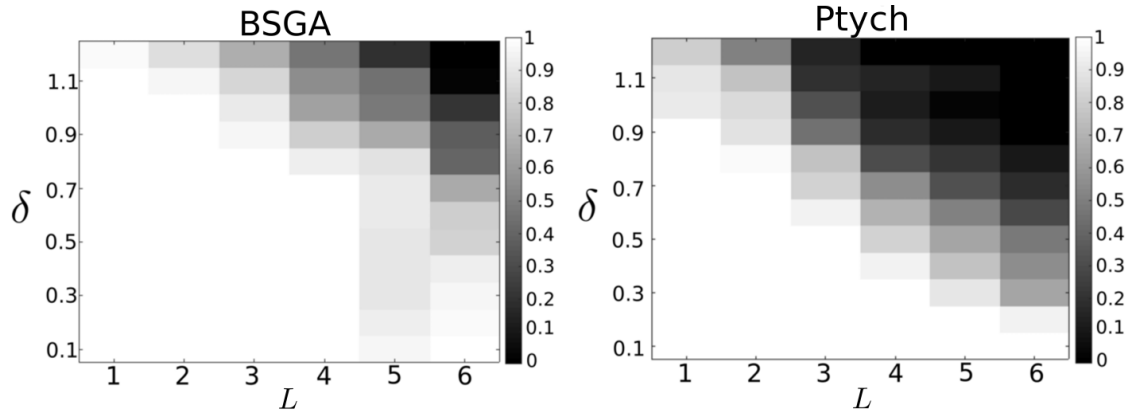

Figura 17. Empirical success rate comparison between BSGA and Ptych as a function of L and δ in the absence of noise.

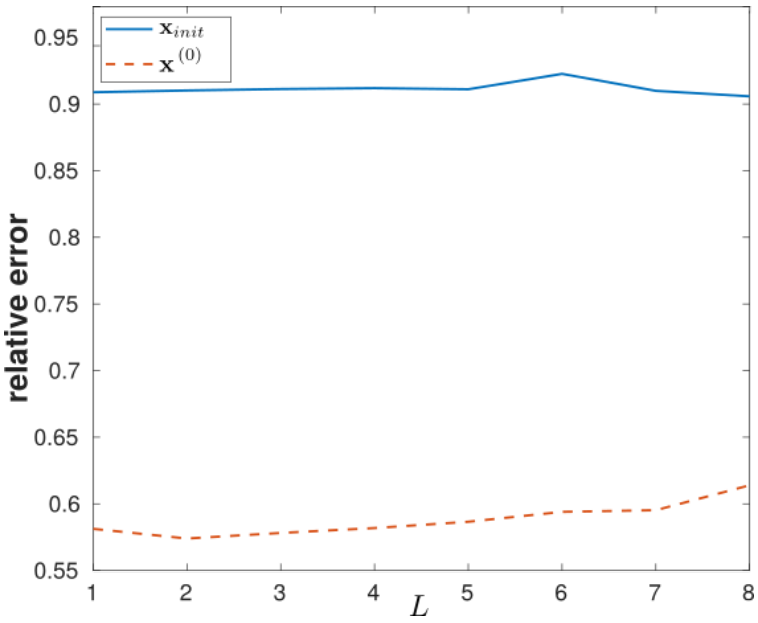

Figura 18. Relative error comparison between the initial vector $\mathbf{x}_{ini\_pty}$ as defined in (102), and the returned initial guess $\mathbf{x}^{(0)}$ for different values of L in the absence of noise. For each value of L, an average of the relative error was computed among 100 trials.

9.3.2. Relative Error of the Initialization Procedure. This section examines the impact of the designed initialization described in Algorithms 9 and 6, under noisy and noiseless scenarios. We compare the relative error between the starting vector in (102), and the returned solution $\mathbf{x}^{(0)}$

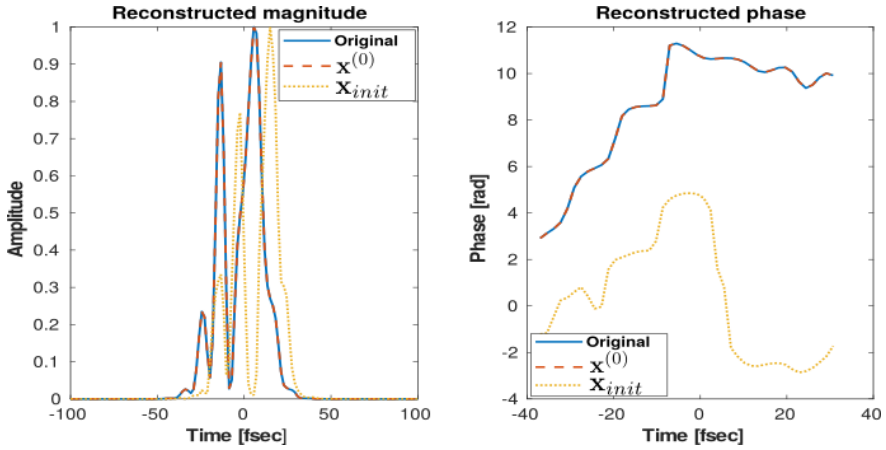

Figura 19. Reconstructed pulses from the FROG trace with $L = 4$ using Algorithm 7 initialized by $\mathbf{x}_{ini\_pty}$ and the returned vector $\mathbf{x}^{(0)}$ using Algorithm 6.

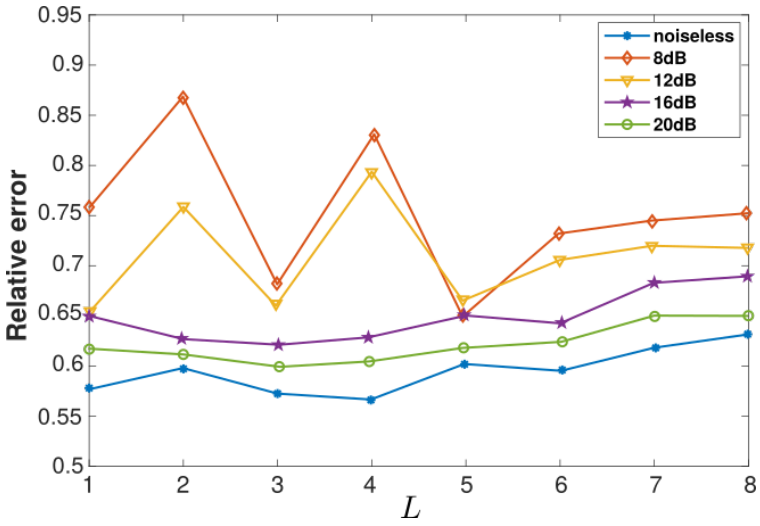

Figura 20. Performance of the proposed initialization described in Algorithms 9 and 6 at different SNR levels, with L ranging from 1 to 8. For each value of L, the relative error was averaged over 100 trials.

of the proposed initialization procedure. The number of iterations to attain the vector $\mathbf{x}^{(0)}$ using the designed initialization was fixed as $T = 2$, and we numerically determine the relative error averaged over 100 trials. These numerical results are summarized in Fig. 18, and indicate that the proposed initialization algorithm outperforms $\mathbf{x}_{ini\_pty}$.

In order to illustrate the effect of the initial guesses, we ran Algorithm 7 initialized by $\mathbf{x}_{ini\_pty}$ and $\mathbf{x}^{(0)}$ with $L = 4$. Fig. 19 shows the attained reconstructions. Notice that the proposed reconstruction algorithm fails in estimating the input pulse when it was initialized by $\mathbf{x}_{ini\_pty}$.

We numerically determine the performance of the proposed initialization at different SNR levels, with L ranging from 1 to 8. Specifically, we added white noise to the FROG measurements at different SNR levels: SNR = 8dB, 12dB, 16dB and 20dB. Fig. 20 displays the relative error attained by the proposed initialization for different SNR and L values.

From Fig. 20 it can be seen that the returned initialization at levels of SNR $\leq$ 16dB is,

approximately, independent of the value of L when $L \leq 6$. Combining these numerical results with Fig. 30, we conclude that BSGA is able to better estimate the underlying pulse (up to trivial ambiguities) if $L \leq 4$ for both noiseless and noisy scenarios compared to Ptych.

Finally, we numerically determine the empirical success rate of BSGA with increasing L, in the absence of noise, when Algorithm 7 is initialized with $\mathbf{x}_{ini\_pty}$, a random vector and $\mathbf{x}^{(0)}$. A trial is declared successful when the returned estimate attains a relative error as in (86) that is smaller than 1×10^{-6}. The results are summarized in Fig. 21, where the number of iterations that BSGA requires to reach the given relative error for $L = 1$ is also presented. The success rate and the number of iterations are averaged over 100 pulses. The reported results show the effectiveness of Algorithm 7 when it is initialized by $\mathbf{x}^{(0)}$ for $L > 1$.

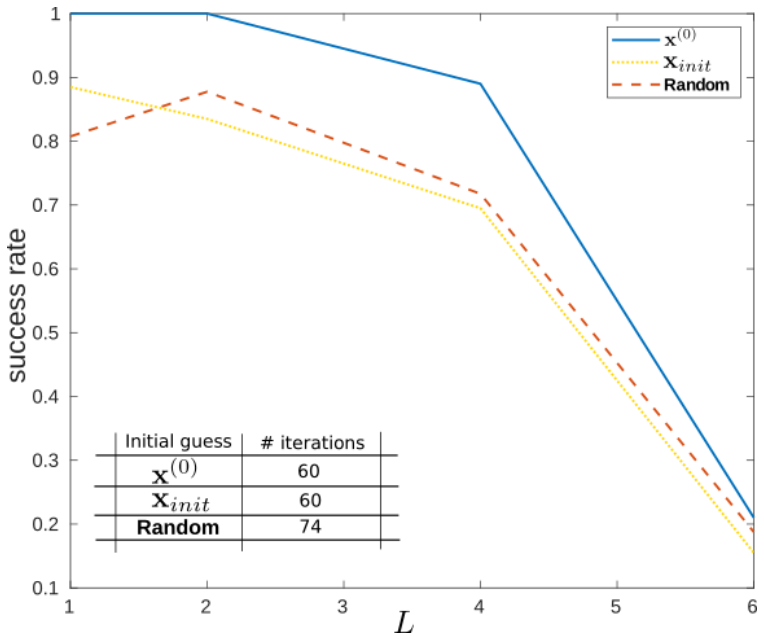

Figura 21. Empirical success rate of Algorithm 7 when it is initialized by $\mathbf{x}^{(0)}$, $\mathbf{x}_{ini\_pty}$ and a random vector as a function of L in the absence of noise.

9.3.3. Pulse Reconstruction Examples for $L = 1$. In this section we show the performance of BSGA in recovering two pulses under noiseless and noisy scenarios for $L = 1$. The results are presented in Figs. 22, and 23, respectively, where the attained relative errors by BSGA and Ptych are included. For the second scenario, the FROG trace is corrupted by Gaussian noise with SNR = 20dB.

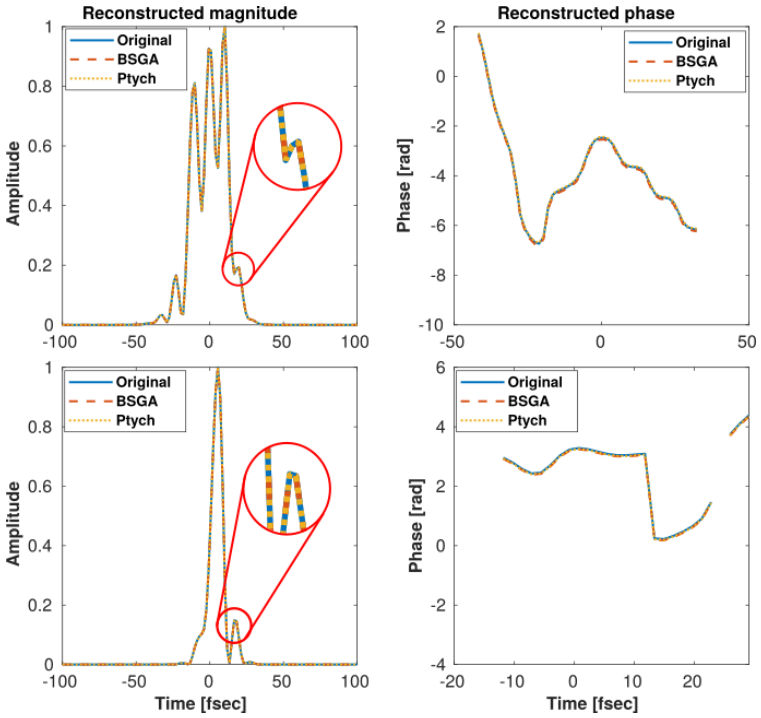

Figura 22. Reconstructed pulses from complete FROG data ($L = 1$), in the absence of noise. The attained error for both BSGA and Ptych was 1×10^{-6}.

From the results in Fig. 22 it can be observed that both methods, BSGA and Ptych, are able to estimate the pulses and provide similar results for the noiseless case.

On the other hand, in Fig. 23, the attained reconstructions, for the noisy scenario, indicate that BSGA is able to better estimate the pulse compared to Ptych. This advantage is obtained

because of the effectiveness of the proposed smoothing update step and initialization strategy from complete data as reported in Fig. 30, and Figs. 18, 10, respectively.

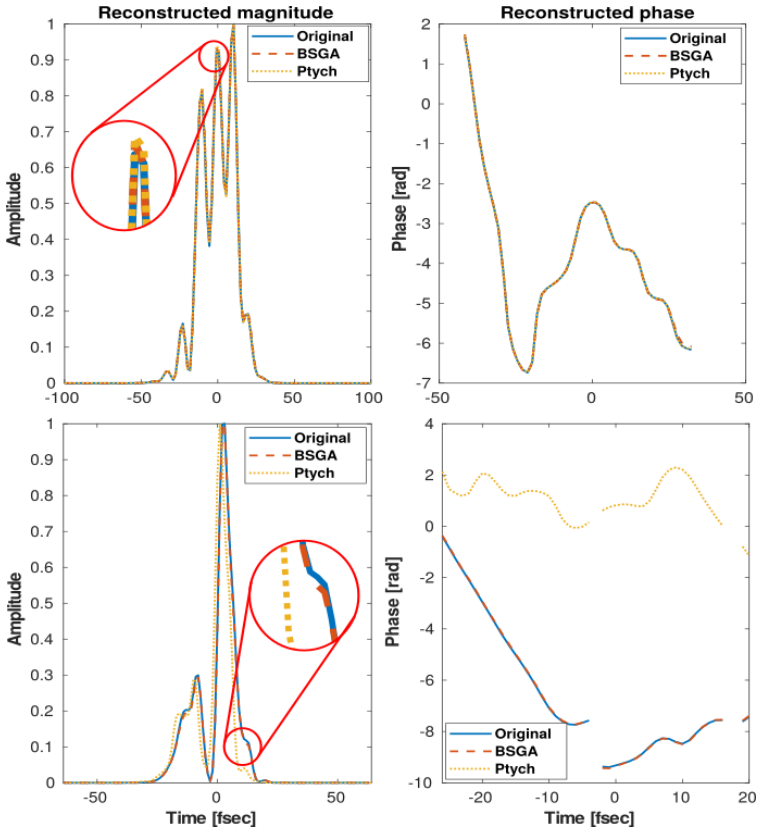

Figura 23. Reconstructed pulses from complete noisy FROG data ($L = 1$), with SNR = 20dB. The attained relative error for the top pulse for both BSGA and Ptych was 5×10^{-2}. For the bottom pulse the attained errors were 5×10^{-2} and 2×10^{-1} for BSGA and Ptych respectively.

9.3.4. Pulse Reconstruction Examples for $L > 1$. Next, we examine the recovery performance of BSGA from noisy incomplete data by adding Gaussian noise with SNR = 20dB, for $L \in \{2,4,8\}$. Figs. 24 and 25 illustrate the attained reconstructions for BSGA and Ptych; their attained relative errors are also reported in Fig. 25. These figures suggest that BSGA better estimates the pulse and its FROG trace compared to Ptych over a range of values of L. This advantage is obtained because of the effectiveness of the proposed smoothing update step and initialization

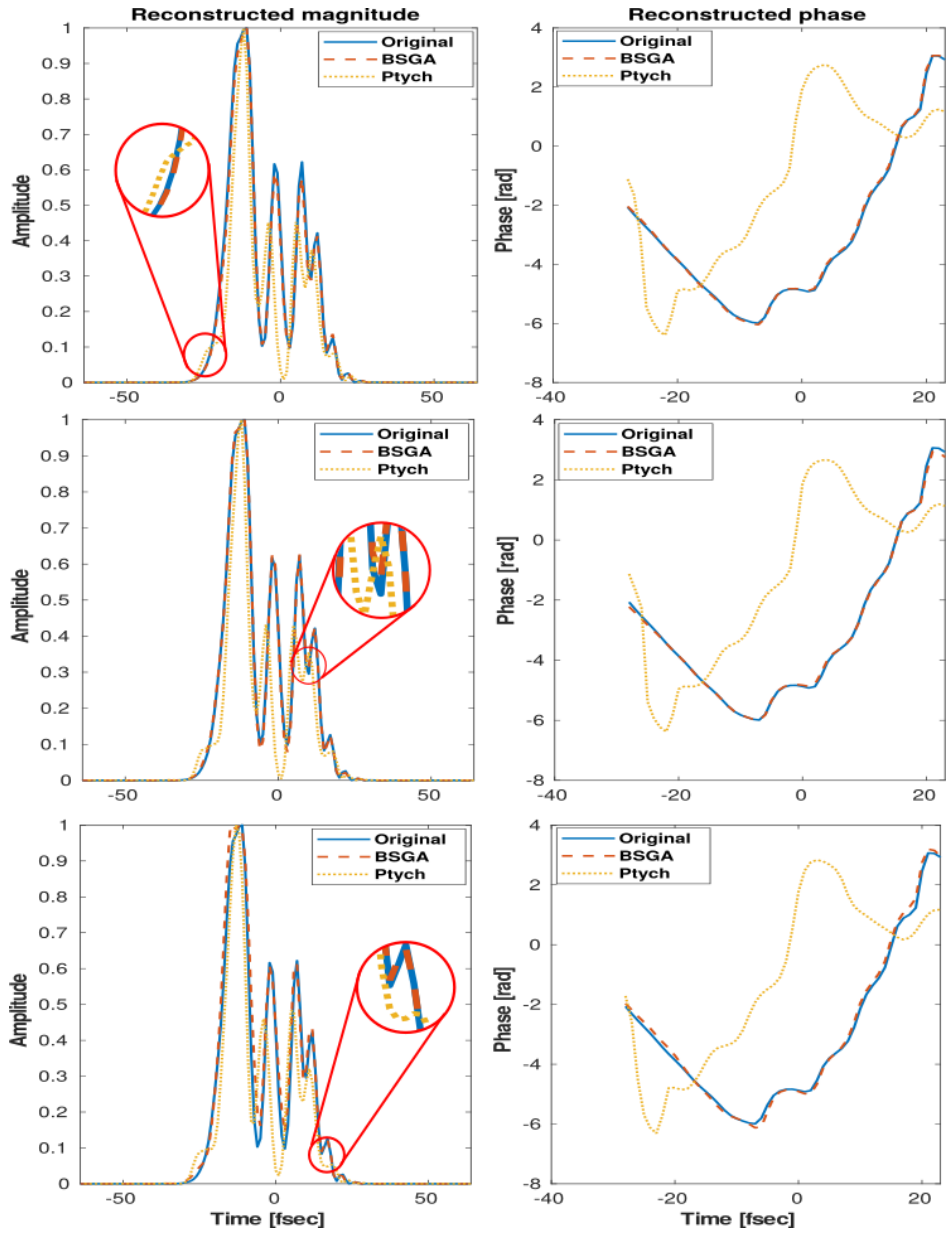

Figura 24. Reconstructed pulses from incomplete noisy FROG traces (SNR = 20dB), for different values of L. (a) $L = 2$, (b) $L = 4$, and (c) $L = 8$.

strategy from incomplete data as reported in Fig. 30, and Figs. 18, 10, respectively.

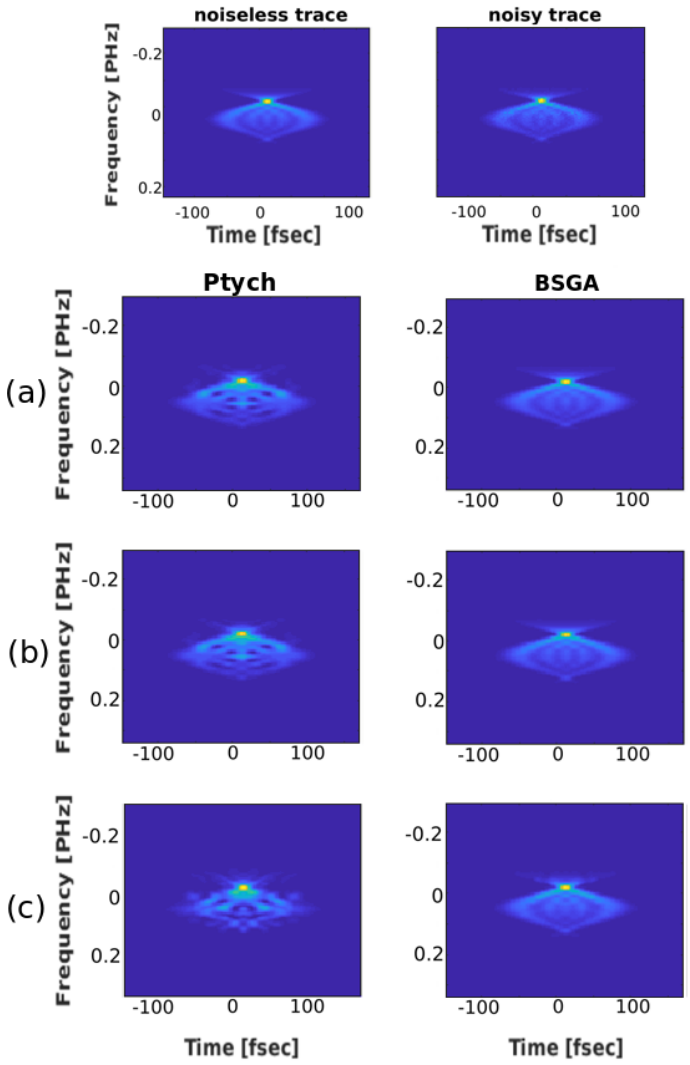

Figura 25. Reconstruction of full FROG traces from incomplete noisy data for all methods. Top row shows the desirable full FROG trace without and with noise of SNR = 20dB. (a) $L = 2$, (b) $L = 4$, and (c) $L = 8$. The attained errors for BSGA and Ptych were 5×10^{-2} and 2×10^{-1}, respectively, for all the reconstructed FROG traces.

9.3.5. Computational Complexity. Simulations were conducted to compare the speed of convergence of the algorithms in the absence of noise, for $L = 1, 2$ and 4. Table 4 reports the number of iterations and running time required by BSGA and Ptych to achieve a relative error of 1×10^{-6}, averaged over 100 pulses. The experiment shows that BSGA is similar in time and number of iterations compared to Ptych for a range of values of L.

Table 4
Comparison of iteration count and time cost

	Algorithms	**Iterations**	**Time (s)**
$L=1$	BSGA	60	1.451
	Ptych	36	1.325
$L=2$	BSGA	111	1.567
	Ptych	125	1.954
$L=4$	BSGA	265	1.772
	Ptych	300	2.013

9.4. Numerical Results for Radar

This section evaluates the numerical performance of the proposed method. We used the following parameters for Algorithm 7: $\gamma_1 = 0.1$, $\gamma = 0.1$, $\alpha = 0.6$, $\mu_0 = 65$, and $\varepsilon = 1 \times 10^{-10}$. The number of indices that are chosen uniformly at random is fixed as $Q = N$.

The signals used in the simulations were constructed as follows. For all tests, we built a set of $\left\lceil \frac{N-1}{2} \right\rceil$-band-limited and time-limited signals that conform to a Gaussian power spectrum centered at 800 nm. Specifically, each signal ($N = 128$ grid points) is produced via the Fourier transform of a complex vector with a Gaussian-shaped amplitude with a cutoff frequency of 150 microseconds^{-1} (usec^{-1}). Next, we multiply the obtained power spectrum by a uniformly distributed random phase. In the experiments we used the inverse Fourier of this signal as the underlying signal.[2]

---

[2] All simulations were implemented in Matlab R2019a on an Intel Core i7 3.41Ghz CPU with 32 GB RAM.

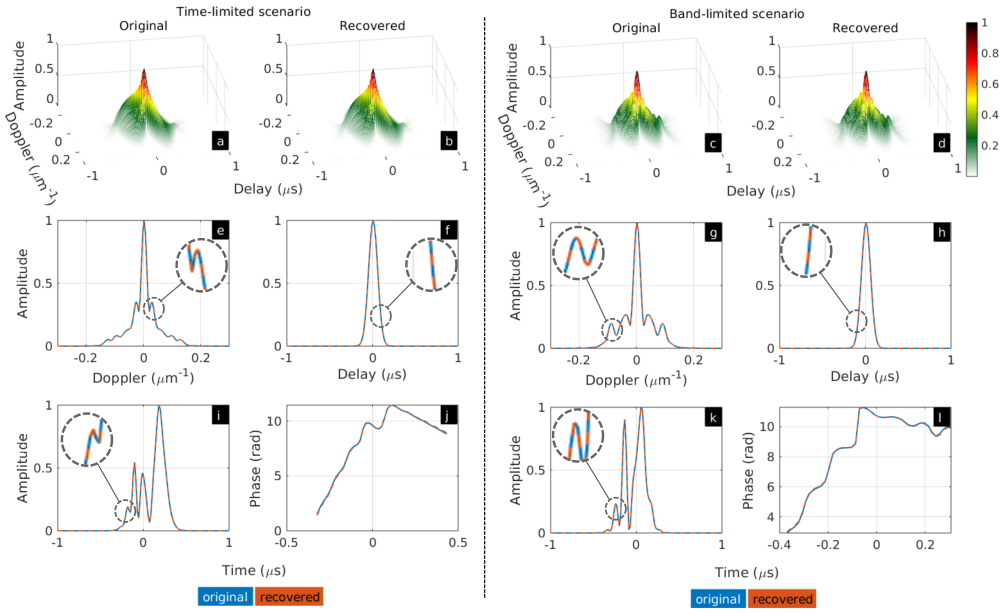

Figura 26. Reconstructed time and band-limited signals with their ambiguity functions in the absent of noise. The attained relative error as in (86) was 1×10^{-6} for both signals. (a), (c) and (b), (d) are the original and recovered ambiguity functions, respectively. (e), (g), and (f), (h) are 1D slices of the ambiguity functions for the time and Doppler dimensions, respectively. (i), (k) and (j), (l) correspond to the recovered magnitude and phase of the estimated signals, respectively.

The tests are divided in three sections to study the performance of Algorithm 7 for complete and incomplete AF, and additional type of time, band limited signals under noisy and noiseless scenarios at different values of signal-to-noise-ratio (SNR), defined as $\text{SNR} = 10\log_{10}(\|\mathbf{A}\|_{\text{F}}^2/\|\sigma\|_2^2)$, where σ is the variance of the noise. The radar function is incomplete when only few shifts or Fourier frequencies are considered. In the first section we examine the ability of Algorithm 7 to recover the signal from complete data. The second section assesses the performance of Algorithm 7 to recover the underlying signal when the AF is incomplete. The last test studies the ability of the proposed method to estimate different type of radar signals than the described above.

9.4.1. Signal Reconstruction from Complete Data. The performance of Algorithm 7 is presented to recover time and band-limited signals under noiseless and noisy scenarios using the

complete radar ambiguity function. The results are presented in Figs. 27, 28, where the attained relative errors by the proposed algorithm are included. For the second scenario, the radar ambiguity function trace is corrupted by Gaussian noise with SNR = 20dB. Specifically, in the noisy case we are assuming that ambiguity function is not perfectly designed which allows to evaluate the robustness of Algorithm 7. The results in Figs. 27 and 28 suggest that the proposed method is able to estimate the signals.

9.4.2. Signal Reconstruction from Incomplete Data. The success rate of Algorithm 7 is evaluated when the ambiguity function is incomplete. To this end, Algorithm 7 is initialized at $\mathbf{x}^{(0)} = \mathbf{x} + \delta\zeta$, where δ is a fixed constant and ζ takes values on $\{-1, 1\}$ with equal probability, while a percentage of the delays are set to zero. A trial is declared successful when the returned

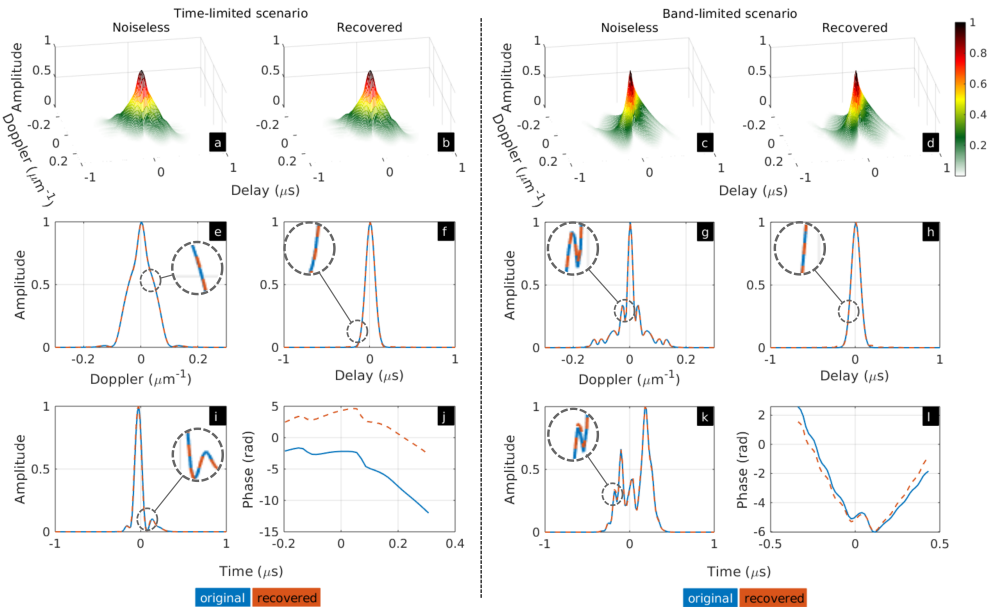

Figura 27. Reconstructed time and band-limited signals with their ambiguity functions in the present of noise with SNR = 20dB. The attained relative error as in (86) was 5×10^{-2} for both signals. (a), (c) and (b), (d) are the noiseless (ideal) and recovered ambiguity functions, respectively. (e), (g), and (f), (h) are 1D slices of the ambiguity functions for the time and Doppler dimensions, respectively. (i), (k) and (j), (l) correspond to the recovered magnitude and phase of the estimated signals, respectively.

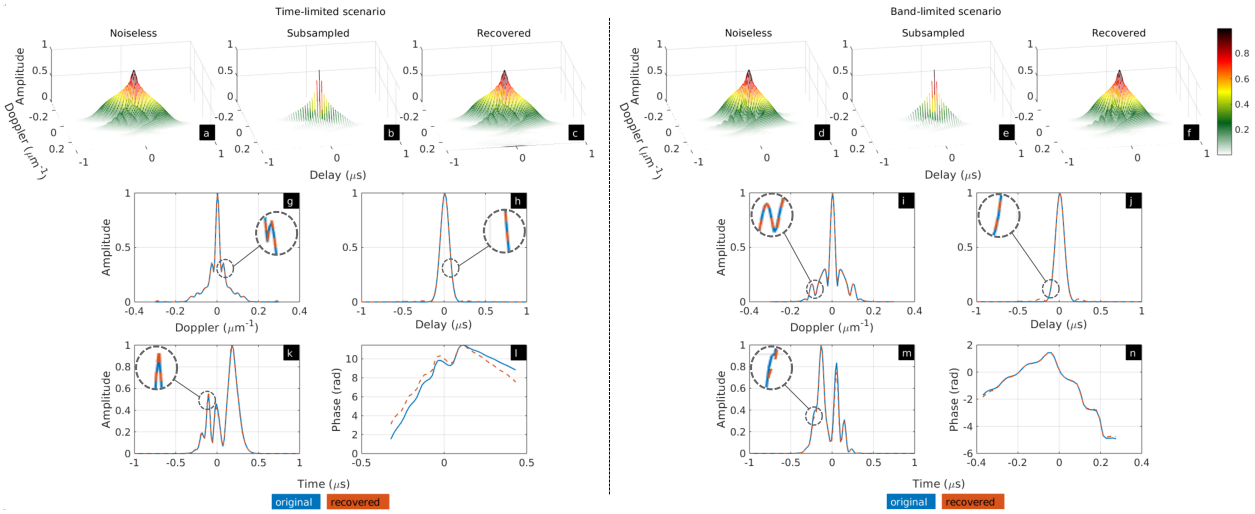

Figura 28. Reconstructed time and band-limited signals when a 50% of the delays of their ambiguity functions are uniformly removed. The incomplete AFs' were corrupted by noise with SNR = 20dB. The attained relative error as in (86) was 5×10^{-2} for both signals. (a),(d); (b),(e); and (c),(f) are the original, sub-sampled and recovered ambiguity functions, respectively. (g), (h), and (i), (j) are 1D slices of the ambiguity functions for the time and Doppler dimensions, respectively. (k), (m) and (j), (l) correspond to the recovered magnitude and phase of the estimated signals, respectively.

estimate attains a relative error as in (86) that is smaller than 10^{-6}. We numerically determine the empirical success rate among 100 trials. Fig. 30 summarizes these results, and shows that Algorithm 7 is able to estimate the pulse when the AF is incomplete.

In Fig. 30 the % of removed delays are performed uniformly, which means for instance in the case of 50% every two delays, starting from the first one, are preserved. Additionally, in the case of 75% means every three delays, starting from the first one, are preserved. From these results it is numerically validate Proposition 4 (in consequence Corollary 1) that not all the delays are need to estimate the underlying signal. To illustrate this, Figs. 27 and 29 show the estimated time and band-limited signals from noisy incomplete AF (50% and 75% of the delays are removed respectively). Observe that Algorithm 7 is able to return a close estimation of the signal even when the incomplete AF is assumed imperfectly designed, suggesting the effectiveness of Algorithm 7.

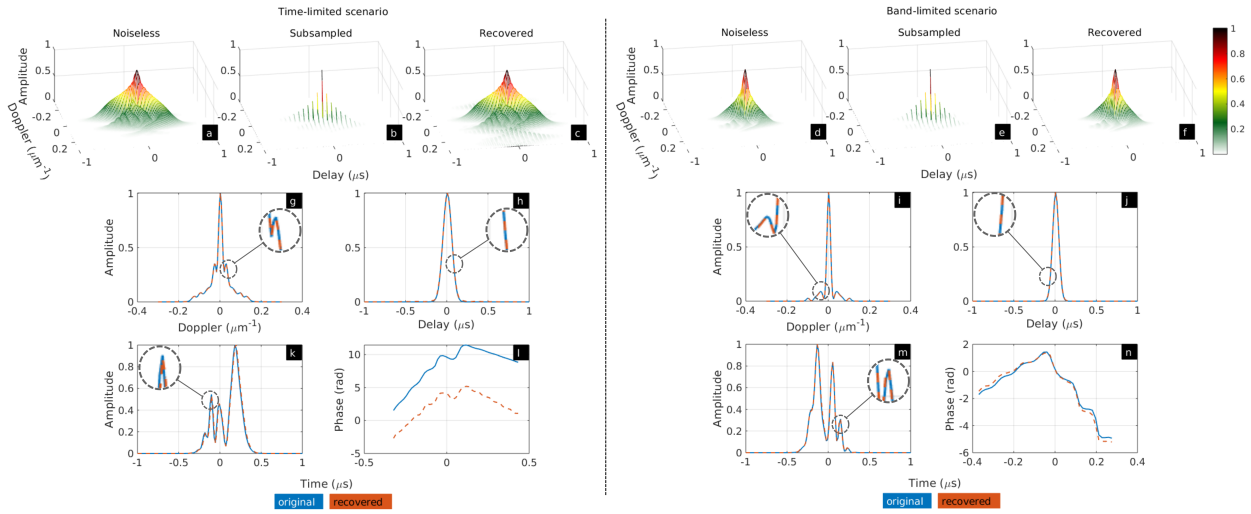

Figura 29. Reconstructed time and band-limited signals when a 75% of the delays of their ambiguity functions are uniformly removed. The incomplete AFs' were corrupted by noise with SNR = 20dB. The attained relative error as in (86) was 5×10^{-2} for both signals. (a),(d); (b),(e); and (c),(f) are the original, sub-sampled and recovered ambiguity functions, respectively. (g), (h), and (i), (j) are 1D slices of the ambiguity functions for the time and Doppler dimensions, respectively. (k), (m) and (j), (l) correspond to the recovered magnitude and phase of the estimated signals, respectively.

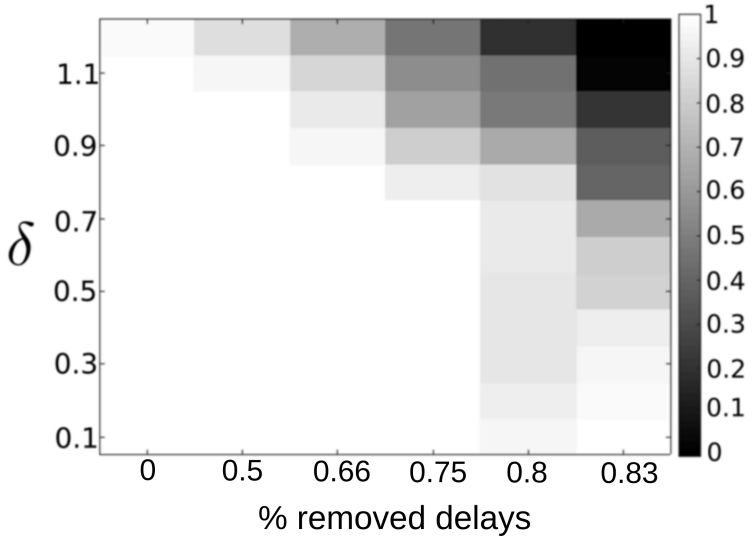

Figura 30. Empirical success rate of Algorithm 7 as a function of% removed delays (uniformly) and δ in the absence of noise.

To complement the results in Figs. 30, 28 and 29, here is also presented the performance of Algorithm 7 when a percentage of the delays and the Fourier frequencies of the AF are non-

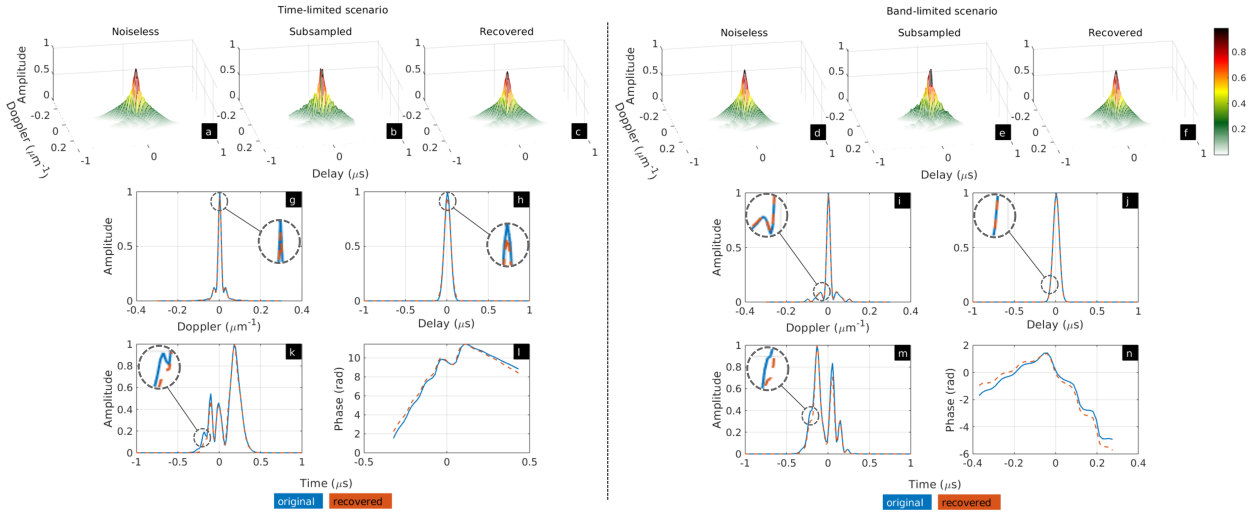

Figura 31. Reconstructed time and band-limited signals when a 57% of the delays of their ambiguity functions are non-uniformly removed. The incomplete AFs' were corrupted by noise with SNR = 20dB. The attained relative error as in (86) was 9×10^{-2} for both signals. (a),(d); (b),(e); and (c),(f) are the original, sub-sampled and recovered ambiguity functions, respectively. (g), (h), and (i), (j) are 1D slices of the ambiguity functions for the time and Doppler dimensions, respectively. (k), (m) and (j), (l) correspond to the recovered magnitude and phase of the estimated signals, respectively.

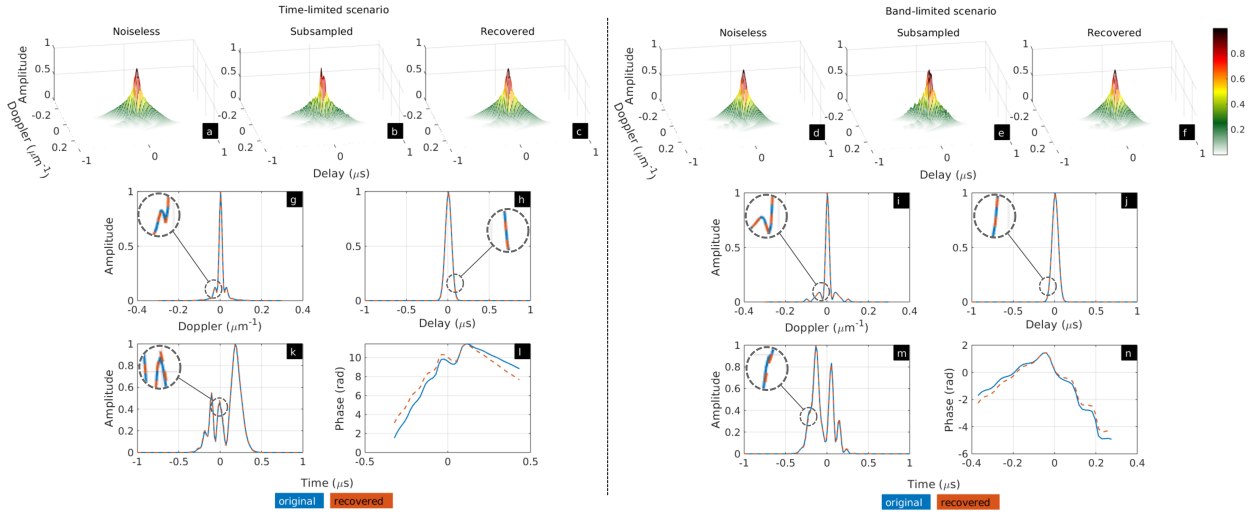

Figura 32. Reconstructed time and band-limited signals when a 57% of the Fourier frequencies of their ambiguity functions are non-uniformly removed. The incomplete AFs' were corrupted by noise with SNR = 20dB. The attained relative error as in (86) was 6×10^{-2} for both signals. (a),(d); (b),(e); and (c),(f) are the original, sub-sampled and recovered ambiguity functions, respectively. (g), (h), and (i), (j) are 1D slices of the ambiguity functions for the time and Doppler dimensions, respectively. (k), (m) and (j), (l) correspond to the recovered magnitude and phase of the estimated signals, respectively.

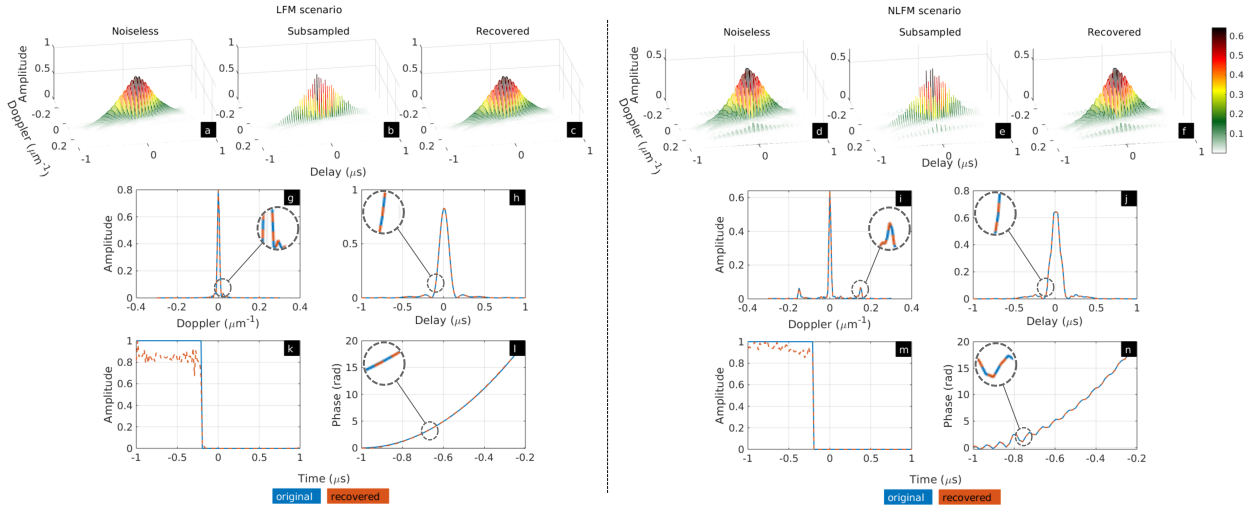

Figura 33. Reconstructed time and band-limited signals when a 50% of the delays of their ambiguity functions are uniformly removed. The incomplete AFs' were corrupted by noise with SNR = 20dB. The attained relative error as in (86) was 6×10^{-2} for both signals. (a),(d); (b),(e); and (c),(f) are the original, sub-sampled and recovered ambiguity functions, respectively. (g), (h), and (i), (j) are 1D slices of the ambiguity functions for the time and Doppler dimensions, respectively. (k), (m) and (j), (l) correspond to the recovered magnitude and phase of the estimated signals, respectively.

uniformly removed, illustrated in Figs. 31 and 32. Specifically, 28% of the first and last delays/ frequencies of the AF were set to zero in Fig. 31/32, respectively. These results suggest that a non-uniform selection of the delays to be removed reduces the ability of Algorithm 7 to estimate the analyzed pulse compared with a uniform strategy. In contrast, in the case of a non-uniform modality to remove frequencies it can be concluded that the performance of Algorithm 7 is close to the uniform selection of the delays to be removed.

9.4.3. Additional Type of Signals. In this section we investigate the performance of Algorithm 7 to estimate Linear/Non-linear Frequency Modulated (LFM/NLFM) pulses from its incomplete noisy ambiguity function. These kind of signals are modeled as

$$\mathbf{x}[n] = \mathbf{a}[n]e^{j\pi\varphi[n]}, \tag{128}$$

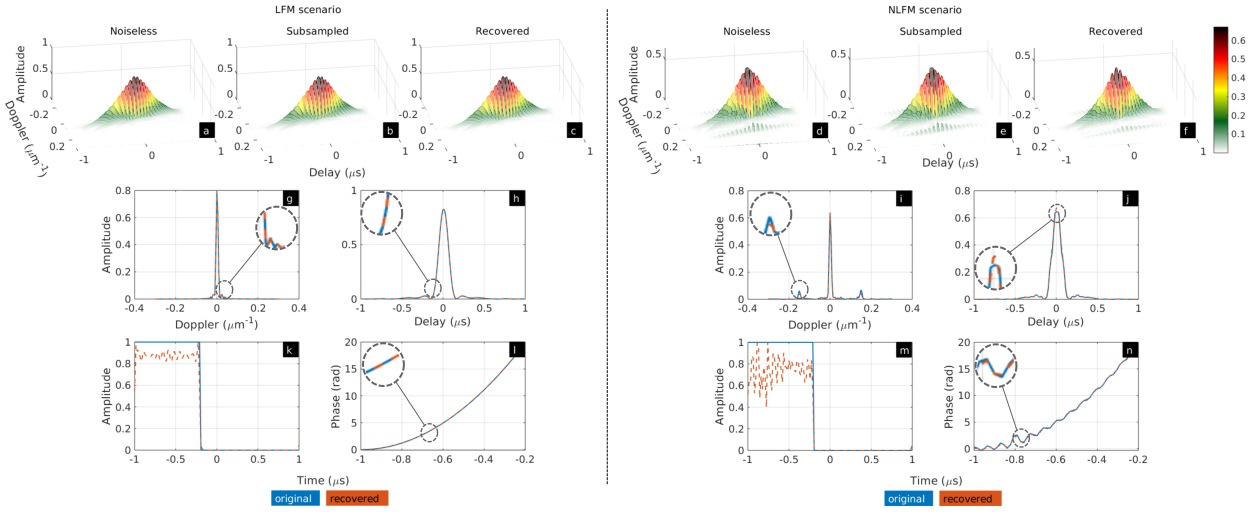

Figura 34. Reconstructed time and band-limited signals when a 19% of the first and last Fourier frequencies of their ambiguity functions are removed. The incomplete AFs' were corrupted by noise with SNR = 20dB. The attained relative error as in (86) was 9×10^{-2} for both signals. (a),(d); (b),(e); and (c),(f) are the original, sub-sampled and recovered ambiguity functions, respectively. (g), (h), and (i), (j) are 1D slices of the ambiguity functions for the time and Doppler dimensions, respectively. (k), (m) and (j), (l) correspond to the recovered magnitude and phase of the estimated signals, respectively.

where $\varphi[n]$ is given by

$$\begin{aligned} \varphi[n] &= \pi k(\Delta t n)^2, && \text{(LFM)} \\ \varphi[n] &= \pi k t^2 + \sum_{l=1}^{L} \alpha_l \cos(2\pi l \Delta t n / T) && \text{(NLFM)}, \end{aligned} \tag{129}$$

with T as the duration of the pulse, Δt as the sampling size in time, $k = \frac{\Delta f}{T}$ such that Δf is the swept bandwidth, and $L > 0$ is an integer. The values for α_l are given by $\alpha_l = \frac{0.4T}{l}$. For this experiment $\Delta f = 128 \times 10^3$, and $\Delta t = 0.4 \times 10^{-6}$. The values of $\mathbf{a}[n]$ for both kind of pulses model a rectangular envelope which is given by

$$\mathbf{a}[n] = \begin{cases} 1 & 0 \leq \Delta t n \leq T \\ 0 & \text{otherwise} \end{cases}. \tag{130}$$

In this experiment two noisy scenarios are considered: first, the 50% of the delays are uniformly removed from the the AF, second, 19% of the first and last Fourier frequencies of the AF are removed. The results are summarized in Fig. 33, and 34, where $SNR = 20$dB, and the attained relative error is also presented. These results suggest that Algorithm 7 is able to estimate accurately the phase of the pulses, while the reconstructed magnitudes present some artifacts. This limitation comes from the fact that their AF is significantly wide such that the removed information is enough to limit the reconstruction quality.

9.5. Numerical Results for Target Detection

In this section, the performance of Algorithm 10 to correctly detect a target is analyzed when its line 2 is replaced by the output of Algorithm 9 and the alternatives OPI, WMCI and TSI. This analysis is carried out for noiseless and noisy measurements, under three different scenarios, as illustrated in Fig. 35. The first scenario employs four objects (toys) with the same magnitude information, two of them containing non-constant phase information. The second scenario uses four objects (two pears and two apples), three of them containing non-constant phase information. The third scenario combines three different objects (toy, apple and pear), with non-constant phase information. In particular, the purpose of this experiment is to detect the toy, and the apple with non-constant phase information, respectively, which are highlighted as illustrated in Fig. 35.

From the results shown in Fig. 35, it can be concluded that the proposed TD methodology is able to detect a target using Algorithm 9 with a single snapshot, while using WMCI requires at least four snapshots to correctly detect the target. Here, only WMCI results are shown since this procedure returns a closer approximation of the optical field compared to OPI and TSI. These

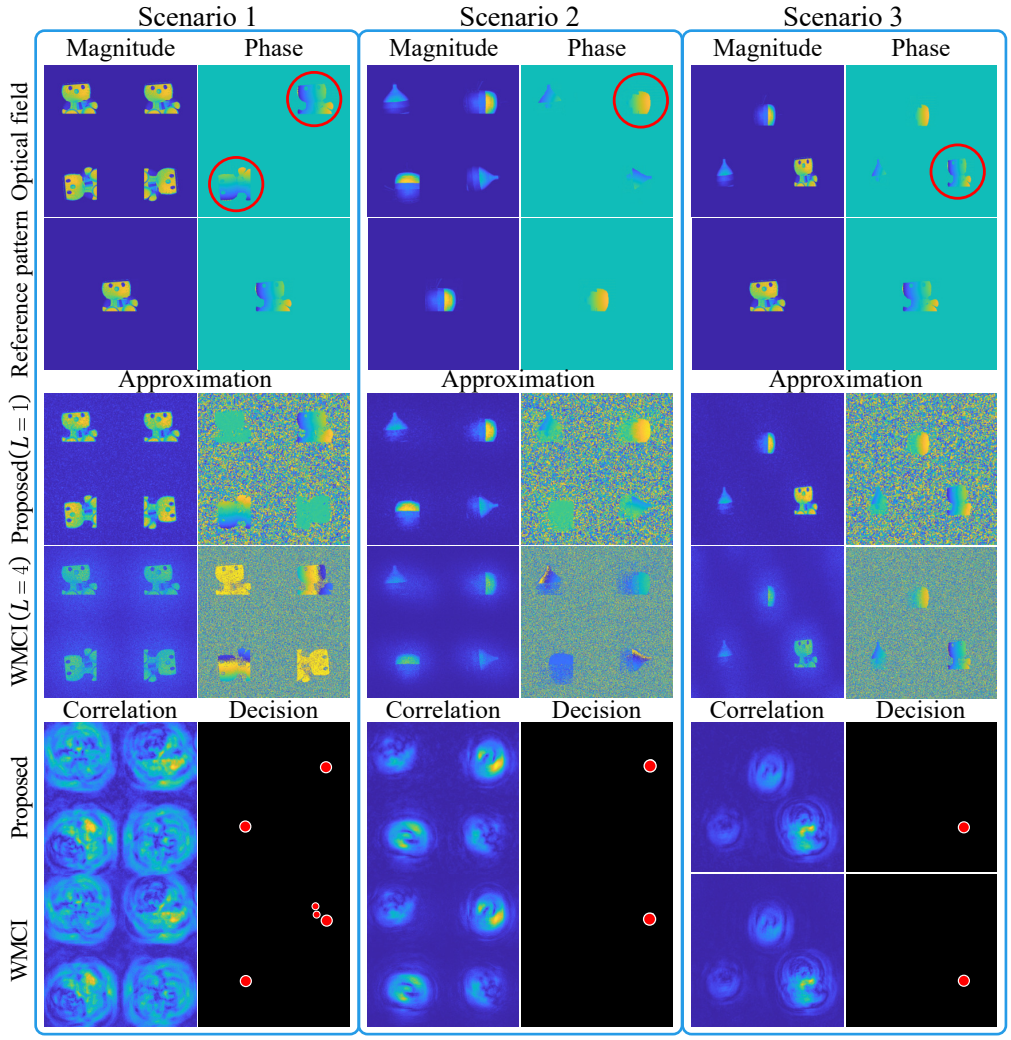

Figura 35. Performance of the proposed TD methodology using Algorithm 9 and WMCI approach for $L \in \{1,4\}$ and SNR $= 30$[dB], under three different scenarios.

results show the effectiveness of Algorithm 9 with noisy data. To complement the results in Fig. 35, Fig. 36 presents the detection rate, from noiseless measurements, of Algorithm 10 when its line 2 is replaced by the output of Algorithm 9 and the alternatives OPI, WMCI and TSI, varying the number of snapshots. The proposed TD methodology using the filtered spectral method achieves better performance using less than four snapshots compared to traditional approximation procedures.

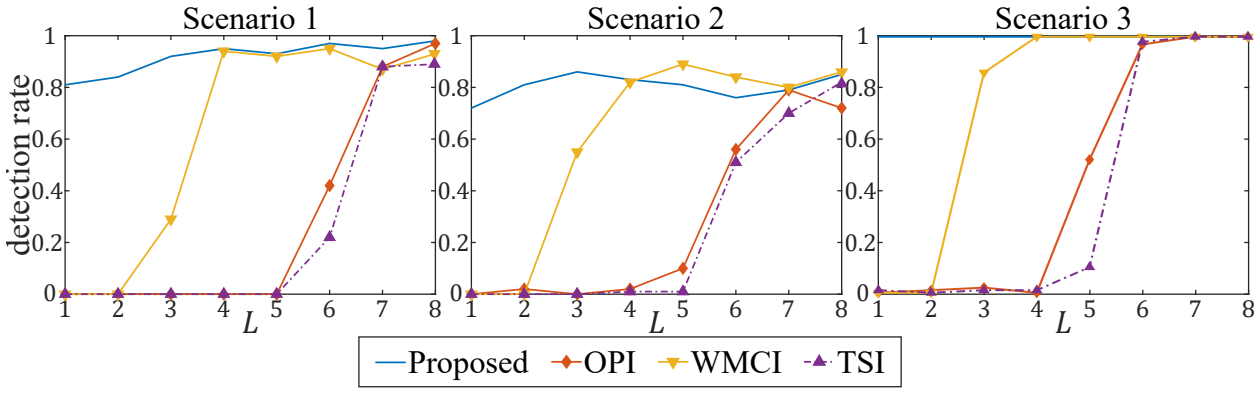

Figura 36. Detection rate of the proposed TD methodology through different approximation procedures varying the number of snapshots and using noiseless measurements.

Finally, Fig. 37 displays the detection rate in grayscale color of Algorithm 10 for the different analyzed approximation methods, where a lighter color indicates superior detection rate. These experiments assume that the measurements are corrupted by Gaussian noise, at different noise levels and number of snapshots. Observe that these numerical results suggest that a single

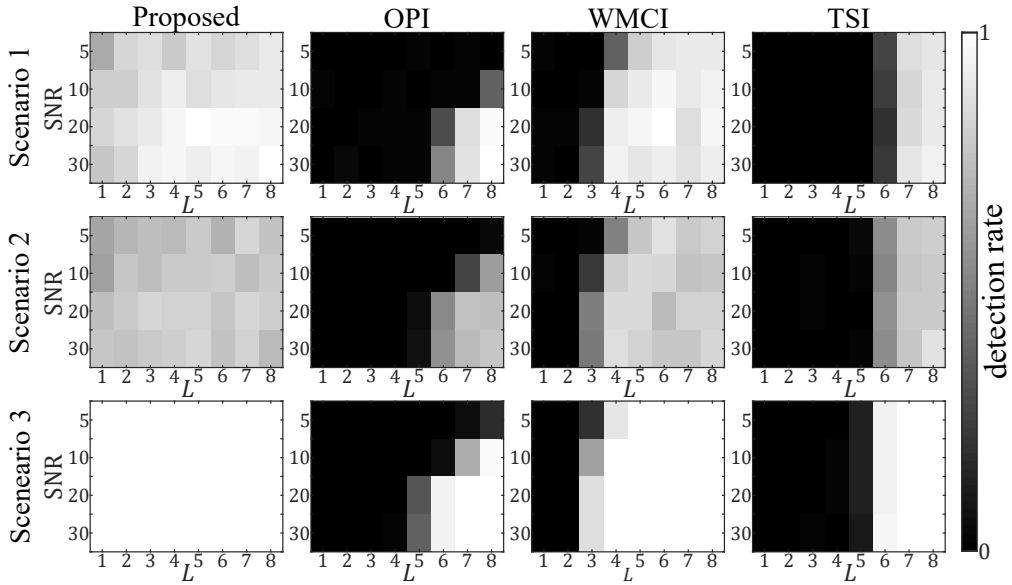

Figura 37. Detection rate of the proposed TD methodology using different approximation procedures with several noise levels and number of snapshots.

snapshot is enough to detect a target with a detection rate of up to 84 % using Algorithm 9 in the tested datasets, even when the noise level increases. Further, the proposed TD methodology using traditional approximation procedures with a single snapshot cannot detect the target of interest, and the detection rate is significantly low.

10. Conclusions and future directions

This thesis presented theoretical recovery guarantees of a scene from coded diffraction patterns acquired at the near, middle and far zones. In addition, a strategy to design the set of coded apertures under admissible coding variables was introduced. In particular, three different coding random variables were used to modulate complex scenes. Numerical experiments were conducted to evaluate the performance of the proposed design methodology in terms of the initialization, successful recovery and reconstruction quality of the phase from CDP. Specifically, simulations show that the designed coded apertures attain a reduction in terms of the relative error of up to 50 % in the initialization compared with non-designed ensembles. Further, by using designed coded apertures the scene is successfully recovered employing up to 40 % less measurements compared with non-designed ensembles. The effectiveness of the proposed method to recover the phase from CDP under additive Gaussian noise using the designed coded apertures was numerically verified. Finally, under sparsity assumptions it was validated that an admissible random variable d satisfying $\mathbb{E}[d] \neq 0$ attains a better performance estimating the non-zero coefficients of θ.

Additionally, the smoothing gradient method, result of this thesis, was extended to the phase retrieval problem in Frequency Resolver Optical Gating (FROG). The results show improvements in recovering the pulse for both magnitude and phase, from noisy incomplete data. Additionally,

the numerical results suggest the effectiveness of the proposed initialization under both noiseless and noisy scenarios with incomplete data. Future work should include implementing BSGA on real data to further validate its performance. Another interesting research direction is to examine similar strategies for blind FROG in which two signals are estimated simultaneously Trebino (2012).

This thesis also analytically demonstrates that time and band-limited signals can be estimated (up to trivial ambiguities) from its ambiguity function. We explore a trust region gradient method to estimate these kind of signals under complete/incomplete noisy and noiseless scenarios, and we verify that these signals can be estimated in a polynomial time with enough accuracy when the data is complete. In the case of incomplete data, we found that although Proposition 4, and Corollary 1 suggest that the full AF is not required to guarantee uniqueness much more work can be done here in order to better estimate the pulses from incomplete data. In fact, numerical results suggest that pulses producing wide AF are not desire in order to reduce the required data to be analyzed. Additionally, these result also validated Proposition 4, and Corollary 1 for three kind of signals.

There are several limitations of our current reconstruction algorithm. First, the initialization strategy employed is simple and can be improve in several ways. The optimization problem that our initialization pursuits to solve is highly non-convex, and since we use an alternating approach the present of saddle points and local minimum should be avoided. Second, we currently fix the parameters by simply cross-validation, however they can be learned from the kind of signals.

Bibliography

Arguello, H. and Arce, G. R. (2014). Colored coded aperture design by concentration of measure in compressive spectral imaging. *IEEE Transactions on Image Processing*, 23(4):1896–1908.

Bagirov, A., Karmitsa, N., and Mäkelä, M. M. (2014). *Introduction to Nonsmooth Optimization: theory, practice and software*. Springer.

Bandeira, A. S., Chen, Y., and Mixon, D. G. (2014). Phase retrieval from power spectra of masked signals. *Information and Inference: a Journal of the IMA*, 3(2):83–102.

Beinert, R. and Plonka, G. (2018). Enforcing uniqueness in one-dimensional phase retrieval by additional signal information in time domain. *Applied and Computational Harmonic Analysis*, 45(3):505–525.

Bendory, T., Beinert, R., and Eldar, Y. C. (2017a). Fourier phase retrieval: Uniqueness and algorithms. In *Compressed Sensing and its Applications*, pages 55–91. Springer.

Bendory, T., Edidin, D., and Eldar, Y. C. (2018a). On signal reconstruction from FROG measurements. *Appl. and Compu. Harmon. Anal.*

Bendory, T., Edidin, D., and Eldar, Y. C. (2019a). Blind phaseless short-time fourier transform recovery. *IEEE Transactions on Information Theory*.

Bendory, T., Edidin, D., and Eldar, Y. C. (2019b). Blind phaseless short-time fourier transform recovery. *IEEE Transactions on Information Theory*.

Bendory, T., Eldar, Y. C., and Boumal, N. (2018b). Non-convex phase retrieval from STFT measurements. *IEEE Trans. on Inf. Theory*, 64(1):467–484.

Bendory, T., Sidorenko, P., and Eldar, Y. C. (2017b). On the uniqueness of frog methods. *IEEE Signal Processing Letters*, 24(5):722–726.

Candes, E. J., Eldar, Y. C., Strohmer, T., and Voroninski, V. (2015a). Phase retrieval via matrix completion. *SIAM review*, 57(2):225–251.

Candès, E. J. and Li, X. (2014). Solving quadratic equations via phaselift when there are about as many equations as unknowns. *Foundations of Computational Mathematics*, 14(5):1017–1026.

Candes, E. J., Li, X., and Soltanolkotabi, M. (2015b). Phase retrieval from coded diffraction patterns. *Applied and Computational Harmonic Analysis*, 39(2):277–299.

Candes, E. J., Li, X., and Soltanolkotabi, M. (2015c). Phase retrieval via wirtinger flow: Theory and algorithms. *IEEE Transactions on Information Theory*, 61(4):1985–2007.

Candès, E. J., Li, X., and Soltanolkotabi, M. (2015). Phase retrieval via wirtinger flow: Theory and algorithms. *IEEE Transactions on Information Theory*, 61(4):1985–2007.

Candès, E. J. and Wakin, M. B. (2008). An introduction to compressive sampling. *IEEE Signal Processing Magazine*, 25(2):21–30.

Chandu, K., Stanich, M., Wu, C. W., and Trager, B. (2013). Direct binary search (dbs) algorithm

with constraints. In *Color Imaging XVIII: Displaying, Processing, Hardcopy, and Applications*, volume 8652, page 86520K. International Society for Optics and Photonics.

Chang, H., Lou, Y., Duan, Y., and Marchesini, S. (2018). Total variation–based phase retrieval for poisson noise removal. *SIAM Journal on Imaging Sciences*, 11(1):24–55.

Chen, Y. and Candes, E. (2015). Solving random quadratic systems of equations is nearly as easy as solving linear systems. In *Advances in Neural Information Processing Systems*, pages 739–747.

Chen, Y. and Candès, E. (2015). Solving random quadratic systems of equations is nearly as easy as solving linear systems. In *Advances in Neural Information Processing Systems*, pages 739–747.

Clarke, F. H. (1990). *Optimization and nonsmooth analysis*. SIAM.

Correa, C. V., Arguello, H., and Arce, G. R. (2016). Spatiotemporal blue noise coded aperture design for multi-shot compressive spectral imaging. *JOSA A*, 33(12):2312–2322.

Dürig, U., Pohl, D. W., and Rohner, F. (1986). Near-field optical-scanning microscopy. *Journal of applied physics*, 59(10):3318–3327.

Eriksson, K., Estep, D., and Johnson, C. (2013). *Applied mathematics: Body and soul: Volume 1: Derivatives and geometry in IR3*. Springer Science & Business Media.

Fienup, C. and Dainty, J. (1987). Phase retrieval and image reconstruction for astronomy. *Image Recovery: Theory and Application*, 231.

Ghadimi, S. and Lan, G. (2013). Stochastic first-and zeroth-order methods for nonconvex stochastic programming. *SIAM J. on Opti.*, 23(4):2341–2368.

Gonzalez, R. C. and Wintz, P. (1977). Digital image processing(book). *Reading, Mass., Addison-Wesley Publishing Co., Inc.(Applied Mathematics and Computation*, (13):451.

Goodman, J. W. (2005). Introduction to fourier optics. *Introduction to Fourier optics, 3rd ed., by JW Goodman. Englewood, CO: Roberts & Co. Publishers, 2005*, 1.

Gross, D., Krahmer, F., and Kueng, R. (2017). Improved recovery guarantees for phase retrieval from coded diffraction patterns. *Applied and Computational Harmonic Analysis*, 42(1):37–64.

Gualdron, O. and Arsenault, H. H. (1993). Phase derived circular harmonic filter. *Opt. Comm.*, 104(1-3):32–34.

Guerrero, A., Pinilla, S., and Arguello, H. (2020). Phase recovery guarantees from designed coded diffraction patterns in optical imaging. *IEEE Trans. on Image Proc*, 29:5687–5697.

Hess, H., Betzig, E., Harris, T., Pfeiffer, L., and West, K. (1994). Near-field spectroscopy of the quantum constituents of a luminescent system. *Science*, 264(5166):1740–1745.

Horisaki, R., Matsui, H., and Tanida, J. (2017). Single-pixel compressive diffractive imaging with structured illumination. *Applied optics*, 56(14):4085–4089.

Hunger, R. (2007). *An introduction to complex differentials and complex differentiability*. Munich University of Technology, Inst. for Circuit Theory and Signal Processing.

Jaganathan, K., Eldar, Y. C., and Hassibi, B. (2015). Phase retrieval: An overview of recent developments. *arXiv preprint arXiv:1510.07713*.

Jahncke, C., Paesler, M., and Hallen, H. (1995). Raman imaging with near-field scanning optical microscopy. *Applied physics letters*, 67(17):2483–2485.

Jaming, P. (2010). The phase retrieval problem for the radar ambiguity function and vice versa. In *2010 IEEE Radar Conference*, pages 230–235. IEEE.

Jensen, J. R. and Lulla, K. (1987). Introductory digital image processing: a remote sensing perspective.

Kolte, R. and Özgür, A. (2016). Phase retrieval via incremental truncated wirtinger flow. *arXiv preprint arXiv:1606.03196*.

Kreyszig, E. (1989). *Introductory functional analysis with applications*, volume 1. wiley New York.

Loewen, E. G. and Popov, E. (2018). *Diffraction gratings and applications*. CRC Press.

Mejia, Y. and Arguello, H. (2018). Binary codification design for compressive imaging by uniform sensing. *IEEE Transactions on Image Processing*.

Millane, R. P. (1990). Phase retrieval in crystallography and optics. *JOSA A*, 7(3):394–411.

Mojica, E., Pertuz, S., and Arguello, H. (2017). High-resolution coded-aperture design for com-

pressive x-ray tomography using low resolution detectors. *Optics Communications*, 404:103–109.

Nocedal, J. and Wright, S. (2006). *Numerical optimization*. Springer Science & Business Media.

Parikh, N. and Boyd, S. (2014). Proximal algorithms. *Foundat. and TrendsÂ® in Optim.*, 1(3):127–239.

Pauwels, E. J. R., Beck, A., Eldar, Y. C., and Sabach, S. (2018). On Fienup methods for sparse phase retrieval. *IEEE Trans. on Signal Process.*, 66(4):982–991.

Pinilla, S., Bacca, J., and Arguello, H. (2018a). Phase retrieval algorithm via nonconvex minimization using a smoothing function. *IEEE Transactions on Signal Processing*, 66(17):4574–4584.

Pinilla, S., Bacca, J., and Arguello, H. (2018b). Sprsf: Sparse phase retrieval via smoothing function. *arXiv preprint arXiv:1807.09703*.

Pinilla, S., García, H., Díaz, L., Poveda, J., and Arguello, H. (2018c). Coded aperture design for solving the phase retrieval problem in x-ray crystallography. *Journal of Computational and Applied Mathematics*, 338:111–128.

Pinilla, S., Poveda, J., and Arguello, H. (2018d). Coded diffraction system in x-ray crystallography using a boolean phase coded aperture approximation. *Optics Communications*, 410:707–716.

Pohl, D. W. and Courjon, D. (2012). *Near field optics*, volume 242. Springer Science & Business Media.

Poon, T.-C. and Liu, J.-P. (2014). *Introduction to modern digital holography: with MATLAB*. Cambridge University Press.

Prémont, G. and Sheng, Y. (1993). Fast design of circular-harmonic filters using simulated annealing. *App. Opt.*, 32(17):3116–3121.

Rodenburg, J. M. (2008). Ptychography and related diffractive imaging methods. *Advances in imaging and electron physics*, 150:87–184.

Saad, Y. (2003). *Iterative methods for sparse linear systems*, volume 82. siam.

Sao, M., Nakamura, Y., Tajima, K., and Shimano, T. (2018). Lensless close-up imaging with fresnel zone aperture. *Japanese Journal of Applied Physics*, 57(9S1):09SB05.

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 processing magazine*, 32(3):87–109.

Shevkunov, I., Katkovnik, V., Petrov, N., and Egiazarian, K. (2018). Super-resolution microscopy for biological specimens: lensless phase retrieval in noisy conditions. *Biomedical optics express*, 9(11):5511–5523.

Shimano, T., Nakamura, Y., Tajima, K., Sao, M., and Hoshizawa, T. (2018). Lensless light-field imaging with fresnel zone aperture: quasi-coherent coding. *Applied optics*, 57(11):2841–2850.

Sidorenko, P., Lahav, O., Avnat, Z., and Cohen, O. (2016). Ptychographic reconstruction algo-

rithm for frequency-resolved optical gating: super-resolution and supreme robustness. *Optica*, 3(12):1320–1330.

Spall, J. C. (2005). *Introduction to stochastic search and optimization: estimation, simulation, and control*, volume 65. John Wiley & Sons.

Stephens, D. J. and Allan, V. J. (2003). Light microscopy techniques for live cell imaging. *science*, 300(5616):82–86.

Thibault, P., Dierolf, M., Bunk, O., Menzel, A., and Pfeiffer, F. (2009). Probe retrieval in ptychographic coherent diffractive imaging. *Ultramicroscopy*, 109(4):338–343.

Trebino, R. (2012). *Frequency-resolved optical gating: the measurement of ultrashort laser pulses*. Springer Science & Business Media.

Vershynin, R. (2010). Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*.

Wang, G., Giannakis, G. B., and Chen, J. (2017a). Scalable solvers of random quadratic equations via stochastic truncated amplitude flow. *IEEE Transactions on Signal Processing*, 65(8):1961–1974.

Wang, G., Giannakis, G. B., and Eldar, Y. C. (2016a). Solving systems of random quadratic equations via truncated amplitude flow. *arXiv preprint arXiv:1605.08285*.

Wang, G., Giannakis, G. B., and Eldar, Y. C. (2018a). Solving systems of random quadratic equations via truncated amplitude flow. *IEEE Transactions on Information Theory*, 64(2):773–794.

Wang, G., Giannakis, G. B., Saad, Y., and Chen, J. (2018b). Phase retrieval via reweighted amplitude flow. *IEEE Transactions on Signal Processing*, 66(11):2818–2833.

Wang, G., Zhang, L., Giannakis, G. B., Akçakaya, M., and Chen, J. (2016b). Sparse phase retrieval via truncated amplitude flow. *arXiv preprint arXiv:1611.07641*.

Wang, G., Zhang, L., Giannakis, G. B., Akçakaya, M., and Chen, J. (2017b). Sparse phase retrieval via truncated amplitude flow. *IEEE Trans. on Signal Proc.*, 66(2):479–491.

Wright, S. J. and Nocedal, J. (1999). Numerical optimization. *Springer Science*, 35(67-68):7.

Yuan, Z., Wang, Q., and Wang, H. (2017). Phase retrieval via sparse wirtinger flow. *arXiv preprint arXiv:1704.03286*.

Zhang, C. and Chen, X. (2009). Smoothing projected gradient method and its application to stochastic linear complementarity problems. *SIAM Journal on Optimization*, 20(2):627–649.

Zhang, H. and Liang, Y. (2016). Reshaped wirtinger flow for solving quadratic system of equations. In *Advances in Neural Information Processing Systems*, pages 2622–2630.

Zi-Liang, P. and Dalsgaard, E. (1995). Synthetic circular-harmonic phase-only filter for shift, rotation, and scaling-invariant correlation. *App. Opt.*, 34(32):7527–7531.

Appendices

Appendix A. Proof of Theorem 3.1.1

The proof of Theorem 3.1.1 is divided into two parts. First, the right inequality in (11) is proved, and then, as a second part, the left inequality is proved. As $\mathbf{W} \in \mathscr{T}_{\mathbf{x}}$ has rank at most two, we can choose normalized vectors $\mathbf{u}, \mathbf{v} \in \mathbb{C}^n$ such that $\mathbf{W} = \lambda_1 \mathbf{u}\mathbf{u}^H + \lambda_2 \mathbf{v}\mathbf{v}^H$.

Then, considering the definition of the linear maps $\mathscr{A}_k$ in (8) it can be obtained

$$\begin{aligned} \|\mathscr{A}_k(\mathbf{W})\|_1 &= \sum_{i=1}^{m} \left| \lambda_1 |\langle \mathbf{a}_{i,k}, \mathbf{u} \rangle|^2 + \lambda_2 |\langle \mathbf{a}_{i,k}, \mathbf{v} \rangle|^2 \right| \\ &\leq |\lambda_1| \|\mathbf{A}_k \mathbf{u}\|_2^2 + |\lambda_2| \|\mathbf{A}_k \mathbf{v}\|_2^2 \\ &\leq (|\lambda_1| + |\lambda_2|) \, \|\mathbf{A}_k\|_2^2 = \|\mathbf{W}\|_1 \|\mathbf{A}_k\|_\infty^2, \end{aligned} \tag{131}$$

in which the first and second inequalities are obtained using the triangular inequality, and matrices $\mathbf{A}_k$ as was defined in (4). Also, $|\lambda_1| + |\lambda_2| = \|\mathbf{W}\|_1$. Further, considering the definition of matrices $\mathbf{A}_k$ in (4) and the sampling vectors in (7) one can find that

$$\mathbf{A}_k^H \mathbf{A}_k = \sum_{\ell=1}^{L} \overline{\mathbf{D}}_\ell \mathbf{D}_\ell, \tag{132}$$

Remark that given the fact that any admissible random variable is bounded, then it is sub-Gaussian Vershynin (2010). Further, considering condition (132) it can be obtained that $\frac{1}{\sqrt{r}}\mathbf{A}_k$ is an isotropic sub-Gaussian matrix, since $\mathbb{E}[\mathbf{A}_k^H \mathbf{A}_k] = r\mathbf{I}$, with $L \geq c_0 n$ for some sufficiently large constant $c_0 > 0$.

Then, from Theorem 5.39 in Vershynin (2010) it can be obtained

$$\mathscr{P}\left(\left\|\frac{1}{\sqrt{r}}\mathbf{A}_k\right\|_\infty \geq \sqrt{m}+C\sqrt{n}+t\right) \leq 2e^{-ct^2}, \tag{133}$$

for constants $c, C > 0$ and any $t > 0$. Then, taking $L \geq C^2\varepsilon^{-2}n$ and $t = \sqrt{nL}\varepsilon$ for any $\varepsilon \in (0, 1/2)$, it can be found from (133) that

$$\mathscr{P}\left(\frac{1}{rnL}\|\mathbf{A}_k\|_\infty^2 \leq 1+\delta\right) \leq 1-2e^{-cnL\varepsilon^2}, \tag{134}$$

for $\delta = 2\varepsilon$. Thus, combining (131) and (134) yields

$$\frac{1}{rnL}\|\mathscr{A}_k(\mathbf{W})\|_1 \leq (1+\delta)\|\mathbf{W}\|_1, \tag{135}$$

for any $\delta \in (0,1)$.

On the other hand, from (131) it can also concluded that

$$\begin{aligned}\|\mathscr{A}_k(\mathbf{W})\|_1 &\geq \sum_{i=1}^{m} \lambda_1|\langle \mathbf{a}_{i,k}, \mathbf{u}\rangle|^2 + \lambda_2|\langle \mathbf{a}_{i,k}, \mathbf{v}\rangle|^2 \\ &= \lambda_1\|\mathbf{A}_k\mathbf{u}\|_2^2 + \lambda_2\|\mathbf{A}_k\mathbf{v}\|_2^2 \\ &= (\lambda_1+\lambda_2)\, r = r\|\mathbf{W}\|_1, \end{aligned} \tag{136}$$

in which the second equality comes from observation in (132), also using that $|\lambda_1| + |\lambda_2| = \lambda_1 +$

$\lambda_2 = \|\mathbf{W}\|_1$ because $\mathbf{W}$ is assumed positive semidefinite. Further, if $r \leq L$, then from (136)

$$\frac{1}{rnL}\|\mathscr{A}_k(\mathbf{W})\|_1 \geq \frac{1}{nL}(1-\delta)\|\mathbf{W}\|_1, \tag{137}$$

for any $\delta \in (0,1)$. Thus, combining (135) and (137) the result holds.

Appendix B. Proof of Theorem 8.1.2

Due to homogeneity in (26), it suffices to work with the case where $\|\mathbf{x}\| = 1$. Instrumental in proving Theorem 8.1.2 is the following result.

Lemma 10.0.1. Consider the noiseless data $|\mathbf{a}_{i,k}^H \mathbf{x}|$. For any unit vector $\mathbf{x} \in \mathbb{C}^n$, there exists a vector $\mathbf{u} \in \mathbb{C}^n$ with $\mathbf{u}^H \mathbf{x} = 0$ and $\|\mathbf{u}\| = 1$, such that

$$\frac{1}{2}\|\mathbf{x}\mathbf{x}^H - \mathbf{z}_0\mathbf{z}_0^H\|_F^2 \leq \frac{\|\mathbf{S}_k\mathbf{u}\|_2^2}{\|\mathbf{S}_k\mathbf{x}\|_2^2}, \tag{138}$$

where $\mathbf{S}_k = \left[\frac{\mathbf{a}_{i_1,k}}{\|\mathbf{a}_{i_1,k}\|_2}, \cdots, \frac{\mathbf{a}_{i_J,k}}{\|\mathbf{a}_{i_J,k}\|_2}\right]^H$ for $i_p \in \mathscr{I}_0^c$, where $J = card(\mathscr{I}_0^c)$.

Demostración. Notice that

$$\begin{aligned}\frac{1}{2}\|\mathbf{x}\mathbf{x}^H - \mathbf{z}_0\mathbf{z}_0^H\|_F^2 &= \frac{1}{2}\|\mathbf{x}\|_2^4 + \frac{1}{2}\|\mathbf{z}_0\|_2^4 - |\mathbf{x}^H\mathbf{z}_0|^2 \\ &= 1 - |\mathbf{x}^H\mathbf{z}_0|^2 = 1 - \cos^2(\theta),\end{aligned} \tag{139}$$

where $\theta \in [0, \pi/2]$ is the angle between the spaces spanned by $\mathbf{x}$ and $\mathbf{z}_0$. Then one can write

$$\mathbf{x} = \cos(\theta)\mathbf{z}_0 + \sin(\theta)\mathbf{z}_0^\perp, \tag{140}$$

where $\mathbf{z}_0^\perp \in \mathbb{C}^n$ is a unit vector orthogonal to $\mathbf{z}_0$ and the real part of its inner product with $\mathbf{x}$ is

non-negative. Then from (140) one can find

$$\mathbf{x}^{\perp} = -\sin(\theta)\mathbf{z}_0 + \cos(\theta)\mathbf{z}_0^{\perp}, \tag{141}$$

in which $\mathbf{x}^{\perp} \in \mathbb{C}^n$ is a unit vector orthogonal to $\mathbf{x}$. Thus, considering (140) and (141), then appealing to Lemma 1 in Wang et al. (2018a)

$$\frac{1}{2}\|\mathbf{x}\mathbf{x}^H - \mathbf{z}_0\mathbf{z}_0^H\|_F^2 \leq \frac{\|\mathbf{S}_k\mathbf{x}^{\perp}\|_2^2}{\|\mathbf{S}_k\mathbf{x}\|_2^2}. \tag{142}$$

Then, taking $\mathbf{u} = \mathbf{x}^{\perp}$ the result holds. □

Turning out to prove Theorem 8.1.2. The first step consists in upper-bounding the term on the right hand-side of (142). Specifically, its numerator term will be upper bounded, and the denominator term lower bounded, which are summarized in the following lemmas.

Lemma 10.0.2. In the setup of Lemma 10.0.1, if $card(\mathscr{I}_0^c) \geq C_I n$, then the next

$$\|\mathbf{S}_k\mathbf{u}\|_2^2 \leq (1+\delta-\zeta)c_k card(\mathscr{I}_0^c) \tag{143}$$

holds for $\delta, \zeta \in (0,1)$ with probability at least $1-2e^{-Cn}$ provided that L is sufficiently large.

Demostración. The proof this lemma proceeds by cases. Also, remark that it is assumed that the set of coded apertures satisfy $\sum_{\ell=1}^{L} \mathbf{D}_\ell^H \mathbf{D}_\ell = r\mathbf{I}$ for some $r > 0$, with $L \geq c_0 n$ for some sufficiently large constant $c_0 > 0$.

As was discussed in Section 4.1, if the assumption over the set of coded apertures is assumed, then

$$\mathbf{E}_k = \sum_{i=1}^{nL} \frac{\mathbf{a}_{i,k}\mathbf{a}_{i,k}^H}{\|\mathbf{a}_{i,k}\|_2^2} \approx c_k \mathbf{I},$$

for the three diffraction zones, for some constants $c_k > 0$. Then, from standard concentration inequality on the sum of random positive semi-definite matrices with sub-Gaussian rows Vershynin (2010), it can be obtained that

$$(1-\delta) \leq \sigma_{min}\left(\frac{1}{c_k}\mathbf{E}_k\right) \leq \sigma_{max}\left(\frac{1}{c_k}\mathbf{E}_k\right) \leq (1+\delta), \tag{144}$$

with probability at least $1-2e^{-Cn}$ as long as L is sufficiently large, for some constant $\delta \in (0,1)$ and $C > 0$, where $\sigma_{max}(\cdot)$ and $\sigma_{min}(\cdot)$ denote the largest and smallest singular value, respectively. Given the fact that $\mathbf{S}_k$ is a sub-matrix of $\mathbf{E}_k$ for each $k = 1,2,3$, from (220)

$$\begin{aligned}\sigma_{max}\left(\frac{1}{c_k\, card(\mathscr{I}_0^c)}\mathbf{S}_k\right) &\leq \sigma_{max}\left(\frac{1}{c_k\, card(\mathscr{I}_0^c)}\mathbf{E}_k\right) - \zeta \\ &\leq 1+\delta-\zeta,\end{aligned} \tag{145}$$

for some constant $\zeta \in (0,1)$, and some $\delta \in (0,1)$. Thus, from (221)

$$\|\mathbf{S}_k\mathbf{u}\|_2^2 = \left|\mathbf{u}^H\mathbf{S}_k^H\mathbf{S}_k\mathbf{u}\right| \leq (1+\delta-\zeta)c_k card(\mathscr{I}_0^c), \tag{146}$$

holds with probability at least $1-2e^{-Cn}$, provided that L is sufficiently large. Considering the fact

that $\mathbf{x}^{\perp} = \mathbf{u}$, then

$$\|\mathbf{S}_k\mathbf{x}^{\perp}\|_2^2 \leq (1+\delta-\zeta)c_k card(\mathscr{I}_0^c), \tag{147}$$

with high probability. Thus, the result holds. □

Lemma 10.0.3. In the setup of Lemma 10.0.1, the following holds with probability at least $1-2e^{-Cn}$

$$\|\mathbf{S}_k\mathbf{x}\|_2^2 \geq (1-\delta)c_k card(\mathscr{I}_0^c), \tag{148}$$

with $\delta \in (0,1)$ provided that L is sufficiently large.

Demostración. Notice that the left side term in (148) can be seen as

$$\|\mathbf{S}_k\mathbf{x}\|_2^2 = \sum_{i \in \mathscr{I}_0^c} \frac{|\mathbf{a}_{i,k}^H\mathbf{x}|^2}{\|\mathbf{a}_{i,k}\|_2^2}. \tag{149}$$

Given the fact that $\mathbf{S}_k$ is a sub-matrix of $\mathbf{E}_k$ for each $k = 1,2,3$, from (220) it can be obtained

$$\begin{aligned}\sigma_{min}\left(\frac{1}{c_k\, card(\mathscr{I}_0^c)}\mathbf{S}_k\right) &\geq \sigma_{min}\left(\frac{1}{c_k\, card(\mathscr{I}_0^c)}\mathbf{E}_k\right) \\ &\geq 1-\delta,\end{aligned} \tag{150}$$

for some constant $\delta \in (0,1)$ with probability at least $1-2e^{-Cn}$ as long as L is sufficiently large,

for some constant $C > 0$. Thus, from (150) it can be concluded that

$$\sum_{i \in \mathscr{I}_0^c} \frac{|\mathbf{a}_{i,k}^H \mathbf{x}|^2}{\|\mathbf{a}_{i,k}\|_2^2} \geq (1-\delta) c_k card(\mathscr{I}_0^c), \tag{151}$$

with probability at least $1 - 2e^{-Cn}$ as long as L is sufficiently large. Thus, from (151)

$$\|\mathbf{S}_k \mathbf{x}\|_2^2 \geq (1-\delta) c_k card(\mathscr{I}_0^c), \tag{152}$$

holds with probability at least $1-2e^{-Cn}$, provided that L is sufficiently large. Therefore, from (152) the result holds. □

Hence, putting together (139) and (152) it can be concluded that

$$\frac{\|\mathbf{S}_k \mathbf{u}\|_2^2}{\|\mathbf{S}_k \mathbf{x}\|_2^2} \leq \frac{1+\delta-\zeta}{1-\delta} \triangleq \kappa < 1, \tag{153}$$

by taking $\delta < \zeta/2$. Thus, putting together (138) and (153) it can be obtained that

$$\sin^2(\theta) = 1 - \cos^2(\theta) \leq \kappa. \tag{154}$$

On the other hand, notice that

$$\begin{aligned} dist^2(\mathbf{x},\mathbf{z}_0) &= \|\mathbf{x}\|_2^2 + \|\mathbf{z}_0\|_2^2 - 2|\mathbf{x}^H\mathbf{z}_0| \\ &= \|\mathbf{x}\|_2^2 + \|\mathbf{z}_0\|_2^2 - 2\cos(\theta) \\ &\leq 2(1-\sqrt{1-\kappa}). \end{aligned} \tag{155}$$

Then, combining (237) and (236) it can be finally concluded that

$$dist^2(\mathbf{x},\mathbf{z}_0) < 1. \tag{156}$$

Thus, in (156) the result holds.

Appendix C. Proof of Lemma 8.1.1

Let consider some notation before proving Lemma 8.1.1. For two integers a and b we use $a \overset{n}{\equiv} b$ to denote congruence of a and b modulo n (n divides $a-b$). Define $z_p^{(k)} = \frac{1}{n}\sum_{q=1}^{n} Z_{q,p}^{k}$, and without of loss of generality assume that $L = 1$. Further, to develop this analysis, take $\mu = \mathbb{E}[d]$, $w = d - \mu$, $\mathbb{E}\left[|w|^2\right] = \rho_1$, and $\mathbb{E}\left[|w|^4\right] = \rho_2$.

For simplicity x_a, d_a, and q_a refer as the a-th entry of vector $\mathbf{x}$, the diagonal matrices $\mathbf{D}$, and $\mathbf{Q}$, respectively. Also, $\Psi_{a,b}$ is the element at row a and column b of the matrix Ψ. Also, define ρ_3 as

$$\rho_3 = (\rho_2 - \rho_1^2)\left(\min_{c\in\{1,\cdots,n\}} |\Psi_{p,c}|^2\right),$$

and ρ_4 as

$$\rho_4 = \rho_1^2\left(\min_{i\in\{1,\cdots,n\}} \sum_{h\neq i}^{n} |\Psi_{p,h}|^2\right).$$

The proof of this Lemma 8.1.1 proceeds by cases as follow.

Near-zone: Considering the case when $k = 1$, then from (7) and (27) we have $z_p^{(1)}$ is given by

$$\begin{aligned}
&\frac{1}{n^5}\sum_{q=1}^{n} |\mathbf{f}_q^H \mathbf{T}\mathbf{F}^H \mathbf{D}\mathbf{x}|^2 |(\Psi\overline{\mathbf{D}}\mathbf{F}\overline{\mathbf{T}}\mathbf{f}_q)_p|^2 \\
&= \frac{1}{n^3}\sum_{c,e,h,t=1}^{n} d_c\overline{d}_e\overline{d}_h d_t \Psi_{p,h}\overline{\Psi}_{p,t} x_c \overline{x}_e \mathbb{1}_{\{e+h\overset{n}{\equiv}c+t\}} + \upsilon_1^{(1)},
\end{aligned} \tag{157}$$

where $\upsilon_1^{(1)}$ can be considered as a constant for this analysis. Considering the definition of w, then

from (157) it can be obtained that $\mathbb{E}[z_p^{(1)}]$ can be written as

$$\begin{aligned}&\frac{1}{n^3}\sum_{c,e,h,t=1}^{n}\mathbb{E}\left[w_c\overline{w}_e\overline{w}_h w_t\right]\Psi_{p,h}\overline{\Psi}_{p,t}x_c\bar{x}_e\mathbb{1}_{\{e+h\overset{n}{\equiv}c+t\}}\\&+\frac{|\mu|^4}{n^3}\sum_{c,e,h,t=1}^{n}x_c\bar{x}_e\overline{\Psi}_{p,t}\Psi_{p,h}\mathbb{1}_{\{e+h\overset{n}{\equiv}c+t\}}+\upsilon_2^{(1)},\end{aligned}\tag{158}$$

where $\upsilon_2^{(1)}$ can be considered as a constant for this analysis. Observe that $\mathbb{E}\left[w_c\overline{w}_e w_t\overline{w}_h\right]=0$ unless $(c=e,h=t)$, or $(e=t,h=c,e\neq c)$, where these two conditions also satisfy that $e+h\overset{n}{\equiv}c+t$. Thus

- $(e=t,h=c)$: Then, the first term in (158) can be expressed as

$$\begin{aligned}&\frac{\rho_2}{n^3}\sum_{c=1}^{n}|x_c|^2|\Psi_{p,c}|^2+\frac{\rho_1^2}{n^3}\sum_{c=1}^{n}\sum_{e\neq c}^{n}x_c\bar{x}_e\Psi_{p,c}\overline{\Psi}_{p,e}\\&=\frac{(\rho_2-\rho_1^2)}{n^3}\sum_{c=1}^{n}|x_c|^2|\Psi_{p,c}|^2+\frac{\rho_1^2}{n^2}|(\theta)_p|^2\\&\geq\frac{\rho_3}{n^3}\|\mathbf{x}\|_2^2+\frac{\rho_1^2}{n^2}|(\theta)_p|^2,\end{aligned}\tag{159}$$

  Observe that by the Jensen's inequality $\rho_2-\rho_1^2\geq 0$.

- $(c=e,h=t,c\neq h)$: Thus, the first term in (158) can be rewritten as

$$\frac{\rho_1^2}{n^3}\sum_{c=1}^{n}\sum_{h\neq c}^{n}|x_c|^2|\Psi_{p,h}|^2\geq\frac{\rho_4}{n^3}\|\mathbf{x}\|_2^2,\tag{160}$$

Thus, combining (159), and (160) it can be concluded that

$$\mathbb{E}[z_p] \geq \frac{(\rho_3+\rho_4)}{n^3}\|\mathbf{x}\|_2^2 + \frac{(\rho_1^2+|\mu|^4)}{n^3}|(\theta)_p|^2 + \upsilon_2^{(1)}. \tag{161}$$

Middle-zone: Considering the case when $k=2$, then from (7) and (27) it can be obtained that $z_p^{(2)}$ is given by

$$\begin{aligned}&\frac{1}{n^3}\sum_{q=1}^{n}|\mathbf{f}_q^T\mathbf{QDx}|^2|(\Psi\overline{\mathbf{DQf}}_q)_p|^2\\&=\frac{1}{n^2}\sum_{c,e,h,t=1}^{n} q_c\overline{q}_e\overline{q}_h q_t d_c\overline{d}_e\overline{d}_h d_t\Psi_{p,h}\overline{\Psi}_{p,t}x_c\overline{x}_e\mathbb{1}_{\{e+h\overset{n}{\equiv}c+t\}},\end{aligned} \tag{162}$$

where the notation $\mathbb{1}_{\{e+h\overset{n}{\equiv}c+t\}}$ is equal to one if condition $e+h\overset{n}{\equiv}c+t$ is satisfied, and zero otherwise. Considering the definition of w, then from (162), $\mathbb{E}[z_p^{(2)}]$ can be written as

$$\begin{aligned}&\frac{1}{n^2}\sum_{c,e,h,t=1}^{n}\mathbb{E}\left[w_c\overline{w}_e w_t\overline{w}_h\right]q_c\overline{q}_e\overline{q}_h q_t x_c\overline{x}_e\overline{\Psi}_{p,t}\Psi_{p,h}\mathbb{1}_{\{e+h\overset{n}{\equiv}c+t\}}\\&+\frac{|\mu|^4}{n^2}\sum_{c,e,h,t=1}^{n}q_c\overline{q}_e\overline{q}_h q_t x_c\overline{x}_e\overline{\Psi}_{p,t}\Psi_{p,h}\mathbb{1}_{\{e+h\overset{n}{\equiv}c+t\}}+\upsilon_1^{(2)}\\&=\frac{1}{n^2}\sum_{c,e,h,t=1}^{n}\mathbb{E}\left[w_c\overline{w}_e w_t\overline{w}_h\right]q_c\overline{q}_e\overline{q}_h q_t x_c\overline{x}_e\overline{\Psi}_{p,t}\Psi_{p,h}\mathbb{1}_{\{e+h\overset{n}{\equiv}c+t\}}\\&+\frac{|\mu|^4}{n^2}|(\theta)_p|^2+\upsilon_2^{(2)},\end{aligned} \tag{163}$$

where $\upsilon_1^{(2)}$ in the first line can be considered as a constant for this analysis, and $\upsilon_2^{(2)}$ in the second

line is given by

$$\upsilon_1^{(2)} + \frac{|\mu|^4}{n^2} \sum_{c=1}^{n} \sum_{t\neq c}^{n} \sum_{e=1}^{n} \sum_{h\neq e}^{n} q_c \overline{q}_e \overline{q}_h q_t x_c \bar{x}_e \overline{\Psi}_{p,t} \Psi_{p,h} \mathbb{1}_{\{e+h \overset{n}{\equiv} c+t\}}. \tag{164}$$

Observe that $\mathbb{E}[w_c \overline{w}_e w_t \overline{w}_h] = 0$ unless $(c = e, h = t)$, or $(e = t, h = c, e \neq c)$, where these two conditions also satisfy that $e + h \overset{n}{\equiv} c + t$. Thus

- $(e = t, h = c)$: Then, the first term in (169) can be expressed as

$$\begin{aligned} &\frac{\rho_2}{n^2} \sum_{c=1}^{n} |x_c|^2 |\Psi_{p,c}|^2 + \frac{\rho_1^2}{n^2} \sum_{c=1}^{n} \sum_{e\neq c}^{n} x_c \bar{x}_e \Psi_{p,c} \overline{\Psi}_{p,e} \\ &= \frac{(\rho_2 - \rho_1^2)}{n^2} \sum_{c=1}^{n} |x_c|^2 |\Psi_{p,c}|^2 + \frac{\rho_1^2}{n^2} |(\theta)_p|^2 \\ &\geq \frac{\rho_3}{n^2} \|\mathbf{x}\|_2^2 + \frac{\rho_1^2}{n^2} |(\theta)_p|^2. \end{aligned} \tag{165}$$

  Observe that by the Jensen's inequality $\rho_2 - \rho_1^2 \geq 0$.

- $(c = e, h = t, c \neq h)$: Thus, the first term in (169) can be rewritten as

$$\frac{\rho_1^2}{n^2} \sum_{c=1}^{n} \sum_{h\neq c}^{n} |x_c|^2 |\Psi_{p,h}|^2 \geq \frac{\rho_4}{n^2} \|\mathbf{x}\|_2^2, \tag{166}$$

  Thus, combining (169), and (166) it can be concluded that

$$\mathbb{E}[z_p] \geq \frac{(\rho_3 + \rho_4)}{n^2} \|\mathbf{x}\|_2^2 + \frac{(\rho_1^2 + |\mu|^4)}{n^2} |(\theta)_p|^2 + \upsilon_2^{(2)}. \tag{167}$$

Far-zone: Considering the case when $k = 3$, then from (7) and (27), $z_p^{(3)}$ is given by

$$\begin{aligned}
&\frac{1}{n^3}\sum_{q=1}^{n}|\mathbf{f}_q^H\mathbf{D}\mathbf{x}|^2|(\Psi\overline{\mathbf{D}}\mathbf{f}_q)_p|^2 \\
&=\frac{1}{n^2}\sum_{c,e,h,t=1}^{n} d_c\overline{d}_e d_t\overline{d}_h x_c\overline{x}_e\overline{\Psi}_{p,t}\Psi_{p,h}\mathbb{1}_{\{e+h\overset{n}{\equiv}c+t\}},
\end{aligned} \tag{168}$$

Therefore, from (168), $\mathbb{E}[z_p^{(3)}]$ can be written as

$$\begin{aligned}
&\frac{1}{n^2}\sum_{c,e,h,t=1}^{n}\mathbb{E}\left[w_c\overline{w}_e w_t\overline{w}_h\right]x_c\overline{x}_e\overline{\Psi}_{p,t}\Psi_{p,h}\mathbb{1}_{\{e+h\overset{n}{\equiv}c+t\}} \\
&+\frac{|\mu|^4}{n^2}\sum_{c,e,h,t=1}^{n}x_c\overline{x}_e\overline{\Psi}_{p,t}\Psi_{p,h}\mathbb{1}_{\{e+h\overset{n}{\equiv}c+t\}}+\upsilon_1^{(3)} \\
&=\frac{1}{n^2}\sum_{c,e,h,t=1}^{n}\mathbb{E}\left[w_c\overline{w}_e w_t\overline{w}_h\right]x_c\overline{x}_e\overline{\Psi}_{p,t}\Psi_{p,h}\mathbb{1}_{\{e+h\overset{n}{\equiv}c+t\}} \\
&+\frac{|\mu|^4}{n^2}|(\theta)_p|^2+\upsilon_2^{(3)},
\end{aligned} \tag{169}$$

where $\upsilon_1^{(3)}$ in the first can be considered as a constant for this analysis, and $\upsilon_2^{(3)}$ in the second line is given by

$$\upsilon_2^{(3)}+\frac{|\mu|^4}{n^2}\sum_{c=1}^{n}\sum_{t\neq c}^{n}\sum_{e=1}^{n}\sum_{h\neq e}^{n}x_c\overline{x}_e\overline{\Psi}_{p,t}\Psi_{p,h}\mathbb{1}_{\{e+h\overset{n}{\equiv}c+t\}}. \tag{170}$$

Observe that $\mathbb{E}\left[w_c\overline{w}_e w_t\overline{w}_h\right]=0$ unless $(c=e,h=t)$, or $(e=t,h=c,e\neq c)$, where these two conditions also satisfy that $e+h\overset{n}{\equiv}c+t$. Thus

- $(e = t, h = c)$: Then, the first term in (169) can be expressed as

$$\begin{aligned}
&\frac{\rho_2}{n^2}\sum_{c=1}^{n}|x_c|^2|\Psi_{p,c}|^2+\frac{\rho_1^2}{n^2}\sum_{c=1}^{n}\sum_{e\neq c}^{n}x_c\bar{x}_e\Psi_{p,c}\overline{\Psi}_{p,e}\\
&=\frac{(\rho_2-\rho_1^2)}{n^2}\sum_{c=1}^{n}|x_c|^2|\Psi_{p,c}|^2+\frac{\rho_1^2}{n^2}|(\theta)_p|^2\\
&\geq\frac{\rho_3}{n^2}\|\mathbf{x}\|_2^2+\frac{\rho_1^2}{n^2}|(\theta)_p|^2, \qquad (171)
\end{aligned}$$

- $(c = e, h = t, c \neq h)$: Thus, the first term in (169) can be rewritten as

$$\frac{\rho_1^2}{n^2}\sum_{c=1}^{n}\sum_{h\neq c}^{n}|x_c|^2|\Psi_{p,h}|^2\geq\frac{\rho_4}{n^2}\|\mathbf{x}\|_2^2, \qquad (172)$$

Thus, combining (169), and (172) it can be concluded that

$$\mathbb{E}[z_p]\geq\frac{(\rho_3+\rho_4)}{n^2}\|\mathbf{x}\|_2^2+\frac{(\rho_1^2+|\mu|^4)}{n^2}|(\theta)_p|^2+\upsilon_2^{(3)}. \qquad (173)$$

Then, since ρ_1, and ρ_2 are always greater than zero, it is clear from (161), (167), and (173) that when $\mathbb{E}[d] \neq 0$, then the non-zero coefficients can be better estimated.

Appendix D. Proof of Theorem 5.1.3

Demostración. From (31) and (44) it can be obtained that

$$|K(\mathbf{x},\mu)| = \frac{1}{m}\left|\sum_{k=1}^{m}\left(\varphi_\mu(|\mathbf{a}_k^H\mathbf{x}|)-q_k\right)^2-\left(\varphi_0(|\mathbf{a}_k^H\mathbf{x}|)-q_k\right)^2\right|, \tag{174}$$

where $K(\mathbf{x},\mu) = g(\mathbf{x},\mu) - f(\mathbf{x})$. Note that the right hand side of the equality in (174) can be rewritten as

$$\frac{1}{m}\left|\sum_{k=1}^{m}\varphi_\mu^2(|\mathbf{a}_k^H\mathbf{x}|)-\varphi_0^2(|\mathbf{a}_k^H\mathbf{x}|)-2q_k\left(\varphi_\mu(|\mathbf{a}_k^H\mathbf{x}|)-\varphi_0(|\mathbf{a}_k^H\mathbf{x}|)\right)\right|. \tag{175}$$

By definition of the function $\varphi_\mu(\cdot)$ in (40), and from (175) it can be concluded that

$$\varphi_\mu^2(|\mathbf{a}_k^H\mathbf{x}|)-\varphi_0^2(|\mathbf{a}_k^H\mathbf{x}|) = \mu^2. \tag{176}$$

By combining (175) and (176), and applying the triangular inequality, it can be obtained that

$$|g(\mathbf{x},\mu)-f(\mathbf{x})| \leq \frac{1}{m}\sum_{k=1}^{m}\mu^2+2q_k\left|\varphi_\mu(|\mathbf{a}_k^H\mathbf{x}|)-\varphi_0(|\mathbf{a}_k^H\mathbf{x}|)\right|. \tag{177}$$

Using the fact that the function $\varphi_\mu(\cdot)$ uniformly approximates the function $\varphi_0(\cdot)$ as was proved in Lemma 5.1.2, the above inequality can be expressed as

$$|g(\mathbf{x},\mu)-f(\mathbf{x})| \leq \frac{1}{m}\sum_{k=1}^{m}\mu^2+2q_k\mu. \tag{178}$$

Therefore by taking $q^{max} = \text{máx}\{q_k | k = 1, \cdots, m\}$, from (178) it can be obtained that

$$|g(\mathbf{x}, \mu) - f(\mathbf{x})| \leq \frac{1}{m} \left(\sum_{k=1}^{m} \mu^2 + 2\mu q^{max} \right) = \mu \kappa_1, \tag{179}$$

where $\kappa_1 = (\mu + 2q^{max})$. Thus, the result holds.

On the other hand, given the fact that for a fixed $\mu > 0$, the Wirtinger derivative $\partial g(\mathbf{z}, \mu)$ in (95) is continuous in $\mathbf{z}$. In fact, if $\mu = 0$, then (95) becomes the update direction in Wang et al. (2016a) which is non-continuous. Then, the function g in (44) is smooth in $\mathbf{z}$ because $\partial g(\mathbf{z}, \mu)$ is continuous in $\mathbf{z}$. □

Appendix E. Proof of Theorem 5.2.2

To prove Theorem 5.2.2, the following Lemma 10.0.4 is introduced, which is useful to prove item 2).

Lemma 10.0.4. Assume that f_1 and f_2 are Lipschitz continuous functions on a bounded set I with constants L_1 and L_2, and there is a constant $v > 0$ such that $f_2(x) \geq v$ for all $x \in I$. Then f_1/f_2 is Lipschitz continuous on I. (The proof of Lemma 10.0.4 can be found in Chapter 12 in Eriksson et al. (2013)).

Demostración. 1) Suppose that $S_\mu(\mathbf{w})$ in (51) is unbounded, then there exists a sequence $\{\mathbf{x}_\ell\} \subseteq S_\mu(\mathbf{w})$ such that $\|\mathbf{x}_\ell\|_2 \to \infty$. From the definition of the level set $S_\mu(\mathbf{w})$, it can be obtained that

$$g(\mathbf{x}_\ell, \mu) \leq g(\mathbf{w}, \mu) < \infty, \forall \ell \in \mathbb{N}. \tag{180}$$

However, if $\text{span}(\mathbf{a}_1, \cdots, \mathbf{a}_m) = \mathbb{C}^n$, then the fact that $\|\mathbf{x}_\ell\|_2 \to \infty$ implies that the sequence $g(\mathbf{x}_\ell, \mu) \to \infty$ according to the definition of function g. Then $g(\mathbf{x}_\ell, \mu) \to \infty$ is a contradiction, because $g(\mathbf{x}_\ell, \mu) < \infty$, $\forall \ell \in \mathbb{N}$. Thus, $S_\mu(\mathbf{w})$ is a bounded set.

2) To prove the second part of Assumption 1, we proceed to show that for each function $h_{k,\mu}(\mathbf{x}) = \left(\varphi_\mu(|\mathbf{a}_k^H \mathbf{x}|) - q_k\right)^2$ its Wirtinger derivative is Lipschitz. Thus, since $g(\mathbf{x}, \mu)$ is the sum of the functions $h_{k,\mu}(\mathbf{x})$, then the Writinger derivative of $g(\mathbf{x}, \mu)$ is Lipschitz as it is proven in Chapter 12 in Eriksson et al. (2013).

Notice that, Wirtinger derivative of $h_{k,\mu}$ at point $\mathbf{w} \in \mathbb{C}^n$ is given by

$$\begin{aligned}\partial h_{k,\mu}(\mathbf{w}) &= 2\left(\varphi_\mu\left(|\mathbf{a}_k^H\mathbf{w}|\right) - q_k\right)\frac{\mathbf{a}_k^H\mathbf{w}}{\varphi_\mu(|\mathbf{a}_k^H\mathbf{w}|)}\mathbf{a}_k \\ &= 2\left((\mathbf{a}_k^H\mathbf{w})\mathbf{a}_k - q_k\frac{\mathbf{a}_k^H\mathbf{w}}{\varphi_\mu(|\mathbf{a}_k^H\mathbf{w}|)}\mathbf{a}_k\right). \end{aligned} \tag{181}$$

By definition of $d_r(\cdot,\cdot)$ in Eq.(3), it can be obtained that

$$d_r(\partial h_{k,\mu}(\mathbf{w}_1), \partial h_{k,\mu}(\mathbf{w}_2)) \leq \|e^{-j\theta}\partial h_{k,\mu}(\mathbf{w}_1) - \partial h_{k,\mu}(\mathbf{w}_2)\|_2, \tag{182}$$

for any $\mathbf{w}_1, \mathbf{w}_2 \in S_\mu(\mathbf{w})$ and $\theta \in [0, 2\pi)$. Then, combining (181) and (182), one can write that

$$\begin{aligned} & d_r(\partial h_{k,\mu}(\mathbf{w}_1), \partial h_{k,\mu}(\mathbf{w}_2)) \leq 2\|\mathbf{a}_k\|_2\left|e^{-j\theta}(\mathbf{a}_k^H\mathbf{w}_1) - \mathbf{a}_k^H\mathbf{w}_2\right| \\ & + 2q_k\|\mathbf{a}_k\|_2\left|\frac{e^{-j\theta}(\mathbf{a}_k^H\mathbf{w}_1)}{\varphi_\mu(|\mathbf{a}_k^H\mathbf{w}_1|)} - \frac{\mathbf{a}_k^H\mathbf{w}_2}{\varphi_\mu(|\mathbf{a}_k^H\mathbf{w}_2|)}\right| \\ & \leq 2\|\mathbf{a}_k\|_2^2\|e^{-j\theta}\mathbf{w}_1 - \mathbf{w}_2\|_2 \\ & + \frac{2q_k\|\mathbf{a}_k\|_2}{\mu^2}\left|e^{-j\theta}(\mathbf{a}_k^H\mathbf{w}_1)\varphi_\mu(|\mathbf{a}_k^H\mathbf{w}_2|) - (\mathbf{a}_k^H\mathbf{w}_2)\varphi_\mu(|\mathbf{a}_k^H\mathbf{w}_1|)\right|, \end{aligned} \tag{183}$$

where the first inequality is obtained using the triangular inequality and the second comes from the fact that $\varphi_\mu(t) \geq \mu > 0$ for all $t \in \mathbb{R}$, and using the Cauchy-Schwarz inequality. Then, from (183)

it can be obtained that

$$\begin{aligned}
&\left|e^{-j\theta}(\mathbf{a}_k^H\mathbf{w}_1)\varphi_\mu(|\mathbf{a}_k^H\mathbf{w}_2|)-(\mathbf{a}_k^H\mathbf{w}_2)\varphi_\mu(|\mathbf{a}_k^H\mathbf{w}_1|)\right| \\
&\leq \left|\varphi_\mu(|\mathbf{a}_k^H\mathbf{w}_2|)\left[e^{-j\theta}(\mathbf{a}_k^H\mathbf{w}_1)-\mathbf{a}_k^H\mathbf{w}_2\right]\right| \\
&+\left|(\mathbf{a}_k^H\mathbf{w}_2)\left[\varphi_\mu(|\mathbf{a}_k^H\mathbf{w}_1|)-\varphi_\mu(|\mathbf{a}_k^H\mathbf{w}_2|)\right]\right| \\
&\leq M_{\varphi_\mu}\|\mathbf{a}_k\|_2\|e^{-j\theta}\mathbf{w}_1-\mathbf{w}_2\|_2 \\
&+M_{S_\mu}\|\mathbf{a}_k\|_2\left|\varphi_\mu(|\mathbf{a}_k^H\mathbf{w}_1|)-\varphi_\mu(|\mathbf{a}_k^H\mathbf{w}_2|)\right|, \qquad (184)
\end{aligned}$$

where the second inequality is obtained using the triangular inequality and the following two reasons. First, $\varphi_\mu(|\mathbf{a}_k^H\mathbf{z}|)$ is a bounded function in $S_\mu(\mathbf{w})$ for any $\mathbf{z}\in S_\mu(\mathbf{w})$, since $S_\mu(\mathbf{w})$ is a closed and bounded set as was established in the previous item and $\varphi_\mu(\cdot)$ is a continuous function, *i.e.* $\varphi_\mu(|\mathbf{a}_k^H\mathbf{z}|)\leq M_{\varphi_\mu}$ for some constant $M_{\varphi_\mu}\in\mathbb{R}_+$. Second, $S_\mu(\mathbf{w})$ is a bounded set, then $\|\mathbf{z}\|_2\leq M_{S_\mu},\forall\mathbf{z}\in S_\mu(\mathbf{w})$ for some constant $M_{S_\mu}\in\mathbb{R}_+$. Hence, considering that $\varphi_\mu(\cdot)$ is a Lipschitz function with constant $L_{\varphi_\mu}=1$, then from (184)

$$\begin{aligned}
\left|\varphi_\mu(|\mathbf{a}_k^H\mathbf{w}_1|)-\varphi_\mu(|\mathbf{a}_k^H\mathbf{w}_2|)\right| &\leq \left||\mathbf{a}_k^H\mathbf{w}_1|-|\mathbf{a}_k^H\mathbf{w}_2|\right| \\
&\leq \left|e^{-j\theta}(\mathbf{a}_k^H\mathbf{w}_1)-\mathbf{a}_k^H\mathbf{w}_2\right| \\
&\leq \|\mathbf{a}_k\|_2\|e^{-j\theta}\mathbf{w}_1-\mathbf{w}_2\|_2, \qquad (185)
\end{aligned}$$

where the second and third lines come from the triangular and Cauchy-Schwarz inequality, respec-

tively, and it is valid for all $\theta \in [0, 2\pi)$. Therefore, combining (183), (184) and (185) yields

$$d_r(\partial h_{k,\mu}(\mathbf{w}_1), \partial h_{k,\mu}(\mathbf{w}_2)) \leq \tilde{L}_{h_{k,\mu}} \left\| e^{-j\theta}\mathbf{w}_1 - \mathbf{w}_2 \right\|_2, \tag{186}$$

with $\tilde{L}_{h_{k,\mu}} = 2\|\mathbf{a}_k\|_2^2 + \frac{2q_k M_{\varphi_\mu} \|\mathbf{a}_k\|_2^2}{\mu^2} + \frac{2q_k M_{S_\mu} \|\mathbf{a}_k\|_2^3}{\mu^2}$. Notice that, for the *i.i.d*. Gaussian vectors $\mathbf{a}_k$, $\|\mathbf{a}_k\|_2^2 \leq 2.3n$ holds with probability at least $1 - me^{-n/2}$ Wang et al. (2016a). Then, $\tilde{L}_{h_{k,\mu}} \leq 4.6n + \frac{4.6q_k n}{\mu^2} + \frac{13n^{3/2}M_{S_\mu}}{\mu^2} = L_{h_{k,\mu}}$ with probability exceeding $1 - me^{-n/2}$. Further, taking the value of θ that minimizes the term $\|e^{-j\theta}\mathbf{w}_1 - \mathbf{w}_2\|_2$, from (186), it can be concluded that

$$d_r(\partial h_{k,\mu}(\mathbf{w}_1), \partial h_{k,\mu}(\mathbf{w}_2)) \leq L_{h_{k,\mu}} d_r(\mathbf{w}_1, \mathbf{w}_2), \tag{187}$$

with probability at least $1 - me^{-n/2}$. Thus, from (187) the result holds.

2) Note that, the function $\varphi'_\mu(|\mathbf{a}_k^H\mathbf{z}|)$ can be expressed as $\varphi'_\mu(|\mathbf{a}_k^H\mathbf{z}|) = \frac{f_1(|\mathbf{a}_k^H\mathbf{z}|)}{f_2(|\mathbf{a}_k^H\mathbf{z}|)}$, where $f_1(|\mathbf{a}_k^H\mathbf{z}|) = |\mathbf{a}_k^H\mathbf{z}|$ and $f_2(|\mathbf{a}_k^H\mathbf{z}|) = \sqrt{|\mathbf{a}_k^H\mathbf{z}|^2 + \mu^2}$. Notice that, by definition of f_1 and using (185) one can write that

$$\begin{aligned} \left| f_1(|\mathbf{a}_k^H\mathbf{z}_1|) - f_1(|\mathbf{a}_k^H\mathbf{z}_2|) \right| &\leq \|\mathbf{a}_k\|_2 d_r(\mathbf{z}_1, \mathbf{z}_2) \\ &\leq \sqrt{2.3n} d_r(\mathbf{z}_1, \mathbf{z}_2), \end{aligned} \tag{188}$$

taking the value of θ that minimizes the term $\|e^{-j\theta}\mathbf{z}_1 - \mathbf{z}_2\|_2$ with probability at least $1 - me^{-n/2}$. Then, from (188), $f_1(|\mathbf{a}_k^H\mathbf{z}|)$ is Lipschitz continuous with probability exceeding $1 - me^{-n/2}$. Also, it was previously established that $f_2(|\mathbf{a}_k^H\mathbf{z}|)$ are Lipschitz continuous functions with high probability.

Notice that, $f_2(|\mathbf{a}_k^H\mathbf{z}|) = \sqrt{|\mathbf{a}_k^H\mathbf{z}|^2+\mu^2} \geq \mu > 0$, for any fixed μ. Now, considering the fact that $S_\mu(\mathbf{w})$ is a bounded set from item 1) and the previous conditions over functions $f_1(|\mathbf{a}_k^H\mathbf{z}|)$ and $f_2(|\mathbf{a}_k^H\mathbf{z}|)$, it can be obtained that $\varphi'_\mu(|\mathbf{a}_k^H\mathbf{z}|)$ is a Lipschitz continuous function on $S_\mu(\mathbf{w})$ with probability at least $1-me^{-n/2}$, because the hypotheses in Lemma 10.0.4 are satisfied. □

Appendix F. Proof of Theorem 5.2.2

Denote $\mathscr{K} := \{k|\mu_{k+1} = \gamma_1\mu_k\}$ with $\gamma_1 \in (0,1)$. If $\mathscr{K}$ is finite, then according to Lines 10-12 in Algorithm 1 there exists an integer $\overline{k}$ such that for all $i > \overline{k}$ it can be obtained that $\|\partial g(\mathbf{x}_i, \mu_{i-1})\|_2 \geq \gamma\mu_{i-1}$, where $\mu_i = \mu_{\overline{k}}$ and $\gamma \in (0,1)$. Taking $\overline{\mu} = \mu_{\overline{k}}$, the optimization problem solved by Algorithm 1, reduces to solve

$$\min_{\mathbf{x}\in\mathbb{R}^n/\mathbb{C}^n} g(\mathbf{x}, \overline{\mu}). \tag{189}$$

Hence, assuming the setup of Theorem 2, the function $g(\cdot, \overline{\mu})$ satisfies Assumption 1. Further, since Algorithm 1 implements a backtracking strategy and the Assumption 1 is satisfied, by Theorem 5.7 in Wright and Nocedal (1999) it can be obtained that

$$\liminf_{i\to\infty} \|\partial g(\mathbf{x}_i, \mu_{i-1})\|_2 = 0. \tag{190}$$

Notice that, (190) contradicts the fact that $\|\partial g(\mathbf{x}_i, \mu_{i-1})\|_2 \geq \gamma\mu_{i-1}$ for all $i > \overline{k}$. This shows that $\mathscr{K}$ must be infinite and $\lim_{i\to\infty}\mu_i = 0$. Given that $\mathscr{K}$ is infinite, then $\mathscr{K} = \{k_0, k_1, \cdots,\}$ with

$k_0 < k_1 < \cdots$. Thus

$$\liminf_{i\to\infty} \|\partial g(\mathbf{x}_i, \mu_{i-1})\|_2 \leq \gamma \lim_{i\to\infty} \mu_i = 0. \tag{191}$$

Therefore, from (191) the result holds.

Appendix G. Proof of Theorem 5.2.3

To prove Theorem 4, we need to introduce the following Definition 5 and Lemma 6, to determine $\partial^c f(\mathbf{x})$.

Definition 5. *Regular function* (Definition 3.5 in Clarke (1990)): A function h is said to be regular at $\mathbf{x}$ provided that

1. For all $\mathbf{w} \in \mathbb{R}^n$, the usual one-sided directional derivative $h'(\mathbf{x};\mathbf{w})$ exists.
2. For all $\mathbf{w} \in \mathbb{R}^n$, $h'(\mathbf{x};\mathbf{w}) = \limsup_{\substack{\mathbf{y}\to\mathbf{x} \\ t\downarrow 0}} \frac{h(\mathbf{y}+t\mathbf{w})-h(\mathbf{y})}{t}$.

Lemma 6. The items of this lemma are proved in Theorems 3.13, 3.16 and 3.19 in Clarke (1990), respectively.

1. A Lipschitz continuous function $h : \mathbb{R}^n \to \mathbb{R}$ is regular at $\mathbf{x}$ if h is convex or smooth at $\mathbf{x}$.
2. Suppose that $h_i : \mathbb{R}^n \to \mathbb{R}, i = 1, \cdots, m$ are Lipschitz continuous near $\mathbf{x}$. Then their sum $h = \sum_{i=1}^n \lambda_i h_i$ is also Lipschitz continuous near $\mathbf{x}$ and

$$\partial^c h(\mathbf{x}) = \partial^c \left(\sum_{i=1}^n \lambda_i h_i \right)(\mathbf{x}) \subseteq \sum_{i=1}^n \lambda_i \partial^c h_i(\mathbf{x}), \tag{192}$$

where $\sum_{i=1}^{n} \lambda_i \partial^c h_i(\mathbf{x}) = \{\sum_{i=1}^{n} \lambda_i \mathbf{w}_i : \mathbf{w}_i \in \partial^c h_i(\mathbf{x})\}$. If h is regular at $\mathbf{x}$, equality holds.

3. Let $G(\mathbf{x}) = h(P(\mathbf{x}))$, where $P : \mathbb{R}^n \to \mathbb{R}^m$ is Lipschitz continuous near $\mathbf{x}$ and $h : \mathbb{R}^m \to \mathbb{R}$ is Lipschitz continuous near $P(\mathbf{x})$. Then G is Lipschitz continuous near $\mathbf{x}$ and

$$\partial^c G(\mathbf{x}) \subset \overline{\text{conv}}\Big\{ \sum_{i=1}^{m} \alpha_i \zeta_i | \zeta_i \in \partial^c P(\mathbf{x}), \\ \alpha = (\alpha_1, \cdots, \alpha_m)^T \in \partial^c h(P(\mathbf{x}))\Big\}. \tag{193}$$

If h is regular at $P(\mathbf{x})$ and P is smooth at $\mathbf{x}$, equality holds.

Remark that, any complex vector $\mathbf{w} \in \mathbb{C}^n$ can be uniquely identified with an element $[\mathbf{w}_R, \mathbf{w}_I]^T \in \mathbb{R}^{2n}$, *i.e* $\mathbf{w} \equiv [\mathbf{w}] = \begin{bmatrix} \mathbf{w}_R \\ \mathbf{w}_I \end{bmatrix}$. Now, define $F : \mathbb{R}^2 \to \mathbb{R}_+$ as $F(\mathbf{z}) = \|\mathbf{z}\|_2$. Then, by definition of F, it can be obtained that $|\mathbf{a}_k^H \mathbf{x}| \equiv F([\mathbf{a}_k^H \mathbf{x}])$. Notice that for any $\mathbf{z}_1, \mathbf{z}_2 \in \mathbb{R}^2$ it can be obtained that

$$|F(\mathbf{z}_1) - F(\mathbf{z}_2)| \leq \|\mathbf{z}_1 - \mathbf{z}_2\|_2. \tag{194}$$

Then, from (194) it can be concluded that F is a Lipschitz continuous function. Further, given the fact that F is a convex function, by item 1) in Lemma 6, F is a regular function. Thus, using the result 3) in Lemma 6, it can be concluded that $\partial^c \varphi_0(|\mathbf{a}_k^H \mathbf{x}|) \equiv \partial^c F([\mathbf{a}_k^H \mathbf{x}])$, where from Theorem 3.9 in Clarke (1990) it can be found

$$\partial^c F([\mathbf{a}_k^H \mathbf{x}]) = \left\{ \begin{bmatrix} \mathbf{a}_{k,R} & -\mathbf{a}_{k,I} \\ \mathbf{a}_{k,I} & \mathbf{a}_{k,R} \end{bmatrix} \mathbf{z} : \mathbf{z} \in [-1, 1]^2 \right\}, \tag{195}$$

with $\mathbf{a}_{k,R}, \mathbf{a}_{k,I} \in \mathbb{R}^n$ are the real and imaginary parts of vector $\mathbf{a}_k$, respectively, *i.e.* $\mathbf{a}_k = \mathbf{a}_{k,R} + j\mathbf{a}_{k,I}$. Therefore, given the fact that F is a regular function, from 2) in Lemma 6 one can write that

$$\partial^c f(\mathbf{x}) \equiv \frac{2}{m} \sum_{k=1}^{m} (\varphi_0(|\mathbf{a}_k^H \mathbf{x}|) - q_k) \partial^c F([\mathbf{a}_k^H \mathbf{x}]), \tag{196}$$

where the sum is calculated as defined in 2) in Lemma 6. Now, we prove Theorem 4.

Demostración. We proceed by proving that each function $\varphi_\mu(|\mathbf{a}_k^H \mathbf{x}|)$ satisfies the gradient consistency property. Notice that $\partial \varphi_\mu(|\mathbf{a}_k^H \mathbf{x}|)$ is given by

$$\partial \varphi_\mu(|\mathbf{a}_k^H \mathbf{x}|) = \left(\frac{\mathbf{a}_k^H \mathbf{x}}{\sqrt{|\mathbf{a}_k^H \mathbf{x}|^2 + \mu^2}} \right) \mathbf{a}_k. \tag{197}$$

Then, (197) can be equivalently expressed as $\partial \varphi_\mu(|\mathbf{a}_k^H \mathbf{x}|) \equiv \varphi'_\mu(\|[\mathbf{a}_k^H \mathbf{x}]\|_2)$, where

$$\varphi'_\mu(\|[\mathbf{a}_k^H \mathbf{x}]\|_2) = \begin{bmatrix} \mathbf{a}_{k,R} & -\mathbf{a}_{k,I} \\ \mathbf{a}_{k,I} & \mathbf{a}_{k,R} \end{bmatrix} \frac{[\mathbf{a}_k^H \mathbf{x}]}{\sqrt{\|[\mathbf{a}_k^H \mathbf{x}]\|_2^2 + \mu^2}}. \tag{198}$$

Considering (198), the gradient consistency property for $\varphi_\mu(|\mathbf{a}_k^H \mathbf{x}|)$ can be equivalently formulated as

$$\left\{ \lim_{\substack{\mathbf{x} \to \mathbf{x}^* \\ \mu \downarrow 0}} \partial \varphi_\mu(|\mathbf{a}_k^H \mathbf{x}|) \right\} \equiv \left\{ \lim_{\substack{\mathbf{x} \to \mathbf{x}^* \\ \mu \downarrow 0}} \varphi'_\mu(\|[\mathbf{a}_k^H \mathbf{x}]\|_2) \right\}. \tag{199}$$

Therefore, from (195) and (198), the gradient consistency property for $\varphi_\mu(|\mathbf{a}_k^H\mathbf{x}|)$ holds if

$$\left\{ \lim_{\substack{\mathbf{x}\to\mathbf{x}^* \\ \mu\downarrow 0}} \frac{[\mathbf{a}_k^H\mathbf{x}]}{\sqrt{\|[\mathbf{a}_k^H\mathbf{x}]\|_2^2+\mu^2}} \right\} = [-1,1]^2. \tag{200}$$

Notice that, we need to establish that all limit points of the left hand side term in (200) belong to the set $[-1,1]^2$. To do that, we proceed by cases. First considering $\mathbf{x}^* \neq \mathbf{0}$ from (200) we have that

$$\lim_{\mu\downarrow 0, \mathbf{x}\to\mathbf{x}^*} \frac{[\mathbf{a}_k^H\mathbf{x}]}{\sqrt{\|[\mathbf{a}_k^H\mathbf{x}]\|_2^2+\mu^2}} = \frac{[\mathbf{a}_k^H\mathbf{x}^*]}{\|[\mathbf{a}_k^H\mathbf{x}^*]\|_2}, \tag{201}$$

where $\frac{[\mathbf{a}_k^H\mathbf{x}^*]}{\|[\mathbf{a}_k^H\mathbf{x}^*]\|_2} \in [-1,1]^2$ for any $\mathbf{x}^* \neq \mathbf{0}$. Now, for $\mathbf{x}^* = \mathbf{0}$, we can rewrite (201) as

$$\lim_{\substack{\mathbf{x}\to\mathbf{0} \\ \mu\downarrow 0}} \frac{[\mathbf{a}_k^H\mathbf{x}]}{\sqrt{\|[\mathbf{a}_k^H\mathbf{x}]\|_2^2+\mu^2}} \left(\frac{1/\mu}{1/\mu}\right) = \lim_{\substack{\mathbf{x}\to\mathbf{0} \\ \mu\downarrow 0}} \frac{[\mathbf{a}_k^H\mathbf{x}]/\mu}{\sqrt{\|[\mathbf{a}_k^H\mathbf{x}]\|_2^2/\mu^2+1}}. \tag{202}$$

Then, from (202) we have that

$$\begin{aligned}
&\lim_{\substack{\mathbf{x}\to\mathbf{0} \\ \mu\downarrow 0}} \frac{[\mathbf{a}_k^H\mathbf{x}]/\mu}{\sqrt{\|[\mathbf{a}_k^H\mathbf{x}]\|_2^2/\mu^2+1}} \\
&= \lim_{\substack{\mathbf{x}\to\mathbf{0} \\ \mu\downarrow 0}} \left[\frac{\frac{z_R}{\mu}}{\sqrt{(\frac{z_R}{\mu})^2+(\frac{z_I}{\mu})^2+1}}, \frac{\frac{z_I}{\mu}}{\sqrt{(\frac{z_R}{\mu})^2+(\frac{z_I}{\mu})^2+1}} \right]^T \\
&= \begin{cases}
\mathbf{1}, \text{ if } \frac{z_R}{\mu}, \frac{z_I}{\mu} \to \infty \\
\alpha, \text{ if } \left(|\frac{z_R}{\mu}| \to \infty \vee |\frac{z_R}{\mu}| < \infty\right) \wedge \left(|\frac{z_I}{\mu}| \to \infty \vee |\frac{z_I}{\mu}| < \infty\right), \\
-\mathbf{1}, \text{ if } \frac{z_R}{\mu}, \frac{z_I}{\mu} \to -\infty
\end{cases}
\end{aligned} \tag{203}$$

where the vector $\alpha \in [-1,1]^2$ since $\|\varphi'_\mu(\|[\mathbf{a}_k^H\mathbf{x}]\|_2)\|_2 \leq 1$ from (198), $\vee/\wedge$ are the or/and logic operations respectively, $\mathbf{1} = [1,1]^T$, and $[\mathbf{a}_k^H\mathbf{x}] = [z_R, z_I]^T$. Further, given that $\varphi'_\mu(|\mathbf{a}_k^H\mathbf{z}|)$ and $\|\cdot\|_2$ are Lipschitz continuous function on any $S_\mu(\mathbf{w})$ according to Theorem 5.2.2 and (194), respectively, then the composed function $\varphi'_\mu(\|[\mathbf{a}_k^H\mathbf{x}]\|_2)$ is Lipschitz continuous as it is shown in Chapter 12 in Eriksson et al. (2013). Hence, considering that $\varphi'_\mu(\|[\mathbf{a}_k^H\mathbf{x}]\|_2)$ is Lipschitz continuous, from Theorem 3.9 in Clarke (1990) and (201) and (203), it can be concluded that

$$\left\{ \lim_{\substack{\mathbf{x}\to\mathbf{x}^* \\ \mu\downarrow 0}} \frac{[\mathbf{a}_k^H\mathbf{x}]}{\sqrt{\|[\mathbf{a}_k^H\mathbf{x}]\|_2^2 + \mu^2}} \right\} = [-1,1]^2. \tag{204}$$

In order to prove the gradient consistency property of the function $g(\mathbf{x},\mu)$, then consider the Wirtinger derivative of $g(\mathbf{x},\mu)$. From (204) we have that $\varphi_\mu(|\mathbf{a}_k^H\mathbf{x}|)$ satisfies the gradient consistency property, then one can write

$$\begin{aligned}
&= \left\{ \lim_{\substack{\mathbf{x}\to\mathbf{x}^* \\ \mu\downarrow 0}} \partial g(\mathbf{x},\mu) \right\} = \left\{ \lim_{\substack{\mathbf{x}\to\mathbf{x}^* \\ \mu\downarrow 0}} \frac{2}{m} \sum_{k=1}^{m} (\varphi_\mu(|\mathbf{a}_k^H\mathbf{x}|) - q_k) \partial \varphi_\mu(|\mathbf{a}_k^H\mathbf{x}|) \right\} \\
&\equiv \left\{ \lim_{\substack{\mathbf{x}\to\mathbf{x}^* \\ \mu\downarrow 0}} \frac{2}{m} \sum_{k=1}^{m} (\varphi_\mu(|\mathbf{a}_k^H\mathbf{x}|) - q_k) \begin{bmatrix} \mathbf{a}_{k,R} & -\mathbf{a}_{k,I} \\ \mathbf{a}_{k,I} & \mathbf{a}_{k,R} \end{bmatrix} \frac{[\mathbf{a}_k^H\mathbf{x}]}{\sqrt{\|[\mathbf{a}_k^H\mathbf{x}]\|_2^2 + \mu^2}} \right\} \\
&= \frac{2}{m} \sum_{k=1}^{m} (\varphi_0(|\mathbf{a}_k^H\mathbf{x}^*|) - q_k) \partial^c F([\mathbf{a}_k^H\mathbf{x}^*]) \\
&\equiv \partial^c f(\mathbf{x}^*),
\end{aligned} \tag{205}$$

where the third equality comes from (195). This shows that the gradient consistency property holds for the smoothing function $g(\mathbf{x},\mu)$.

On the other hand, considering the result in Theorem 5.2.2 we have that

$$\liminf_{i\to\infty} \|\partial g(\mathbf{x}_i, \mu_{i-1})\|_2 = 0$$

for the sequences $\{\mu_i\}$ and $\{\mathbf{x}_i\}$ generated by Algorithm 7. This means that there exist $\mathbf{x}^*$ such that $\lim_{i\to\infty} \mathbf{x}_i = \mathbf{x}^*$. Given the fact that $g(\mathbf{x}, \mu)$ satisfies the gradient consistency property according to (205), then one can conclude that $\mathbf{0} \in \partial^c f(\mathbf{x}^*)$. □

Appendix H. Proof of Theorem 7.2.1

Let us define the search set as

$$\mathscr{J} := \{\mathbf{z} \in \mathbb{C}^N, B\text{-bandlimited} : \text{dist}(\mathbf{x}, \mathbf{z}) \leq \rho, B \leq N/2\}, \tag{206}$$

for some small constant $\rho > 0$. Recall that $\mathbf{z}$ is a B-bandlimited signal if there exists k such that $\tilde{\mathbf{z}}[k] = \cdots = \tilde{\mathbf{z}}[N+k+B-1] = 0$, where $\tilde{\mathbf{z}}$ is the Fourier transform of $\mathbf{z}$. The bandlimit condition guarantees that we have unique solution, according to Proposition 4.

In order to prove Theorem 7.2.1, the function $h(\mathbf{z}, \mu)$ in (87) needs to satisfy the four requirements stated in the following lemma, which are used in the analysis of convergence for stochastic gradient methods Ghadimi and Lan (2013).

Lemma 10.0.5. The function $h(\mathbf{z}, \mu)$ in (87) and its Wirtinger derivative in (95) satisfy the following properties.

1. The cost function $h(\mathbf{z}, \mu)$ in (87) is bounded below.

2. The set $\mathscr{J}$ as defined in (206) is closed and bounded.

3. There exists a constant $U > 0$, such that

$$\left\| \frac{\partial h(\mathbf{z}_1, \mu)}{\partial \bar{\mathbf{z}}} - \frac{\partial h(\mathbf{z}_2, \mu)}{\partial \bar{\mathbf{z}}} \right\|_2 \leq U \left\| \mathbf{z}_1 - \mathbf{z}_2 \right\|_2, \tag{207}$$

holds for all $\mathbf{z}_1, \mathbf{z}_2 \in \mathscr{J}$.

4. For all $\mathbf{z} \in \mathscr{J}$

$$\mathbb{E}_{\Gamma_{(t)}}\left[\left\|\mathbf{d}_{\Gamma_{(t)}} - \frac{\partial h(\mathbf{z},\mu)}{\partial \overline{\mathbf{z}}}\right\|_2^2\right] \leq \zeta^2, \tag{208}$$

for some $\zeta > 0$, where $\mathbf{d}_{\Gamma_{(t)}}$ is as in Line 9 of Algorithm 7.

Demostración. See Appendix 8. □

To prove Theorem 7.2.1, denote the set $\mathscr{K}_1 := \{t | \mu^{(t+1)} = \gamma_1 \mu^{(t)}\}$ with $\gamma_1 \in (0,1)$, which is a tunable parameter Zhang and Chen (2009). If the set $\mathscr{K}_1$ is finite, then according to Lines 13-16 in Algorithm 7 there exists an integer $\grave{t}$, such that, for all $t > \grave{t}$

$$\left\|\mathbf{d}_{\Gamma_{(t)}}\right\|_2 \geq \gamma \mu^{(\grave{t})}, \tag{209}$$

with $\gamma \in (0,1)$. Taking $\grave{\mu} = \mu^{(\grave{t})}$, the optimization problem (87) reduces to

$$\min_{\mathbf{z} \in \mathbb{C}^N} h(\mathbf{z}, \grave{\mu}). \tag{210}$$

Now, considering the properties stated in Lemma 10.0.5, from (Ghadimi and Lan, 2013, Theorem 2.1) we get

$$\lim_{t\to\infty}\left\|\frac{\partial h(\mathbf{x}^{(t)}, \mu^{(t)})}{\partial \overline{\mathbf{z}}}\right\|_2 = \lim_{t\to\infty}\left\|\mathbb{E}_{\Gamma_{(t)}}\left[\mathbf{d}_{\Gamma_{(t)}}\right]\right\|_2 = 0. \tag{211}$$

It can be readily seen that (211) contradicts the assumption $\left\|\mathbf{d}_{\Gamma_{(t)}}\right\|_2 \geq \gamma \mu^{(\grave{t})}$, for all $t > \grave{t}$. This

shows that $\mathscr{K}_1$ must be infinite and $\lim_{t\to\infty} \mu^{(t)} = 0$.

Given that $\mathscr{K}_1$ is infinite, we deduce that

$$\begin{aligned}\lim_{t\to\infty} \left\| \frac{\partial h(\mathbf{x}^{(t)}, \mu^{(t)})}{\partial \bar{\mathbf{z}}} \right\|_2 &= \lim_{t\to\infty} \left\| \mathbb{E}_{\Gamma_{(t)}} \left[\mathbf{d}_{\Gamma_{(t)}} \right] \right\|_2 \\ &\leq \lim_{t\to\infty} \mathbb{E}_{\Gamma_{(t)}} \left[\left\| \mathbf{d}_{\Gamma_{(t)}} \right\|_2 \right] \leq \gamma \lim_{t\to\infty} \mu^{(t)} = 0, \end{aligned} \tag{212}$$

where the second line follows from the Jensen inequality. Therefore, from (212) the result of Theorem 7.2.1 holds.

Proof of Lemma 10.0.5. The proof of Lemma 10.0.5 is obtained by individually proving the following four requirements.

1) Following from the definition of $h(\mathbf{z}, \mu)$ in (87) it is clear that $h(\mathbf{z}, \mu) \geq 0$ and thus bounded below.

2) This holds by definition.

3) From (93) it follows that the ℓ-th entry of $\frac{\partial h(\mathbf{z},\mu)}{\partial \bar{\mathbf{z}}}$ is given by

$$\frac{\partial h(\mathbf{z}, \mu)}{\partial \bar{\mathbf{z}}}[\ell] = \frac{1}{N^2} \sum_{k,p=0}^{N-1} \left(\mathbf{f}_k^H \mathbf{g}_p(\mathbf{z}) - \upsilon_{k,p} \right) \overline{q}_{\ell,p} e^{2\pi i \ell k/N}, \tag{213}$$

where $\upsilon_{k,p} = \sqrt{\mathbf{Z}[p,k]}\frac{\mathbf{f}_k^H \mathbf{g}_p(\mathbf{z})}{\varphi_\mu\left(\left|\mathbf{f}_k^H \mathbf{g}_p(\mathbf{z})\right|\right)}$, and

$$\overline{q}_{\ell,p} = \overline{\mathbf{z}}[\ell+p] + \overline{\mathbf{z}}[\ell-p]e^{-2\pi ikp/N},$$

$$\mathbf{g}_p(\mathbf{z}) = [\mathbf{z}[0]\mathbf{z}[pL], \cdots, \mathbf{z}[N-1]\mathbf{z}[N-1+pL]]^T.$$

Let $\mathbf{D}_p(\mathbf{z})$ be a diagonal matrix composed of the entries of $\overline{\mathbf{z}}_{pL}[n] = \overline{\mathbf{z}}[n+pL]$. Using (92), the term $\overline{q}_{\ell,p}e^{2\pi i\ell k/N}$ can be rewritten as

$$\overline{q}_{\ell,p}e^{2\pi i\ell k/N} = \left(\mathbf{D}_p(\mathbf{z})\mathbf{f}_k\right)[\ell] + \omega^{-kp}\left(\mathbf{D}_{-p}(\mathbf{z})\mathbf{f}_k\right)[\ell]. \tag{214}$$

Thus,

$$\frac{\partial h(\mathbf{z},\mu)}{\partial \overline{\mathbf{z}}} = \frac{1}{N^2}\sum_{p,k=0}^{N-1} f_{k,p}(\mathbf{z}) + g_{k,p}(\mathbf{z}), \tag{215}$$

where

$$f_{k,p}(\mathbf{z}) = \rho_{k,p}(\mathbf{z})\mathbf{D}_p(\mathbf{z})\mathbf{f}_k,$$

$$g_{k,p}(\mathbf{z}) = \omega^{-kp}\rho_{k,p}(\mathbf{z})\mathbf{D}_{-p}(\mathbf{z})\mathbf{f}_k, \tag{216}$$

and

$$\rho_{k,p}(\mathbf{z}) = \mathbf{f}_k^H \mathbf{g}_p(\mathbf{z}) - \sqrt{\mathbf{Z}[p,k]} \frac{\mathbf{f}_k^H \mathbf{g}_p(\mathbf{z})}{\varphi_\mu\left(\left|\mathbf{f}_k^H \mathbf{g}_p(\mathbf{z})\right|\right)}. \tag{217}$$

To prove 3) we establish that any $f_{k,p}(\mathbf{z})$ and $g_{k,p}(\mathbf{z})$ satisfy

$$\left\| f_{k,p}(\mathbf{z}_1) - f_{k,p}(\mathbf{z}_2) \right\|_2 \leq r_{k,p} \|\mathbf{z}_1 - \mathbf{z}_2\|_2, \tag{218}$$

and

$$\left\| g_{k,p}(\mathbf{z}_1) - g_{k,p}(\mathbf{z}_2) \right\|_2 \leq s_{k,p} \|\mathbf{z}_1 - \mathbf{z}_2\|_2, \tag{219}$$

for all $\mathbf{z}_1, \mathbf{z}_2 \in \mathscr{J}$ and some constants $r_{k,p}, s_{k,p} > 0$. In fact, once we prove (218), it can be performed a similar analysis for $g_{k,p}(\mathbf{z})$, and thus the result of this third part holds.

From the definition of $f_{k,p}(\mathbf{z})$, for any $\mathbf{z}_1, \mathbf{z}_2 \in \mathscr{J}$ we have that

$$\frac{1}{\sqrt{N}} \left\| f_{k,p}(\mathbf{z}_1) - f_{k,p}(\mathbf{z}_2) \right\|_2 \leq \left\| \rho_{k,p}(\mathbf{z}_1)\overline{\mathbf{z}}_1 - \rho_{k,p}(\mathbf{z}_2)\overline{\mathbf{z}}_2 \right\|_2, \tag{220}$$

considering that $\mathbf{D}_p(\mathbf{z}_1)$ and $\mathbf{D}_p(\mathbf{z}_2)$ are diagonal matrices, and $\|\mathbf{f}_k\|_2 = \sqrt{N}$. Observe that from

(217) and (220) it can be obtained that

$$\begin{aligned}
&\frac{1}{\sqrt{N}}\left\|f_{k,p}(\mathbf{z}_1)-f_{k,p}(\mathbf{z}_2)\right\|_2 \\
\leq&\frac{\left|\mathbf{f}_k^H\mathbf{g}_p(\mathbf{z}_1)\right|}{\mu}\left(\varphi_\mu\left(\left|\mathbf{f}_k^H\mathbf{g}_p(\mathbf{z}_1)\right|\right)+\sqrt{\mathbf{Z}[p,k]}\right)\|\mathbf{z}_1-\mathbf{z}_2\|_2 \\
&+\|\mathbf{z}_2\|_2\underbrace{\left|\rho_{k,p}(\mathbf{z}_1)-\rho_{k,p}(\mathbf{z}_2)\right|}_{p_1},
\end{aligned} \tag{221}$$

where the second inequality comes from the fact that $\varphi_\mu\left(\left|\mathbf{f}_k^H\mathbf{g}_p(\mathbf{z}_1)\right|\right)\geq\mu$. The term p_1 in (221) can be upper bounded as

$$\begin{aligned}
p_1&\leq\left|\mathbf{f}_k^H\mathbf{g}_p(\mathbf{z}_1)-\mathbf{f}_k^H\mathbf{g}_p(\mathbf{z}_2)\right|+\sqrt{\mathbf{Z}[p,k]}\left|\frac{\mathbf{f}_k^H\mathbf{g}_p(\mathbf{z}_1)}{\varphi_\mu\left(\left|\mathbf{f}_k^H\mathbf{g}_p(\mathbf{z}_1)\right|\right)}-\frac{\mathbf{f}_k^H\mathbf{g}_p(\mathbf{z}_2)}{\varphi_\mu\left(\left|\mathbf{f}_k^H\mathbf{g}_p(\mathbf{z}_2)\right|\right)}\right| \\
&\leq\left|\mathbf{f}_k^H\mathbf{g}_p(\mathbf{z}_1)-\mathbf{f}_k^H\mathbf{g}_p(\mathbf{z}_2)\right|+\frac{\sqrt{\mathbf{Z}[p,k]}}{\mu^2}\varphi_\mu\left(\left|\mathbf{f}_k^H\mathbf{g}_p(\mathbf{z}_2)\right|\right)\left|\mathbf{f}_k^H\mathbf{g}_p(\mathbf{z}_1)-\mathbf{f}_k^H\mathbf{g}_p(\mathbf{z}_2)\right| \\
&+\frac{\sqrt{\mathbf{Z}[p,k]}}{\mu^2}\left|\mathbf{f}_k^H\mathbf{g}_p(\mathbf{z}_2)\right|\left|\varphi_\mu\left(\left|\mathbf{f}_k^H\mathbf{g}_p(\mathbf{z}_1)\right|\right)-\varphi_\mu\left(\left|\mathbf{f}_k^H\mathbf{g}_p(\mathbf{z}_2)\right|\right)\right|.
\end{aligned} \tag{222}$$

Recall that $\mathscr{J}$ is a closed bounded set, and thus compact. Since $\varphi_\mu(\cdot)$ is a continuous function, there exists a constant M_{φ_μ} such that $\varphi_\mu\left(\left|\mathbf{f}_k^H\mathbf{g}_p(\mathbf{z})\right|\right)\leq M_{\varphi_\mu}$ for all $\mathbf{z}\in\mathscr{J}$. Also, from Lemma 2 in Pinilla et al. (2018a) we have that $\varphi_\mu(\cdot)$ is a 1-Lipschitz function. Combining this with (222) we get

$$\begin{aligned}
p_1&\leq\left|\mathbf{f}_k^H\mathbf{g}_p(\mathbf{z}_1)-\mathbf{f}_k^H\mathbf{g}_p(\mathbf{z}_2)\right|+\frac{\sqrt{\mathbf{Z}[p,k]}M_{\varphi_\mu}}{\mu^2}\left|\mathbf{f}_k^H\mathbf{g}_p(\mathbf{z}_1)-\mathbf{f}_k^H\mathbf{g}_p(\mathbf{z}_2)\right| \\
&+\frac{\sqrt{\mathbf{Z}[p,k]}}{\mu^2}\left|\mathbf{f}_k^H\mathbf{g}_p(\mathbf{z}_2)\right|\left|\left|\mathbf{f}_k^H\mathbf{g}_p(\mathbf{z}_1)\right|-\left|\mathbf{f}_k^H\mathbf{g}_p(\mathbf{z}_2)\right|\right|,
\end{aligned} \tag{223}$$

and thus

$$p_1 \leq \left(\frac{\sqrt{\mathbf{Z}[p,k]} M_{\varphi_\mu}}{\mu^2} + 1 \right) \left| \mathbf{f}_k^H \mathbf{g}_p(\mathbf{z}_1) - \mathbf{f}_k^H \mathbf{g}_p(\mathbf{z}_2) \right| + \frac{\sqrt{\mathbf{Z}[p,k]}}{\mu^2} \left| \mathbf{f}_k^H \mathbf{g}_p(\mathbf{z}_2) \right| \left| \mathbf{f}_k^H \mathbf{g}_p(\mathbf{z}_1) - \mathbf{f}_k^H \mathbf{g}_p(\mathbf{z}_2) \right|, \tag{224}$$

where (224) results from applying the triangular inequality. Putting together (221) and (224) we obtain that

$$\begin{aligned} \frac{1}{\sqrt{N}} \left\| f_{k,p}(\mathbf{z}_1) - f_{k,p}(\mathbf{z}_2) \right\|_2 \leq{}& \frac{\left| \mathbf{f}_k^H \mathbf{g}_p(\mathbf{z}_1) \right|}{\mu} \left(M_{\varphi_\mu} + \sqrt{\mathbf{Z}[p,k]} \right) \|\mathbf{z}_1 - \mathbf{z}_2\|_2 \\ &+ \|\mathbf{z}_2\|_2 \left(\frac{\sqrt{\mathbf{Z}[p,k]} M_{\varphi_\mu}}{\mu^2} + 1 \right) \left| \mathbf{f}_k^H \mathbf{g}_p(\mathbf{z}_1) - \mathbf{f}_k^H \mathbf{g}_p(\mathbf{z}_2) \right| \\ &+ \frac{\|\mathbf{z}_2\|_2 \sqrt{\mathbf{Z}[p,k]}}{\mu^2} \left| \mathbf{f}_k^H \mathbf{g}_p(\mathbf{z}_2) \right| \left| \mathbf{f}_k^H \mathbf{g}_p(\mathbf{z}_1) - \mathbf{f}_k^H \mathbf{g}_p(\mathbf{z}_2) \right|. \end{aligned} \tag{225}$$

Observe that the upper bound in (225) directly depends on a term of the form $\mathbf{f}_k^H \mathbf{g}_p(\mathbf{z})$ for some $\mathbf{z} \in \mathscr{J}$, which might be zero. However, Lemma 10.0.6 proves that $\left| \mathbf{f}_k^H \mathbf{g}_p(\mathbf{z}) \right| > 0$ or equivalently $\mathbf{f}_k^H \mathbf{g}_p(\mathbf{z}) \neq 0$, for almost all $\mathbf{z} \in \mathscr{J}$.

Lemma 10.0.6. Let $\mathbf{z} \in \mathscr{J}$ where $\mathscr{J}$ as defined in (206). Then, for almost all $\mathbf{z} \in \mathscr{J}$ the following holds

$$\left| \mathbf{f}_k^H \mathbf{g}_p(\mathbf{z}) \right| > 0, \tag{226}$$

for all $k, p \in \{0, \cdots, N-1\}$, with $\mathbf{g}_p(\mathbf{z})$ as in (214).

Demostración. We prove this lemma by contradiction. Suppose that $\left|\mathbf{f}_k^H \mathbf{g}_p(\mathbf{z})\right| = 0$. Then, from (60) we have that

$$\begin{aligned}\left|\mathbf{f}_k^H \mathbf{g}_p(\mathbf{z})\right|^2 &= \left|\sum_{n=0}^{N-1} \mathbf{z}[n]\mathbf{z}[n+pL]e^{-2\pi ink/N}\right|^2 \\ &= \sum_{n,m=0}^{N-1} \left(\mathbf{z}[n]\overline{\mathbf{z}}[m]\mathbf{z}[n+pL]\overline{\mathbf{z}}[m+pL]\right) e^{\frac{2\pi i(m-n)k}{N}} = 0. \end{aligned} \tag{227}$$

Observe that (227) is a quartic polynomial equation with respect to the entries of $\mathbf{z}$. However, for almost all signals $\mathbf{z} \in \mathscr{J}$ the left hand side of (227) will not be equal to zero which leads to a contradiction Bendory et al. (2018a). □

Then, proceeding to bound the term $\left|\mathbf{f}_k^H \mathbf{g}_p(\mathbf{z})\right|$, notice that from (60) we have that

$$\begin{aligned}\left|\mathbf{f}_k^H \mathbf{g}_p(\mathbf{z})\right| &= \left|\sum_{n=0}^{N-1} \mathbf{z}[n]\mathbf{z}[n+pL]e^{-2\pi ink/N}\right| \\ &\leq \sum_{n=0}^{N-1} |\mathbf{z}[n]\mathbf{z}[n+pL]| \leq N\|\mathbf{z}\|_2, \end{aligned} \tag{228}$$

in which the second inequality arises from $\|\mathbf{z}\|_2 \leq \sqrt{N}\|\mathbf{z}\|_\infty$ and $\|\mathbf{z}\|_1 \leq \sqrt{N}\|\mathbf{z}\|_2$. Combining (225) and (228) we get

$$\begin{aligned}\frac{1}{\sqrt{N}}\left\|f_{k,p}(\mathbf{z}_1) - f_{k,p}(\mathbf{z}_2)\right\|_2 \leq{}& \frac{N\|\mathbf{z}_1\|_2}{\mu}\left(M_{\varphi_\mu} + \sqrt{\mathbf{Z}[p,k]}\right)\|\mathbf{z}_1 - \mathbf{z}_2\|_2 \\ &+ \|\mathbf{z}_2\|_2\left(\frac{\sqrt{\mathbf{Z}[p,k]}M_{\varphi_\mu}}{\mu^2} + 1\right)\left|\mathbf{f}_k^H \mathbf{g}_p(\mathbf{z}_1) - \mathbf{f}_k^H \mathbf{g}_p(\mathbf{z}_2)\right| \\ &+ \frac{N\|\mathbf{z}_2\|_2^2\sqrt{\mathbf{Z}[p,k]}}{\mu^2}\left|\mathbf{f}_k^H \mathbf{g}_p(\mathbf{z}_1) - \mathbf{f}_k^H \mathbf{g}_p(\mathbf{z}_2)\right|. \end{aligned} \tag{229}$$

Now, we have to analyze the term $\left|\mathbf{f}_k^H \mathbf{g}_p(\mathbf{z}_1) - \mathbf{f}_k^H \mathbf{g}_p(\mathbf{z}_2)\right|$ in (229). Specifically, from (60) it can be obtained that

$$\begin{aligned}\left|\mathbf{f}_k^H \mathbf{g}_p(\mathbf{z}_1) - \mathbf{f}_k^H \mathbf{g}_p(\mathbf{z}_2)\right| &\leq \sum_{n=0}^{N-1} \left|\mathbf{z}_1[n]\mathbf{z}_1[n+pL] - \mathbf{z}_2[n]\mathbf{z}_2[n+pL]\right| \\ &\leq N\left(\|\mathbf{z}_1\|_2 + \|\mathbf{z}_2\|_2\right)\|\mathbf{z}_1 - \mathbf{z}_2\|_2, \end{aligned} \tag{230}$$

where the second inequality results from $\|\mathbf{z}\|_2 \leq \sqrt{N}\|\mathbf{z}\|_\infty$ and $\|\mathbf{z}\|_1 \leq \sqrt{N}\|\mathbf{z}\|_2$. Combining (229) and (230) we obtain that

$$\left\|f_{k,p}(\mathbf{z}_1) - f_{k,p}(\mathbf{z}_2)\right\|_2 \leq r_{k,p}\|\mathbf{z}_1 - \mathbf{z}_2\|_2, \tag{231}$$

where $r_{k,p}$ is given by

$$\begin{aligned} r_{k,p} = &\frac{N\sqrt{N}\|\mathbf{z}_1\|_2}{\mu}\left(M_{\varphi_\mu} + \sqrt{\mathbf{Z}[p,k]}\right) + N^2\sqrt{N}\left(\|\mathbf{z}_1\|_2 + \|\mathbf{z}_2\|_2\right)\frac{\|\mathbf{z}_2\|_2^2\sqrt{\mathbf{Z}[p,k]}}{\mu^2} \\ &+ N\sqrt{N}\left(\|\mathbf{z}_1\|_2 + \|\mathbf{z}_2\|_2\right)\|\mathbf{z}_2\|_2\left(\frac{\sqrt{\mathbf{Z}[p,k]}M_{\varphi_\mu}}{\mu^2} + 1\right). \end{aligned} \tag{232}$$

Since the set $\mathscr{J}$ is bounded, then $\|\mathbf{z}\|_2 < \infty$ for all $\mathbf{z} \in \mathscr{J}$. Therefore, $0 < r_{k,p} < \infty$, and from (231) the result holds.

4) We proceed to prove (208). Observe that

$$\mathbb{E}_{\Gamma_{(t)}}\left[\left\|\mathbf{d}_{\Gamma_{(t)}}-\frac{\partial h(\mathbf{z},\mu)}{\partial \overline{\mathbf{z}}}\right\|_2^2\right]\leq \mathbb{E}_{\Gamma_{(t)}}\left[2\left\|\mathbf{d}_{\Gamma_{(t)}}\right\|_2^2\right]+2\left\|\frac{\partial h(\mathbf{z},\mu)}{\partial \overline{\mathbf{z}}}\right\|_2^2, \tag{233}$$

in which the inequality comes from the fact that $\|\mathbf{w}_1+\mathbf{w}_2\|_2^2\leq 2\left(\|\mathbf{w}_1\|_2^2+\|\mathbf{w}_2\|_2^2\right)$ for any $\mathbf{w}_1,\mathbf{w}_2\in\mathbb{C}^N$. Combining (207) and (233) we have that

$$\mathbb{E}_{\Gamma_{(t)}}\left[\left\|\mathbf{d}_{\Gamma_{(t)}}-\frac{\partial h(\mathbf{z},\mu)}{\partial \overline{\mathbf{z}}}\right\|_2^2\right]\leq \mathbb{E}_{\Gamma_{(t)}}\left[2\left\|\mathbf{d}_{\Gamma_{(t)}}\right\|_2^2\right]+2U\|\mathbf{z}\|_2^2, \tag{234}$$

for some $U>0$. Recall that $\Gamma_{(t)}$ is sampled uniformly at random from all subsets of $\{1,\cdots,N\}\times\{1\cdots,R\}$ with cardinality Q. From the definition of $\mathbf{d}_{\Gamma_{(t)}}$ in Line 9 of Algorithm 7, it can be concluded that

$$\begin{aligned}\mathbb{E}_{\Gamma_{(t)}}\left[2\left\|\mathbf{d}_{\Gamma_{(t)}}\right\|_2^2\right]&\leq\frac{4Q}{N^2}\sum_{p,k=0}^{N-1}\left\|f_{k,p}(\mathbf{z})+g_{k,p}(\mathbf{z})\right\|_2^2\\&\leq\frac{8Q}{N^2}\sum_{p,k=0}^{N-1}\left\|f_{k,p}(\mathbf{z})\right\|_2^2+\left\|g_{k,p}(\mathbf{z})\right\|_2^2,\end{aligned} \tag{235}$$

using the fact that $\|\mathbf{w}_1+\mathbf{w}_2\|_2^2\leq 2\left(\|\mathbf{w}_1\|_2^2+\|\mathbf{w}_2\|_2^2\right)$ for any $\mathbf{w}_1,\mathbf{w}_2\in\mathbb{C}^N$. Furthermore, since $f_{k,p}(\mathbf{z})$ and $g_{k,p}(\mathbf{z})$ satisfy (218) and (219), respectively, we conclude that

$$\mathbb{E}_{\Gamma_{(t)}}\left[2\left\|\mathbf{d}_{\Gamma_{(t)}}\right\|_2^2\right]\leq\frac{8Q\|\mathbf{z}\|_2^2}{N^2}\sum_{p,k=0}^{N-1}r_{k,p}^2+s_{k,p}^2, \tag{236}$$

for some constants $r_{k,p}, s_{k,p} > 0$. Thus, combining (234) and (236) we obtain that

$$\mathbb{E}_{\Gamma_{(t)}}\left[\left\|\mathbf{d}_{\Gamma_{(t)}} - \frac{\partial h(\mathbf{z},\boldsymbol{\mu})}{\partial \overline{\mathbf{z}}}\right\|_2^2\right] \leq \zeta^2, \tag{237}$$

where ζ is defined as

$$\zeta = \|\mathbf{z}\|_2 \sqrt{\frac{8Q}{N^2}\sum_{p,k=0}^{N-1} r_{k,p}^2 + s_{k,p}^2 + 2U}. \tag{238}$$

Notice $\zeta < \infty$ because the set $\mathscr{J}$ is bounded. Thus, from (237) the result holds.

Appendix I. Proof of Proposition 4

We begin the proof by reformulating the measurement model to a more convenient structure. Applying the inverse Fourier transform we write $\mathbf{x}[n] = \frac{1}{N}\sum_{k=0}^{N-1} \tilde{\mathbf{x}}[k] e^{2\pi i k n/N}$. Then, according to (85), we have

$$\mathbf{A}[p,k] = |\mathbf{S}[p,k]|^2, \tag{239}$$

where $\mathbf{S}[p,k]$ is defined as

$$\begin{aligned}
\mathbf{S}[p,k] &= \sum_{n=0}^{N-1} \mathbf{x}[n]\overline{\mathbf{x}[n-p]} e^{-2i\pi nk/N} \\
&= \frac{1}{N^2}\sum_{n=0}^{N-1}\left(\sum_{\ell_1=0}^{N-1}\tilde{\mathbf{x}}[\ell_1]e^{2\pi i\ell_1 n/N}\right)\times\left(\sum_{\ell_2=0}^{N-1}\overline{\tilde{\mathbf{x}}[\ell_2]}e^{-2\pi i\ell_2 n/N}e^{2\pi i\ell_2 p/N}\right)e^{-2\pi ikn/N} \\
&= \frac{1}{N^2}\sum_{\ell_1,\ell_2=0}^{N-1}\tilde{\mathbf{x}}[\ell_1]\overline{\tilde{\mathbf{x}}[\ell_2]}e^{2\pi i\ell_2 p/N}\sum_{n=0}^{N-1}e^{2\pi in(\ell_1-\ell_2-k)/N} = \frac{1}{N}\sum_{\ell=0}^{N-1}\tilde{\mathbf{x}}[\ell+k]\overline{\tilde{\mathbf{x}}[\ell]}e^{2\pi i\ell p/N},
\end{aligned} \tag{240}$$

since the later sum is equal to N if $\ell_1 = k + \ell_2$ and zero otherwise.

Assume that $B = N/2$, N is even, that $\tilde{\mathbf{x}}[n] \neq 0$ for $k = 0 \ldots, B-1$, and that $\tilde{\mathbf{x}}[n] = 0$ for $k = N/2, \ldots, N-1$. If the signal's nonzero Fourier coefficients are not in the interval $0, \ldots, N/2-1$, then we can cyclically reindex the signal without affecting the proof. If N is odd, then one should replace $N/2$ by $\lfloor N/2 \rfloor$ everywhere in the sequel. Clearly, the proof carries through for any $B \leq N/2$.

Considering (240), the band-limit assumption on the signal forms a "inverted pyramid"structure. Here, each row represents fixed k and varying ℓ of $\tilde{\mathbf{x}}[\ell + k]\overline{\tilde{\mathbf{x}}[\ell]}$ for $k = 0, \ldots, N/2-1$

$$
\begin{aligned}
&|\tilde{\mathbf{x}}[0]|^2, |\tilde{\mathbf{x}}[1]|^2, \ldots, |\tilde{\mathbf{x}}[B-1]|^2, 0, \ldots, 0 \\
&\overline{\tilde{\mathbf{x}}[0]}\tilde{\mathbf{x}}[1], \overline{\tilde{\mathbf{x}}[1]}\tilde{\mathbf{x}}[2], \ldots, \overline{\tilde{\mathbf{x}}[B-2]}\tilde{\mathbf{x}}[B-1], 0, \ldots, 0 \\
&\qquad\qquad \vdots \\
&\overline{\tilde{\mathbf{x}}[0]}\tilde{\mathbf{x}}[B-1], 0, \ldots, 0, 0, \ldots, 0 \\
&0, 0, \ldots, \tilde{\mathbf{x}}[0]\overline{\tilde{\mathbf{x}}[B-1]}, 0, \ldots, 0 \\
&\qquad\qquad \vdots \\
&0, \mathbf{x}[0]\overline{\tilde{\mathbf{x}}[1]}, \tilde{\mathbf{x}}[1]\overline{\tilde{\mathbf{x}}[2]}, \ldots, \tilde{\mathbf{x}}[B-2]\overline{\tilde{\mathbf{x}}[B-1]}, 0, \ldots, 0.
\end{aligned} \tag{241}
$$

Then, $\mathbf{S}[p,k]$ as in (240) is a subsample of the Fourier transform of each one of the pyramid's rows.

Step 0: From the $(B-1)$-th row of (241), we see that

$$
|\mathbf{S}[p,B]| = |\tilde{\mathbf{x}}[0]||\tilde{\mathbf{x}}[B-1]|, \forall p = 0, \ldots, N-1. \tag{242}
$$

Considering that in the radar phase retrieval problem the translation ambiguity is continuous (second ambiguity in Proposition 3), we can set $\tilde{\mathbf{x}}[0]$ to be real and, without loss of generality it can be assumed that $\tilde{\mathbf{x}}[0] = 1$ Bendory et al. (2018a). Note that in contrast to the FROG phase retrieval problem, this continuity property in the radar scenario is satisfied for general signals. Then, from (242) we obtain that

$$|\mathbf{S}[p, B-1]| = |\tilde{\mathbf{x}}[B-1]|, \forall p = 0, \ldots, N-1. \tag{243}$$

Step 1: From the first row of (241), we conclude the following system of equations

$$|\mathbf{S}[p,0]| = \frac{1}{N}\left|\sum_{\ell=0}^{B-1} |\tilde{\mathbf{x}}[\ell]|^2 e^{2\pi i \ell p/N}\right|, p = 0, \ldots, N-1. \tag{244}$$

Given the fact that from the **Step 0** the entries $|\tilde{\mathbf{x}}[0]|$, $|\tilde{\mathbf{x}}[B-1]|$, and $\{|\mathbf{S}[p,0]|\}_{p=0}^{N-1}$ are known, then appealing to Lemma 10.0.7 for almost all signals we have that $|\tilde{\mathbf{x}}[1]|, \ldots, |\tilde{\mathbf{x}}[B-2]|$ are uniquely determined. It is worth mentioning that this previous argument does not imply that $\tilde{\mathbf{x}}[1], \ldots, \tilde{\mathbf{x}}[B-1]$ are uniquely determined. In fact there are up to 2^{B-1} vectors, modulo global phase, reflection and conjugation that satisfy the constraints in (242), and (244) (Bendory et al., 2017a, Section 3.1).

Step 2: Moving to analyze the second row of (241) we obtain the following system of

equations

$$|\mathbf{S}[p,1]| = \frac{1}{N}\left|\sum_{\ell=0}^{B-2}\tilde{\mathbf{x}}[\ell+1]\overline{\tilde{\mathbf{x}}[\ell]}e^{2\pi i\ell p/N}\right|, p = 0,\ldots,N-1. \tag{245}$$

Fix one of the possible solutions for $\tilde{\mathbf{x}}[1]$ from **Step 1**. Then, since $\tilde{\mathbf{x}}[0]$ is known, Lemma 10.0.7 states that for almost all signals $\overline{\tilde{\mathbf{x}}[1]}\tilde{\mathbf{x}}[2],\ldots,\tilde{\mathbf{x}}[B-1]\overline{\tilde{\mathbf{x}}[B-2]}$ are uniquely determined.

Step 3: Considering the fact that $\tilde{\mathbf{x}}[0]$, and $\tilde{\mathbf{x}}[1]$ are known, from **Step 2** we can estimate $\tilde{\mathbf{x}}[2]$. Thus, since $\overline{\tilde{\mathbf{x}}[0]}\tilde{\mathbf{x}}[2]$ is known, appealing to Lemma 10.0.7 for almost all signals $\overline{\tilde{\mathbf{x}}[1]}\tilde{\mathbf{x}}[3],\ldots,\overline{\tilde{\mathbf{x}}[B-3]}\tilde{\mathbf{x}}[B-1]$ are uniquely determined. However, remark that at this stage the 2^{B-1} possible solutions from **Step 2** remains.

Despite the large amount of possible solutions, we can prove that at this step there is only one vector (up to trivial ambiguities) out of the 2^{B-1} possibilities of **Step 2**, that is consistent with the constraints in (242), (244), and (245). To see this, from **Step 1** we have that $|\tilde{\mathbf{x}}[0]|,\ldots,|\tilde{\mathbf{x}}[B-1]|$ are uniquely determined. Therefore, from the knowledge of $\{|\tilde{\mathbf{x}}[\ell]|\}_{\ell=0}^{B-1}$, and $\{|\mathbf{S}[p,0]|\}_{p=0}^{N-1}$, from Lemma 10.0.8 we have that $\overline{\tilde{\mathbf{x}}[1]}\tilde{\mathbf{x}}[2],\ldots,\tilde{\mathbf{x}}[B-1]\overline{\tilde{\mathbf{x}}[B-2]}$ are uniquely determined for almost all signals. This previous fact leads to a unique selection (up to trivial ambiguities) of $\tilde{\mathbf{x}}[1]$ in **Step 2**, and in consequence a unique selection of $\tilde{\mathbf{x}}[2]$ in this step.

Step $B-1$**:** Considering that from the $B-2$ previous steps the entries $\tilde{\mathbf{x}}[0],\ldots,\tilde{\mathbf{x}}[B-2]$ were uniquely determined (up to trivial ambiguities), appealing again to Lemma 10.0.7 we have that $\tilde{\mathbf{x}}[B-1]$ can be also uniquely determined.

Finally, analyzing the construction process described above we have that at **Step 2**, the signal $\tilde{\mathbf{x}}$ can be uniquely determined, which means that $m \geq 3B$ measurements are need to solve the radar phase retrieval problem for band-limited signals. If in addition, we have access to the spectrum signal $|\tilde{\mathbf{x}}|$, at **Step 1** we can uniquely determined $\tilde{\mathbf{x}}$ if $N \geq 3$, implying that under this scenario only $m \geq 2B$ measurements are needed.

Lemma 10.0.7. ((Bendory et al., 2019b, Corollary IV.3)) If $m \geq 2|\mathscr{J} - \mathscr{J}| - 1 + 2|\mathscr{J}|$ and $N > |\mathscr{J}|$ (that is, at least one signal entry is known), then almost every $\mathbf{w} \in \mathbb{C}^N$ is determined uniquely by $\{|\tilde{\mathbf{w}}[k]|\}_{k=0}^{m-1}$. Here, $\mathscr{J}$ is the set of indices of the unknowns, $|\mathscr{J}|$ represents its cardinality, and $\mathscr{J} - \mathscr{J} = \{n_1 - n_2 | n_1, n_2 \in \mathscr{J}\}$.

Lemma 10.0.8. ((Beinert and Plonka, 2018, Corollary 2)) Almost every complex-valued signal $\mathbf{w} \in \mathbb{C}^N$ can be uniquely recovered from $\{|\tilde{\mathbf{w}}[k]|\}_{k=0}^{N-1}$ and $\{|\mathbf{w}[n]|\}_{n=0}^{N-1}$ up to rotations.

Appendix J. Proof of Corollary 1

Recall that according to (85), we have

$$\mathbf{A}[p,k] = \left| \sum_{n=0}^{N-1} \mathbf{x}[n]\overline{\mathbf{x}[n-p]} e^{-2i\pi nk/N} \right|^2 . \tag{246}$$

Assume that $S = N/2$, N is even, that $\mathbf{x}[n] \neq 0$ for $n = 0\ldots,S-1$, and that $\mathbf{x}[n] = 0$ for $n = N/2,\ldots,N-1$. If the signal's nonzero coefficients are not in the interval $0,\ldots,N/2-1$, then we can cyclically reindex the signal without affecting the proof. If N is odd, then one should replace $N/2$ by $\lfloor N/2 \rfloor$ everywhere in the sequel. Clearly, the proof carries through for any $S \leq N/2$.

Considering (240), the band-limit assumption on the signal forms a "inverted pyramid"structure. Here, each row represents fixed p and varying n of $\overline{\mathbf{x}[n-p]}\mathbf{x}[n]$ for $p = 0,\ldots,N/2-1$

$$\begin{gathered}
|\mathbf{x}[0]|^2, |\mathbf{x}[1]|^2, \ldots, |\mathbf{x}[S-1]|^2, 0, \ldots, 0 \\
0, \overline{\mathbf{x}[0]}\mathbf{x}[1], \overline{\mathbf{x}[1]}\mathbf{x}[2], \ldots, \overline{\mathbf{x}[S-2]}\mathbf{x}[S-1], 0, \ldots, 0 \\
\vdots \\
0, 0, \ldots, 0, \overline{\mathbf{x}[0]}\mathbf{x}[S-1], 0, \ldots, 0, 0, \ldots, 0 \\
\mathbf{x}[0]\overline{\mathbf{x}[S-1]}, 0, 0, \ldots, 0 \\
\mathbf{x}[0]\overline{\mathbf{x}[S-2]}, \mathbf{x}[1]\overline{\mathbf{x}[S-1]}, 0, 0, \ldots, 0 \\
\vdots \\
\mathbf{x}[0]\overline{\mathbf{x}[1]}, \mathbf{x}[1]\overline{\mathbf{x}[2]}, \ldots \mathbf{x}[S-2]\overline{\mathbf{x}[S-1]}, 0, 0, \ldots, 0.
\end{gathered} \tag{247}$$

Then, $\mathbf{A}[p,k]$ as in (246) is a subsample of the Fourier transform of each one of the pyramid's rows.

Therefore, performing an analogous construction procedure as in Appendix 9 over (247) we have that the signal $\tilde{\mathbf{x}}$ can be uniquely determined from $m \geq 3S$ measurements. If in addition, we have access to the spectrum signal $|\tilde{\mathbf{x}}|$, we can uniquely determined $\tilde{\mathbf{x}}$ if $N \geq 3$, implying that under this scenario only $m \geq 2S$ measurements are needed.